Metagenomic analysis of river microbial communities

by

Thea Van Rossum

Bachelor of Science, University of British Columbia, 2010

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2017

Approval

Name: **Thea Van Rossum**

Degree: **Doctor of Philosophy**

Title: **Metagenomic analysis of river microbial communities**

Examining Committee:

Chair: **Dr. Michel Leroux**
Professor

Dr. Fiona Brinkman
Senior Supervisor
Professor

Dr. William Davidson
Supervisor
Professor

Dr. Margo Moore
Supervisor
Professor

Dr. Ryan Morin
Internal Examiner
Assistant Professor

Dr. Steven Hallam
External Examiner
Associate Professor
Microbiology & Immunology
University of British Columbia

Date Defended/Approved: September 22, 2017

Abstract

As concern over the availability of freshwater increases, so does the interest in river microorganisms due to their importance in drinking water safety and signalling environmental contamination. However, foundational understanding of their variability in rivers is lacking, especially for viruses. Here, I present work to improve the understanding of planktonic microbial communities in rivers over time in the context of varying environmental conditions and contrasting land use. DNA-sequencing based metagenomic and phylogenetic marker gene (16S, 18S, g23) approaches were used to profile microbial communities, coupled with measures of environmental and chemical conditions. I analysed microbial community profiles from monthly samples collected over one year from three watersheds with agricultural, urban, or minimal land use. Viral, bacterial, and microeukaryotic planktonic communities were synchronous overall, but had contrasting geographic patterns and the strength of their synchrony, as well as their relationships with environmental conditions, were heterogenous across sampling sites. These differences illustrated that bacteria are important yet insufficient representatives of microbial community dynamics despite their prevalence in microbiome research. However, this emphasis on bacteria has produced richer reference databases, which enabled a gene-specific analysis. Using a reference-based approach, I found that communities with lower water quality due to agricultural activity had higher abundances of nutrient metabolism and bacteriophage gene families. Based on these water quality associated findings and on complementary analyses, I identified potential biomarkers to demonstrate that bacterial river metagenome data could feasibly support the development of new assays for water quality monitoring. To complement these studies of anthropogenic contamination, I studied bacteria in river biofilms across a natural gradient of metal concentrations at a potential mining site. Clear relationships among metal concentrations, pH, and microbiomes were evident and this study provided fundamental knowledge of microbial communities at a potential mine site before disruption from development. Throughout these studies, the scarcity of reference information for microbial communities in lotic freshwater provided an opportunity to identify weaknesses in popular microbiome analysis methods and present approaches better suited to poorly characterised environments. Overall, my work aims to improve the understanding of planktonic river microbial community variability, both for the advancement of basic science and to support future development of more effective water quality monitoring approaches.

Keywords: metagenomics; microbiome; rivers; microbial ecology; water quality

Acknowledgements

I would like to acknowledge Dr. Fiona Brinkman for being a wonderful leader, teacher, supervisor, and mentor. It has been my privilege to receive her support in every endeavour I have pursued. I would also like to acknowledge my committee members, Dr. Margo Moore and Dr. Willie Davidson for their thoughtful discussions and support throughout my graduate studies. A large section of my research was within the context of the GC Watershed Project in collaboration with the BCCDC. I would like to acknowledge the supportive and dynamic atmosphere the leaders of this project created and the cooperative spirit that all members of the project brought to the research. I also enjoyed an excellent collaboration with Dr. Chris Kennedy and his lab members. Throughout my time in the Brinkman lab, my lab mates have provided support, creative input and fun whenever possible. I would especially like to acknowledge Dr. Mike Peabody and Dr. Emma Griffiths for their contributions to my thesis research and Geoff Winsor for guidance in my very first bioinformatics project. Finally, I would also like to acknowledge all the funding agencies that have supported my thesis work: NSERC, Simon Fraser University, Genome Canada, Genome BC, Canadian Water Resources Association, and the MBB department.

To my wonderful parents, I will never be able to say thank you often enough. But I can try: thank you. Your love, confidence and guidance encouraged me to never doubt I could achieve something if I worked consistently and creatively enough. Your unwavering support gave me the courage to risk failing. Your trust in me inspired me to trust in myself. Brett, my brilliant partner, your passion for science and your long view of its ups and downs kept me going strong throughout my graduate studies. Thank you for always reminding me to celebrate the wins, both big and small, and being a calming presence when I needed reassurance. To the rest of my family and friends, through the celebrations and commiserations, you've helped me smile, laugh, and keep moving forward. Thank you all for your wit, kindness, and encouragement.

Table of Contents

List of Tables

List of Figures

List of Acronyms

ADS	GC Watershed Project site: Agricultural downstream
APL	GC Watershed Project site: Agricultural pollution site
AUP	GC Watershed Project site: Agricultural upstream
BLAST	Basic local alignment search tool
bp	Base pairs
CCME	Canadian Council of Ministers of the Environment
kbp	Thousand base pairs
LCA	Lowest common ancestor
Mbp	Million base pairs
N	Nitrogen
NCBI	National Center for Biotechnology Information
NMDS	Non-metric multidimensional scaling
NO_2^-	Nitrite
NO_3^-	Nitrate
ORF	Open reading frame
OTU	Operational taxonomic unit
P	Phosphorous
PCA	Principle component analysis
PCoA	Principle coordinate analysis
PCR	Polymerase chain reaction
PUP	GC Watershed Project site: Protected upstream
qPCR	Quantitative polymerase chain reaction
UDS	GC Watershed Project site: Urban downstream
UPL	GC Watershed Project site: Urban pollution site
WQI	Water Quality Index

Chapter 1.

Introduction

1.1. Brief overview of river microbial ecology

River water is rich in microbial life, including bacteria, viruses, microeukaryotes, fungi, and archaea. These microorganisms collect in rivers from the surrounding upstream area ("drainage basin" or "catchment"), through overland flow, groundwater flow, aerial deposition, and streamflow, either attached to particles or free-living (Wetzel, 2001). Once in the water column, microorganisms either pass through further downstream or attach to the riverbed or banks and establish biofilms. These biofilms can also be a source of microorganisms in the water column when they shed cells. Many organisms shift from free-floating ("planktonic") to river-bed-associated ("benthic") and vice versa with stages in their life cycle or changes in environmental conditions (Sigee, 2005). The composition of river water microbial communities is shaped by environmental conditions through their effect on organisms living in the river, organisms living in the catchment, and transport processes.

1.1.1. Importance of riverine microbial communities

Clean rivers are critical for human and environmental health. They are a major source of drinking water and provide irrigation for agriculture, habitat for fisheries, and space for recreation. Human pressures can compromise these functions by decreasing river "health" in many ways, including chemical contamination (e.g. nutrients, metals, pesticides, industrial solvents) and microbial contamination (e.g. fecal coli, pathogens, invasive species) (Meybeck, 2003). Changes in microbial communities can directly reflect microbial contamination, but they can also be used to monitor chemical contamination. For example, an algae bloom might indicate that chemical contamination has occurred due to fertiliser being washed into a river, or an increase in metals tolerant bacteria might indicate metals contamination. Changes in river microbial communities can reflect contamination occurring in the water directly or in the catchment area, if it affects the incoming microorganisms. In addition to indicating the presence of contaminants, microbial communities in rivers can also transform them. For example, when nitrate or nitrite levels are high, bacteria can use these molecules as electron acceptors instead

of oxygen, which results in the permanent removal of reactive nitrogen (denitrification) (Findlay, 2010). Such processes can affect the impact from nutrient contamination on downstream receiving waters.

River microbial communities are also important components of global nutrient cycles, wherein they transform inorganic nutrients and release nutrients through decomposition of organic materials (Meybeck, 2003; Findlay, 2010). For example, microorganisms can increase phosphate levels in the water column through the decay of organic sediments and direct release of microbial metabolites (Sigee, 2005). Microorganisms also assimilate inorganic phosphorous and nitrogen directly from the water column (reviewed in (Findlay, 2010)). These processes make nutrients accessible to other organisms and support microbial growth, consequently supporting consumers up the food chain.

1.1.2. Ecological shaping forces in riverine microbial communities

1.1.2.1. General ecological shaping forces in microbial communities

In general, the ecological forces that shape the composition of microbial communities can be summarised as dispersal, selection, ecological drift, and speciation (Vellend, 2010). A community's composition changes through ecological drift due to chance demographic fluctuations and through speciation due to the evolution of new species. Dispersal and selection are discussed in more detail below.

1.1.2.1.1. Dispersal

Dispersal is the movement of organisms from one location to another followed by successful establishment (Hanson *et al.*, 2012). Microbial dispersal is often discussed in terms of the Bass Becking hypothesis, which states that "everything is everywhere, but the environment selects". The first clause of this statement asserts that there are no limitations to microbial transport, i.e. due to their small size, microorganisms could be transported long distances by wind and water currents and would not be stopped by mountains, land masses, or oceans. The second clause asserts that local conditions will determine the local microbial community. Whether transported microorganisms replicate to found a new population depends on local conditions, both abiotic and biotic (Koskella and Meaden, 2013). The likelihood of dispersal to result in new populations is increased by microorganisms' ability to persist in the environment in a dormant (metabolically inactive) state (Lennon and Jones, 2011) and the need

for only one cell to found a new population (Kirchman, 2012). This allows microorganisms to be transported over long distances and remain potentially able to found a new population over long time periods, resulting in dispersal over space and time. The accuracy of the Baas Becking hypothesis, especially the unlimited dispersal clause, appears to depend on the geographic scale and biological resolution at which microbial communities are considered (Kirchman, 2012). In general, microorganisms are easily dispersed, but there is doubt in whether this dispersal is unlimited (van der Gast, 2015).

1.1.2.1.2. Selection

Local abiotic and biotic conditions can shape microbial communities through selective processes such as competition, succession, predation, viral lysis, and emigration. Competition is a key process in shaping community composition. When some community members are better than others at using the energy sources available or withstanding local stresses, they can outcompete their neighbours for resources. This allows them to reproduce more frequently, thereby gaining abundance in a community.

Succession is another major factor in community formation, in which an established community changes the local conditions enough such that another set of organisms is better suited to the new environment than the previously successful residents. At this point, a new set of organisms out compete the existing ones and the community composition changes. Succession has limited applicability in the constantly changing water column as there is no opportunity for a standing community to establish itself, but does apply in the land and benthos surrounding the river. For example, catchment soil microbial communities may move through different compositions over time, which would lead to different inputs to the river water.

When predation and viral lysis selectively kill a type of prey, this decreases the prey's abundance. In predation, protists might contribute to shaping a microbial community by selectively grazing on bacteria based on their size or the chemical composition of their cellular surfaces (Kirchman, 2012). In viral lysis, the composition of the viral community can shape the composition of their host communities since most viruses have a narrow host range (Paez-Espino *et al.*, 2016). According to the "kill the winner" hypothesis (Thingstad, 2000), as a particular host population becomes more abundant, the likelihood of chance encounters between host cells and viruses increases, leading to an increase in viral lysis. This decreases the host population, preventing it from outcompeting the other microorganisms in the community and is hypothesised to maintain diversity in host communities. Viral lysis is thought to have a

stronger effect on microbial communities than predation due to the far larger number of viruses than grazers (Kirchman, 2012).

Microbial community composition can also change through cells physically leaving the community location. The mechanism of this movement and the cells affected depends on the environment. In soil, some microorganisms may be swept away by ground water flow. In river water, when considering at a fixed location, microorganisms leave by flowing downstream or depositing to the channel base and sides, possibly becoming part of the benthic community.

1.1.2.2. Shaping forces of free-floating microbial communities in rivers

Because near-surface river water is under constant turn over, the free-floating microbial community at a fixed geographic location is composed of immigrant microorganisms from the planktonic river community upstream and surrounding non-planktonic communities. Microorganisms from these latter sources are introduced through overland and groundwater flow from the upstream catchment land, aerial deposition, and suspension from the river bed and banks. The balance of whether a river community is more influenced by upstream aquatic communities versus benthic or terrestrial communities appears to be related to the distance between sampling point and river source (water residency time) (Read *et al.*, 2014; Niño-García *et al.*, 2016). In streams sampled close to their headwaters (the tributaries close to a stream's source), the microbial community found in the water column primarily reflects those in the surrounding land. In this case, the effect of local conditions on riverine planktonic communities is mostly indirect, and reflects the effect of local conditions on terrestrial or benthic communities. Further downstream in a river, the microbial community has had more time to respond to the local conditions within the water channel, so there may be a direct effect of local conditions on the planktonic community. Given the interplay between microbial community composition, water residency time, geography, and covarying environmental conditions, the driving shaping forces of free-floating microbial communities in river water are complex and variable. Despite this, some trends have been observed and are described below.

1.1.2.2.1. Effect of major environmental conditions on riverine microbial plankton

Determinants of planktonic microbial communities in rivers include major environmental factors such as the hours of sunlight, temperature, and rainfall (Staley *et al.*, 2015; Wang *et al.*, 2015a; Liu *et al.*, 2013). Changes in these major forces influence changes in many catchment properties, such as vegetation, agricultural activity, and runoff intensity, which in turn affect in-

water conditions such as turbidity, rate of water flow, pH levels, and nutrient concentrations. Due to this cascade of effects, it is difficult to predict the specific microbial response to a major environmental change such as higher rainfall levels. Further, the specific effect of these changing conditions on the microbial community can vary based on the specific river site. For example, the effect of a change in runoff intensity can vary based on the catchment (Sigee, 2005). When runoff levels are higher, some rivers have higher dissolved organic carbon (DOC) levels due to influx water bringing in DOC from soil, while some rivers have lower DOC due to dilution from DOC-poor incoming water (e.g. from ice melt or rainfall over paved surfaces). This relationship is further complicated when rainfall is correlated with season, as this can affect DOC levels through high activity of plankton in the summer and leaf litter in the fall, which brings in DOC from the leaves as well as bacteria and fungi that live on the leaves (Sigee, 2005). In summary, while it has been shown that changes in major environmental forces are correlated with changes in microbial communities (Zeglin, 2015), predicting what the resultant change will be may require more specific information, such as catchment conditions.

1.1.2.2.2. Effect of catchment geography and land use on riverine microbial plankton

The geography of a catchment will affect what is carried into the river. First, catchment geography and land use determines what types of particles are available for transport. For example, catchments may vary in terms of soil type, rock type, vegetation, animal life, presence of lakes, and human activity, such as urbanisation, mining, industry, recreation and agriculture. Each of these factors will affect what minerals, chemicals, organisms, or other particles might be introduced into a river. Second, catchment geography contributes to the selection of particles that are transported. For example, in a catchment with steep terrain, overland flow will be faster and thus have more energy, which will enable it to carry larger particles. Ground cover will also affect what is transported and the speed of transport, with paved or rocky areas supporting faster transport and dissolution of road salts or minerals, while forested or thick vegetation will slow water transport. These inputs affect microbial plankton directly, through the influx of organisms, and indirectly, through the change in conditions leading to shifts in competitive advantages and resultant shifts in relative abundances.

1.1.2.2.3. Effect of in-channel conditions on riverine microbial plankton

1.1.2.2.3.1. Channel hydrology

The effect of water flow on the microbial community within the water column of a stream varies according to two main properties: velocity and type of flow (laminar or turbulent) (Wetzel,

2001). Flow in streams and rivers is typically turbulent, with laminar flow only occurring in very shallow, slow-moving water (Wetzel, 2001). This turbulence discourages the establishment of multicellular aggregates.

Water residence time is the length of time something has been in a body of water (lake, stream, river, etc.). It is related to the dendritic distance to the headwaters and the flow rate. Within a catchment, bacterial taxonomic community similarity was better correlated with similarity in dendritic distance from the headwaters than Euclidean spatial distance or catchment areas (Read *et al.*, 2014). In this same study, bacterial communities were found to be more strongly shaped by water residence time than physico-chemical conditions, though some effect of copper and nitrate were seen (Read *et al.*, 2014). The physico-chemical measures taken in this study that did not appear to strongly shape the community included pH, alkalinity, suspended sediments, soluble reactive phosphorus, total dissolved phosphorus, total phosphorus, ammonia, dissolved reactive silicon, fluoride, chloride, nitrite, nitrate, total dissolved nitrogen, sulphate and dissolved organic carbon, sodium, boron, iron, magnesium, zinc, copper and aluminium.

1.1.2.2.3.2. Nutrient concentrations

Nutrient concentrations in the water channel can directly change the river plankton microbial community by affecting the growth dynamics of microorganisms in the water channel, as well as indirectly reflect a change in the incoming water from the catchment, which could be accompanied by a change in the incoming microorganisms and chemicals.

The type and amount of dissolved organic material (DOM) is shaped by the catchment's biota (e.g. plants, human activity) and physico-chemical conditions (e.g. geochemistry: pH of soil, organic matter content of soil; and hydrology: flow path, residence time in soil) (Sigee, 2005). Dissolved organic carbon from the catchment area is the main source of carbon inputs to rivers (Sigee, 2005). In a study of nearly 300 rivers and lakes sampled in northern Canada, the abundances of DOM types had more impact on bacterial communities' gene compositions than did region or ecosystem (Ruiz-González *et al.*, 2015a). In contrast, taxa were most strongly influenced by pH, water temperature, and water residence time (Ruiz-González *et al.*, 2015a).

Nitrogen, phosphorous, and silicon are the main inorganic and limiting nutrients in freshwater environments (Sigee, 2005). Nitrogen is primarily present in freshwater as nitrogen gas, nitrate, nitrite, and ammonia (Sigee, 2005). Nitrite and ammonia are mainly transient

breakdown products and are oxidized to nitrate (Sigee, 2005). Nitrogen's major biological role is as a component of amino acids and proteins. Phosphorous is primarily present in freshwater as orthophosphates and polyphosphates (Sigee, 2005). Its major biological roles are as components of nucleic acids (DNA, RNA) and structural phospholipids, and for energy transformation (ATP). Silicon is primarily present in freshwater as anionic silicates and a major component of cell walls in diatoms and other algae (Sigee, 2005). While technically a limiting nutrient, silicon is unlike nitrogen and phosphorous as it is not widely required beyond the diatoms (Sigee, 2005).

The relationship between nutrient concentrations and microbial community composition is often not straightforward. For example, increases in nitrogen do not necessarily result in increased nitrogen metabolism or protein synthesis. In an experiment where a nitrate pulse was introduced to ponds, large shifts in denitrification were concurrent with an increase in the abundance of genes involved in stress tolerance and fermentation, but changes were not observed in the abundance of nitrogen metabolism genes (Carrino-Kyker *et al.*, 2013). The relationships between nutrient concentrations and microbial communities can also vary by biological group. For example, in marine studies of microbial communities, nitrite, nitrate, and phosphorous were not found to be important drivers of overall gene composition (Sunagawa *et al.*, 2015), but had some correlations with viral communities (Brum *et al.*, 2015). In rivers specifically, some studies have found significant relationships between nutrient concentrations and bacterial community composition (Wang *et al.*, 2015a; Ibekwe *et al.*, 2016), but this is not always the case (Zeglin, 2015).

1.1.2.2.3.3. Water temperature and pH

Water temperature in a stream is affected by the temperature of incoming waters (e.g. from warm shallow ponds or cold melt waters), daily hours of sunlight, amount of shade cover, and air temperature (Sigee, 2005). A change in water temperature can indicate a possible shift in the incoming waters—and thus a coincident shift in incoming organisms and chemicals—and can affect which microorganisms are most fit. For example, different algae are most efficient at photosynthesis at different temperatures (Wetzel, 2001). In addition to its direct effects, water temperature can also affect many other water properties and conditions, such as metabolic rates and photosynthesis production, compound toxicity, dissolved oxygen and other dissolved gas concentrations, conductivity and salinity, oxidation reduction potential, water density, and pH (Fondriest Environmental, 2014).

Water temperature and pH are related chemically, with the pH level of neutral water shifting between 7 at 25°C to 7.5 at 0°C. They are also related biologically, with higher temperatures often coincident with more hours of sunlight, which often increases the amount of photosynthesis occurring, which can decrease acidity (Fondriest Environmental, 2013). Photosynthesis, respiration, and decomposition all influence the concentration of carbon dioxide and can result in diurnal variations in pH (Fondriest Environmental, 2013). Aside from temperature and sunlight, a river's pH can be influenced by catchment rock type and vegetation (e.g. due to acidic decomposing pine needles), acidic precipitation, industrial wastewater, and mining discharges (Fondriest Environmental, 2013).

In a study of nearly 300 rivers and lakes sampled in northern Canada, taxa were most strongly influenced by pH, water temperature, and water residence time (Ruiz-González *et al.*, 2015a). Bacterial taxonomic composition was also correlated with pH in an urban river (Ibekwe *et al.*, 2016). However, other studies have not found pH to be significantly correlated with composition (Read *et al.*, 2014).

1.2. Motivation to improve knowledge of microbial communities in rivers

1.2.1. Lack of foundational characterisation of riverine microplankton

Foundational understanding of microbial communities in rivers is still lacking, in part because of the difficulty in culturing riverine microorganisms (Zeglin, 2015). With the rise of phylogenetic marker gene sequencing approaches, progress has been made in describing bacterial variability (Zeglin, 2015) (reviewed in Section 1.1.2 above). However, characterisation of bacterial communities across time, especially across multiple locations, is still understudied (Zeglin, 2015). Gene-based characterisation, as opposed to taxonomic characterisation, also remains understudied and in those cases where both have been performed, important differences in their temporal variability have been observed (Ruiz-González *et al.*, 2015a).

Further, in contrast to the basic characterisation of bacterial community variability that has been achieved, little is known about the community dynamics of free-floating viruses (virioplankton) in rivers (Middelboe *et al.*, 2008; Peduzzi, 2016; Jacquet *et al.*, 2010). For example, virioplankton metagenomes have only been reported in one study where only two samples were compared (Dann *et al.*, 2016). Viral communities in lakes and oceans are better

studied; however, these communities are likely distinct from those in rivers given their differing hydrology and bacterial community compositions (Jacquet *et al.*, 2010; Aguirre de Cárcer *et al.*, 2015; Niño-García *et al.*, 2016).

Further studies of riverine free-floating microbial communities, especially over time and with consideration for viruses, is required to improve the understanding of their natural variability. This information will also support further characterisation of how their variability might be affected by contamination, for example from agricultural, urban, recreational, or industrial activities.

1.2.2. Towards the development of improved water quality monitoring techniques

Riverine microorganisms are used as sentinels of water quality. However, common approaches to monitor microbial communities in rivers require improvement. A better understanding of the natural composition and variability of free-floating microbial communities in rivers could support the development of improved methods for water quality monitoring (Dunn *et al.*, 2014).

The definition of water quality depends on the water's intended role. For example, water quality can be measured in the context of its ability to support aquatic life, for which nutrient and dissolved oxygen concentrations might be measured, or in the context of providing human drinking or recreational water, for which pathogens or fecal contamination might be measured.

Traditional water quality monitoring usually relies on spot sampling followed by tests for chemical and/or biological indicators of contamination (Allan *et al.*, 2006). The indicator paradigm is based on the idea that testing for a single or small number of characteristics can adequately and efficiently represent a broad spectrum of associated characteristics. Indicator-based water quality tests can be assessed in terms of the accuracy of the target indicator as a proxy for water quality and the accuracy of the test used to measure the indicator's presence or abundance.

A common approach to assess water quality in the context of human health is to test for fecal contamination, as this can be associated with the presence of human pathogens. A common bacterial indicator is *Escherichia coli*. Tests for *E. coli* are generally based on selective-culturing or detection of specific genetic sequences (Griffith *et al.*, 2016). Selective

culturing tests have the benefits of being well-established and only detecting live cells, whereas genetic-based tests might detect genetic material from dead organisms. However, culture-based tests have several important limitations (reviewed in Mendes Silva and Domingues, 2015). Selective culturing can result in false negative results, for example, a test for *E. coli* in bathing waters (ISO 9308-3 or ISO 9308-1) is not able to detect all strains. This includes at least one pathogenic strain (*E. coli* O157:H7) because it does not produce the enzyme that the selective culturing is targeting. This test can also result in false positives, as some non-target species produce this enzyme. Selective culturing can also be slow, with tests for *E. coli* routinely taking 18 to 24 hours. Because of these limitations, developing genetic tests for *E. coli* has been of interest, as have alternative bioindicators.

Genetic tests for water quality in a bathing context include testing for the abundance of marker sequences from bacterial groups or genes associated with feces (*Enterococcus*, total coliforms, fecal coliforms, *E. coli*, Bacteroidales species), *E. coli* bacteriophage (coliphage), humans, non-human animals (gulls, cows), and human viruses. When markers from these categories were tested at three beaches and their abundances were compared to rates of human gastrointestinal illness after exposure, site-specific conditions were found to determine which indicators were most predictive (Griffith *et al.*, 2016), with no indicators being the most predictive at all sites.

To reduce the impact of a particular marker's limitations, it can be paired with complementary assays. For example, testing one indicator each for particulate matter (turbidity, particle counts), organic matter (total organic carbon, dissolved organic carbon), and fecal indicator organisms (fecal coliforms, enterococci) has been shown to give adequate characterisation of water quality in the context of protecting a raw source of drinking water (Plummer and Long, 2007).

The combination of individual measures into an overall score for water quality, i.e. a Water Quality Index (WQI), is well established and multiple WQI formulations exist (Tyagi *et al.*, 2013). For example, the Canadian Council of Ministers of the Environment (CCME) WQI is a framework to evaluate surface water quality for the protection of aquatic life (Canadian Council of Ministers of the Environment, 2007). It requires at least four component measurements and includes guidelines for 27 measures of water chemistry, biological features, nutrients, and metals (chloride, dissolved oxygen, pH, chlorophyll, *E. coli*, total coliforms, ammonia, nitrogen, nitrate, phosphorous, lead, silver, aluminum, arsenic, cadmium, chromium, copper, iron,

mercury, molybdenum, nickel, selenium, thallium, zinc). While a compound score provides simplicity for end-users, its result greatly depends on which criteria are chosen to be included in its calculation.

The CCME WQI is an example of a test for water quality that is focused on environmental health instead of just human health. Riverine microbial communities are under threat of disturbance due to human activities, such as mining, agriculture, urbanisation, and recreation. These activities introduce physical and chemical contamination that, in addition to potentially introducing human pathogens, change local conditions and can result in microbial community shifts. These shifts can then affect ecological communities up the food chain and affect the ability of the microbial community to perform environmental services. Thus, it is also important to monitor water quality with its impact on the environment and microbial communities in mind. While the CCME WQI is well used (Tyagi *et al.*, 2013), many of the measurements required to calculate this index are expensive and slow to collect. Just as water quality monitoring for human health has been improved with new tests based on genetic identification of biomarkers, so too might environmental monitoring be improved (Dunn *et al.*, 2014).

1.2.3. The Applied Metagenomics of the Watershed Microbiome project ("GC Watershed Project")

In British Columbia, at the time of initiation of this research (2012), water quality monitoring was primarily tested after collection and distribution (i.e. at the "tap")[1]. The tests performed rely mainly on the abundance of culturable total coliforms, fecal coliforms and *E. coli* (Krewski *et al.*, 2002), which are assessed using selective culturing conditions. The presence of these microorganisms suggests that water has recently been contaminated with fecal material and indicates there may be a risk to human health due to microbial pathogens. However, as described above, this culture-based approach to evaluating water quality can lead to false positives due to non-target organisms surviving the selective culturing conditions, and false negatives due to threats to human health from contamination from non-fecal sources and non-bacterial pathogens (Goldstein *et al.*, 1996; Krewski *et al.*, 2002). Further, because this test focuses on human health, it fails to monitor for contamination that can threaten environmental health, which is intrinsically valuable as well as economically valuable due to the ecosystem services it provides.

[1] Personal correspondence with Dr. Natalie Prystajeky of the British Columbia Center for Disease Control, the government body responsible for water quality monitoring in British Columbia.

To investigate an approach to improve water quality monitoring, a national collaborative project was funded by Genome Canada and Simon Fraser University called the “Applied Metagenomics of the Watershed Microbiome” project (or the “GC Watershed Project”). This project was a collaboration between government and academic scientists to investigate improving water quality monitoring techniques using culture-free approaches based on DNA sequencing. Specifically, the project’s goal was to develop a panel of biomarkers based on microbial communities that would monitor water quality directly in river water (at the “source”), allowing for an ecosystem-level profile of water quality to monitored. Because these biomarkers were intended to monitor rivers directly, foundational information about microbial communities in rivers was required to identify effective markers.

The work reported in Chapters 2 ,3, and 4 of this thesis were performed within the context of the GC Watershed Project and the methodology involved is described in more detail in section 1.2.3 and 2.4. Briefly, monthly water samples were collected over a one-year period from six sites in three watersheds with agricultural, urban or minimal land use. For each sample, environmental, physical, and chemical conditions were recorded, and microeukaryotic, bacterial, and viral genetic material was collected and sequenced. The resultant data was analysed to characterise the variability of microbial communities in these watersheds and to support biomarker development.

1.3. Microbiome profiling methods based on DNA sequencing

To improve our understanding of microbial communities in rivers, both for basic science and to inform progress towards improved water quality monitoring, it is necessary to study the whole microbial community instead of select proxies, selected either for biological significance or practical ease. This has been challenging due to the traditional reliance on culture-based approaches in microbiology and the difficulty in culturing most environmental microorganisms (Amann *et al.*, 1995). This has led to an incomplete and biased representation of microbial diversity in ecological literature and genome databases. To address this issue, several DNA-sequencing based approaches have been developed that do not rely on culturing. These approaches allow a more complete profile of a microbial community to be compiled, and are described in sections 1.3.1 and 1.3.2 below. These sections are followed by an overview of how these profiles can be used in environmental monitoring research and some of the analysis techniques involved (section 1.4).

1.3.1. Phylogenetic marker gene profiling

Microbiome profiling methods based on phylogenetic marker genes are centered on the idea that if a gene is shared among all organisms in a clade due to vertical transfer, then differences in that gene's DNA sequence can reveal the evolutionary relationships and diversity among a set of individuals from that clade. Further, if the gene is single-copy or the copy numbers are known, then the abundance of the gene can correspond to the abundance of the type of microbe.

1.3.1.1. Phylogenetic marker sequence choice

To use a gene or DNA region as a marker for clade diversity and phylogeny, it needs to ubiquitous within the clade. If the DNA sequence will be studied using PCR, then it also needs to have regions that are highly conserved flanking a region that is highly variable, as this allows PCR primers to be designed against the conserved regions to amplify the variable region. It is beneficial if the sequence only occurs once per genome (single copy), so that a simple one-to-one sequence to cell abundance inference can be made.

In the study of bacteria, the 16S rRNA gene is often used as a phylogenetic marker gene because it mostly satisfies these criteria, with some exceptions. It is ubiquitous and there are multiple highly variable regions that can be chosen; however, the gene's copy number is variable (Stoddard *et al.*, 2015). When copy number variability is not known or not corrected for, it can introduce bias in abundance profiles that can result in bias in downstream comparisons in some, but not all cases (Kembel *et al.*, 2012). Despite this complication, the 16S rRNA gene is the most common choice for studying bacteria. One of the reasons for this is that the reference databases for 16S rRNA genes are well developed, especially for gut microbiomes, and the methods for their analysis are well established (Goodrich *et al.*, 2014). For similar reasons, the eukaryotic homologue, 18S rRNA gene, is used in eukaryotic studies. In fungal studies, the ITS (internal transcribed spacer) sequence is often used. There is no universal gene for viruses; however the major capsid genes (g23) of T4-like bacteriophages can be used to study *Myoviridae*, and RNA-directed RNA polymerase (RdRp) can be used to study RNA viruses.

1.3.1.2. Community fingerprinting / gel-based methods

The first methods developed to look at microbial communities as a whole without culturing are known as "community fingerprinting" methods. This is because they provide a picture of the microbial community that reflects the composition, but does not yield a

comprehensive list of members. Some common approaches include (Nocker *et al.*, 2007): terminal restriction fragment length polymorphism (T-RFLP) (Liu *et al.*, 1997), denaturing gradient gel electrophoresis (DGGE) (Muyzer, 1999), and automated ribosomal intergenic spacer analysis (ARISA) (Fisher and Triplett, 1999).

In all cases, DNA is collected from cells in the environment and a specific region of DNA (i.e. phylogenetic marker sequence) is amplified with PCR. The resultant amplicons are then measured in terms of character and/or abundance. In T-RFLP, amplicons are cut at specific cut sites with restriction enzymes, separated by size with gel or capillary electrophoresis, then abundance is measured by fluorescent labelling of amplicons and testing the strength of the signal (Liu *et al.*, 1997). In DGGE, amplicons are separated by melting behaviour as they move down a gel along a denaturing gradient (Muyzer, 1999). Abundance is not measured in DGGE; however, a particular band can be investigated by recovering DNA from the gel for sequencing. In ARISA, the highly variable internal transcribed spacer region (ITS1) between the 16S rRNA and 23S rRNA genes is amplified and separated by size on a gel, additionally, amplicons can be fluorescently labeled and tested for abundance (Fisher and Triplett, 1999). ARISA is similar to T-RFLP performed on the rRNA 16S rRNA gene, but uses the inherent size variability of ITS1 instead of relying on differentially fragmented amplicons. ARISA can be more sensitive and have higher taxonomic resolution (Danovaro *et al.*, 2006). In both cases, some amplicon sizes are known for organisms and can be classified based on reference databases.

Potential biases in these methods can be due to gene copy number variability leading to skewed abundance and non-unique amplicon characteristics (restriction cut sites, melting behaviour, or ITS1 length) leading to false grouping of organisms. Though physically feasible to perform these methods on many samples, the qualitative data output makes analysis of many samples challenging. These methods provide a fingerprint for a community, but cannot provide a full list of the microorganisms present.

1.3.1.3. DNA-sequencing-based methods

With dramatically decreasing DNA sequencing costs and new methodological developments, gel-based community fingerprinting methods have largely been superseded by targeted gene sequencing. These DNA-sequencing-based methods (also called DNA metabarcoding) are based on the same concept as the fingerprinting methods in that they use one DNA sequence as a proxy to describe the diversity and abundance in a clade, but they

provide more information, are easier to correct for known biases, and produce data that can be processed in a high-throughput manner.

1.3.1.3.1. Sample collection and concentration of biological material

River plankton can be collected many ways, for example, by suctioning water from a few centimeters below the surface or dipping a collection container into the flow, with the former approach more appropriate for collecting large quantities of water. Sampling benthic communities may involve using a biofilm cultivator apparatus planted in the riverbed or swabbing rocks retrieved from the riverbed. In the former case, depending on how long the apparatus is in place for, succession of the biofilm may still be underway but this method has the advantage of highlighting the organisms that are actively colonising at the time of collection.

Since different DNA sequencing or library preparation strategies are most appropriate for viruses, bacteria, and microeukaryotes, it can be useful to create separate “fractions” of a sample that are enriched for each biological group. The typical differences in particle sizes between these groups can be taken advantage of to achieve this separation using size-based filtration. This can be done by sequential filtration, with each successive filter catching cells or particles of smaller sizes (Uyaguari-Diaz *et al.*, 2016).

1.3.1.3.2. DNA extraction

Extraction of DNA from cellular microorganisms often begins when samples are frozen as part of sample storage, which breaks open cells. Additional freeze-thaw cycles are sometimes used to further this lysis. The main DNA extraction is usually done using a kit that typically uses physical (heat, bead beating) and chemical (lysis solutions) approaches to further lyse cells and cellular components (Di Bella J.M., 2013). DNA released from this lysis is then recovered through physical (filter) and/or chemical (silica, chloroform) processes.

The choice of method can significantly influence the yield and composition of extracted DNA due to variable extraction efficiencies resulting in biases, such as overly degrading DNA from sensitive organisms or failing to extract DNA from organisms with thick cell walls (Lu *et al.*, 2015). Kit choice has been shown to have significant effects on microbiome profiles (Henderson *et al.*, 2013; Wesolowska-Andersen *et al.*, 2014), but also has been shown to have less of an impact than host differences (Lu *et al.*, 2015). DNA extraction kits can also introduce contamination (Salter *et al.*, 2014), which is particularly problematic when a small amount of starting material is used.

1.3.1.3.3. Choice of phylogenetic marker sequence and amplification

After DNA has been extracted, the phylogenetic marker sequence of interest is amplified using PCR (see section 1.3.1.1 above for a discussion of phylogenetic marker gene choice). The most commonly used marker gene for bacteria is the 16S rRNA gene but there are multiple highly variable regions within that gene that can be amplified and used as markers: V1 to V9, and combinations thereof. Different regions are best suited to target all bacteria versus subclades (Klindworth *et al.*, 2013). Currently, the most commonly used primers target the V4 region: S-*-Univ-0515-a-S-19 / S-D-Bact-0787-b-A-20 (also known as 515F/806R) (Bikel *et al.*, 2015). These primers have 90% coverage of bacteria, while alternative primers for the V4 region (S-D-Bact-0564-a-S-15/S-D-Bact-0785-b-A-18) have 95% coverage when one mismatch is allowed (Klindworth *et al.*, 2013). The latter primers have been tested with simulations to evaluate their phyla coverage and only failed to detect four phyla when one mismatch was allowed (Chloroflexi, Elusimicrobia, BHI80-139 and Candidate division OP11) (Klindworth *et al.*, 2013).

1.3.1.3.4. Illumina library preparation & DNA sequencing

Illumina sequencing (Illumina, Inc., San Diego, CA) is currently the most common platform for phylogenetic marker gene studies and so I will focus on it here. For a discussion of new long-read sequencing platforms and how they might affect phylogenetic marker gene sequencing, see Chapter 6.

Sequencing libraries for the Illumina platform are prepared by attaching per-sample barcodes and sequencing adapters to amplicons using a process called “tagmentation” then pooling samples so they can be sequenced at the same time (in the same MiSeq cartridge or HiSeq lane). A sequencing library with multiple samples is called a “multiplex” library.

When a library insert sequence (the target DNA for sequencing) is longer than the maximum sequencer read length, paired-end sequencing is often performed. In this case, amplicons are sequenced from both ends sequentially. These “forward” and “reverse” reads can then be merged if there is a sufficient overlap between them (see section 1.3.1.3.6 below).

Illumina sequencers are currently the most common sequencing platform for phylogenetic marker gene studies. There are two broad classes of Illumina sequencers: “benchtop” and “production scale”, with the former more likely to be available at institutions and general research centers, while the latter is more likely only accessible through genomic or

sequencing centers. The most common benchtop and production-scale sequencers are the MiSeq and HiSeq, respectively. The MiSeq can accommodate amplicons up to 560 bp long (from 300 bp paired end reads, with 20bp overlap between paired ends) and produces approximately 25 million reads per run. The HiSeq has higher sequence yields per run (5 billion) but can only accommodate amplicons shorter than 250 bp (150bp reads with 25bp overlap between paired end reads) (Illumina).

1.3.1.3.5. Data quality control

After DNA sequences have been generated, the confidence in the accuracy of the sequences needs to be assessed and low confidence (low quality) sequence regions removed. When Illumina DNA sequencing is performed, a series of fluorescence measurements correspond to a series of nucleotides. When a fluorescence measurement is translated to a specific nucleotide, it is called a “base call”. A sequence of base calls produced from “reading” a DNA molecule is referred to as a “read”. Each base call has an associated confidence score called a “Phred quality score”. These scores range from very low confidence at 10, which corresponds to 90% probability of a correct base call, to very high confidence at 40, which corresponds to 99.99% probability. Due to the mechanics of Illumina sequencing, scores tend to decrease towards the end of a read; however, this is not always the case at a base-by-base resolution.

Removal of low confidence base calls (“quality trimming”) can be performed many ways. The simplest approach is to scan forward along a read and crop the rest of the read off once a base call is encountered that has a Phred score below a certain cut-off. This can be overly conservative if a read happens to have a single low-quality base call close to the beginning of the read with the remainder of the read being high quality. To better handle this situation, a “sliding window” approach can be used in which a read will be scanned *x* bases at a time (e.g. a window of 3 bases) and cropped once the average quality of the bases within the window fall below a certain cut-off value. Typical Phred cut-off values include 20 and 30 (99% and 99.9% probability of a correct call, respectively) depending on the downstream use of reads.

1.3.1.3.6. Paired-end read merging

If paired-end sequencing was performed, the forward and reverse read pairs can be merged together to create one longer read (also called “stitching”). There are many tools that can perform this merging. For example, PEAR (Zhang *et al.*, 2014) uses a statistical approach

to evaluate how long and how exact an overlap match needs to be for two read pairs to be merged.

When merging read pairs together, care must be taken to evaluate the proportion of read pairs that are not merged. If the proportion is large (e.g. 15%), then some of the amplicons may be longer than possible to cover by the reads being merged. In this case, simply using the successfully merged reads can introduce bias as some 16S variable regions differ by length across phyla. For example, the average length of the V3 region (amplified with primers 341f & 518r) ranges from 130 bp to 170 bp, with most Bacteroidetes longer than 140 bp and most Cyanobacteria and Fusobacteria shorter than 140 bp (Van Rossum and Brinkman, unpublished). If read lengths are 70 bp after quality trimming, then reads from the longer amplicons will not be able to be merged. If only merged reads are used, this would result in a bias against phyla with longer variable regions (e.g. Bacteroidetes in the case of the V3 region).

1.3.1.3.7. Amplicon sequence pre-filtering

Amplicon sequences may be pre-filtered to try to remove sequences that are unlikely to represent real 16S rRNA sequences. One method is to discard any sequences less than 60% similar to any 16S sequences in the reference database (this is a default filter in the QIIME pipeline). Another common filter is for chimeric sequences. Chimeras are sequences produced during PCR amplification where one end of the amplicon is from one template DNA molecule, while the other end is from a different template. The joining of these two sequences creates a sequence that does not exist in the original community. One approach to test for chimeric sequences involves testing for similarity of each end of a sequence against a reference database of reads, if the ends are each highly similar to a different reference sequence, the sequence is considered chimeric (Haas *et al.*, 2011). Since it is based on reference sequences, this method can produce false positive calls if the reference database poorly represents the studied environment. *De novo* chimera detection can be performed by using a sequence's source dataset as the reference database (Quince *et al.*, 2011). This will lessen the impact of a poor reference database but is more computationally intensive. In some newer downstream analysis approaches, such as UPARSE-OTU, chimera detection and removal is intrinsic in the method (Edgar, 2013).

1.3.1.3.8. Operation taxonomic unit (OTU) generation

Similar phylogenetic marker sequences are grouped into operation taxonomic units (OTUs), which can be thought of as data-derived taxa. An OTU is a group of sequences that have passed some similarity threshold (often 97% sequence identity) such that it is useful to analyse them as one group (taxonomic unit).

Two main frameworks used for OTU generation and analysis are QIIME (Caporaso *et al.*, 2010b) and mothur (Schloss *et al.*, 2009). The general approach used by each is similar and will be outlined below. There are three main strategies to generate OTUs from amplicon sequences: *de novo* clustering, reference-based clustering, and open-reference clustering.

1.3.1.3.8.1. *De* novo clustering

In this method, reads are clustered without the use of a reference database. *De novo* clustering is the most comprehensive method to generate OTUs and can be the most exact, but is often computationally prohibitive. This approach is recommended by the mothur developers (Westcott and Schloss, 2015). Clustering can be performed with several different algorithms, such as UCLUST (Edgar, 2010), VSEARCH (Rognes *et al.*, 2016), SUMACLUST (Mercier *et al.*, 2013), and swarm (Mahé *et al.*, 2015).

UCLUST (Edgar, 2010) is a closed-source clustering algorithm. It processes sequences in the order that they were input, which is often of decreasing abundance of exact sequence multiples or decreasing length. The first sequence becomes a centroid seed. For every subsequent sequence, if the sequence is more similar to an existing centroid seed than the threshold (e.g. 97%), it is added to that centroid, otherwise it becomes a new centroid seed. The similarity measure is a percent identity and is calculated using the USEARCH (Edgar, 2010) algorithm, where similarity between two sequences is calculated based on the proportion of shared k-mers (nucleotide sequences of length “k”). UPARSE-OTU (Edgar, 2013) is the follow-up clustering approach to UCLUST and has an additional option that if two sections of a sequence are each highly similar to two different existing centroids, the sequence is classified as chimeric and discarded.

VSEARCH (Rognes *et al.*, 2016) and SUMACLUST (Mercier *et al.*, 2013) are open source alternatives to UCLUST (Edgar, 2010) and use a similar algorithm, greedily forming clusters and using k-mers to evaluate dissimilarities between sequences. A difference in VSEARCH is that after it uses k-mer profiles to calculate pairwise similarity between query and

reference sequences, it globally aligns the top reference matches to better select the best match. This is a slow process, but is not unreasonable due to an efficient implementation of the alignment algorithm and the generally short amplicon sequences being compared. SUMACLUST was developed with amplicon sequence clustering in mind and takes a similar approach to VSEARCH. Additionally, SUMACLUST takes the abundance of cluster seed sequences into consideration, with higher abundance sequences more likely to be cluster seeds.

Swarm (Mahé *et al.*, 2015) does not use a single percent identity threshold cut off (e.g. 97% identity) to create clusters, but instead infers OTU boundaries using a single-linkage network of sequences and abundance information. Swarm determines sequence similarity using a k-mer pre-filter followed by calculation of Needleman-Wunsch distances such that distances are only calculated for highly similar sequences. Swarm creates clusters by starting with one sequence as a seed node and creating edges from that node to other sequences (new nodes) that are very similar to that original sequence. This process is continued with each new node as a seed node. In this way, swarm creates a network diagram of the amplicon sequences present in the dataset. Then, abundance information (number of exact copies of each sequence) is overlaid on this network and edges are removed where there are drop-offs in abundance. This creates OTU boundaries that are defined by both sequence identity and abundance in a way that is based on the local structure of each OTU in each dataset. While this method has the definite advantage of removing the requirement for an arbitrarily chosen OTU percent identity cut-off and takes the data's structure into consideration, it may be more difficult to compare results across studies.

As OTUs are a human-conceived analysis framework, comparison of clustering methods in the phylogenetic marker gene context is difficult due to the difficulty in identifying a universal "true" result to compare against. One attempt to compare these methods was performed as a comparison of these methods' abilities to generate biologically relevant OTUs, assessed in the context of the heritability of the human microbiome (Jackson *et al.*, 2016). They found that VSEARCH and SUMACLUST performed well but that differences between methods' results were minimal once reads were collapsed by taxonomic assignment. However, sample diversity estimates were clearly influenced by OTU clustering approach.

1.3.1.3.8.2. Closed reference clustering or reference-based clustering

In this method, reads are compared to a reference database of phylogenetic marker gene sequences. For example, a set of commonly used reference databases for closed-reference clustering are the Greengenes representative datasets (“rep set”). These datasets are composed of representative sequences for clusters that have been made from all the 16S sequences in the Greengenes database at different similarity cut-off levels (e.g. 97%, 94%, etc.). There are many methods to compare sequences against reference databases. Some are general aligners, such as BLASTN and USEARCH (Edgar, 2010), while some have been purpose built for rRNA sequences, such as SORTMeRNA (Kopylova *et al.*, 2012).

In closed-reference clustering, reads are organised into OTUs according only to their similarity to the reference, and not their similarity to each other. This approach means that clustering is fully parallelisable and so can be very fast for extremely large datasets when computational resources are available. This clustering method is potentially acceptable for environments where the real diversity is well represented by reference sequences, such as 16S rRNA sequences in artificially created communities, but will fail to report real diversity when used in environments that are not well characterised, such as freshwater and soil or with phylogenetic marker genes where a good reference database is lacking, such as ITS1. In most cases, there is little justification for discarding sequences due to a lack of matching reference sequence and closed reference clustering is inappropriate.

1.3.1.3.8.3. Open reference clustering

Open reference clustering (Rideout *et al.*, 2014) combines the speed of reference-based clustering with the sensitivity of *de novo* clustering. First, reads are clustered based on a reference database as in reference-based clustering. Any reads that were not similar to a sequence in the reference database are then set aside. From this pool, a small percentage of reads (e.g. 0.1%) are randomly selected and then clustered *de novo*. Representative sequences from each new OTU generated are then used to create a new reference database, which is used for a second round of reference-based clustering performed on the sequences that failed to cluster in the first round of reference-based clustering. Any sequences that still do not cluster after this second round of reference-based clustering are then clustered *de novo*. This approach is recommended by the QIIME developers, but discouraged by the mothur developers (Westcott and Schloss, 2015).

1.3.1.3.9. Taxonomic assignment

OTUs are assigned a taxonomic classification based on the similarity of their member sequences to a reference database. The 16S rRNA gene is represented by three main databases: Greengenes (DeSantis *et al.*, 2006), RDP (Cole *et al.*, 2014), and SILVA (Quast *et al.*, 2013). A recent comprehensive analysis of the differences, advantages and disadvantages between these databases has not been performed and is beyond the scope of this thesis (though see (Schloss, 2015; Santamaria *et al.*, 2012)). Briefly, the main differences are in the number of sequences included in the database, how their alignments were made and curated, and the kinds of sequences represented (e.g. 16S vs 18S, ITS, etc.). RDP is curated and contains approximately 10,000 sequences, whereas SILVA and Greengenes are not curated and contain approximately 150,000 and 400,000 sequences, respectively. SILVA also contains sequences in addition to 16S, including 18S, 23S, and 28S rRNA sequences, so can be used for bacteria, archaea, and eukaryotes. In the QIIME platform, taxonomy is assigned as the most specific lineage description that at least a certain percentage of OTU-member sequences share (e.g. default in QIIME is 90%).

The result of phylogenetic marker gene DNA sequencing described above is a microbiome profile: a list of OTUs present in each sample and, optionally, their taxonomic classification and abundances. Abundance of each OTU can be calculated based on the number of sequences in the OTU cluster. Further analysis of this microbiome profile is described in section 1.4 below.

1.3.2. Shotgun metagenomics

Metagenomic methods aim to characterise a biological community based on an unbiased sample of all the DNA present in an environment. Typically, DNA is collected and sequenced directly, without the use of culturing or amplification. This yields a DNA snapshot of all the genomes present in a community, which is referred to as a “metagenome”. The most established methodology uses shotgun sequencing to produce a collection of short DNA reads and is reviewed below.

Sample collection, concentration of biological material, and DNA extraction are generally performed similarly for shotgun metagenomics as in general phylogenetic marker gene studies, described above (sections 1.3.1.3.1 above and 1.3.1.3.2 above). After DNA extraction, the methodologies diverge, as described below.

1.3.2.1. Amplification

Despite concentration of biological material in the sample collection phase, the amount of extracted DNA might be inadequate for library preparation. This can particularly be an issue in studying viruses. To increase the amount of DNA, random amplification can be performed. This approach uses non-specific primers and PCR to try to increase the amount of DNA while introducing as little bias as possible. Two common methods are multiple displacement amplification (MDA) and sequence-independent single-primer amplification (SISPA). MDA introduces biases towards single-stranded and circular DNA templates, which skews the population away from viruses with double stranded or linear genomes (Kim and Bae, 2011) and can introduce chimeric sequences (Weynberg *et al.*, 2014). While SISPA does not introduce these problems, it can have a bias for amplifying the dominant sequences in a metagenomic sample, which can effectively leads to a bias towards double stranded DNA viruses because they have larger genomes than other viruses (Weynberg *et al.*, 2014). Both methods can introduce template-dependent biases resulting in uneven sequencing depth (Yilmaz *et al.*, 2010; Rosseel *et al.*, 2013).

While ideally amplification would be avoided in virome studies, DNA sequencing platforms still require more material than can generally be collected directly from the environment. Though it has had been popular in virome studies, it is cautioned against using MDA in a study where quantitative comparisons will be made (Weynberg *et al.*, 2014; Marine *et al.*, 2014; Kim and Bae, 2011). RP-SISPA has been recommended over MDA (Weynberg *et al.*, 2014), but whichever method is used, the relevant biases must be considered and caution used when comparing the relative abundances of viral groups (Weynberg *et al.*, 2014).

1.3.2.2. DNA fragmentation & size selection

Because the most common sequencing platform used is Illumina sequencing, DNA must be fragmented before it is prepared as a sequencing library. This is often done physically (using sonication or bead-beating) or enzymatically (non-specific DNase) (Di Bella J.M., 2013). If DNA is “over fragmented” this results in DNA fragments that are shorter than the sequencer’s read length. In this case, sequencing potential will be wasted as the sequencer will attempt capture the expected read length regardless of the library length and will read into the opposite end’s sequencing primers. Fragment sizes can be targeted by adjusting the enzymes or beads used during fragmentation (e.g. the quantity, size, or type of beads). Alternatively, fragments can be selected after fragmentation based on their size using gel-based size separation. When

performed manually, recovery from a gel can result in low yields and the process can be labour intensive, but high-throughput gel-based alternatives exist (e.g. Ranger technology (Coastal Genomics Inc., Burnaby, BC, Canada)).

1.3.2.3. Illumina library preparation & DNA sequencing

Illumina sequencing (Illumina, Inc., San Diego, CA) is currently the most common platform for metagenomic studies, see section 1.3.1.3.4 above. Long-read sequencing platforms (e.g. Pacific Biosciences ("PacBio"), MinION from Oxford Nanopore) are becoming more widely available and have important benefits, discussed in Chapter 6. If paired-end Illumina DNA sequencing was performed, longer "reads" can effectively be created by merging read pairs, as described above (section 1.3.1.3.6 above. If such merging is performed, care should be taken in handling unmerged reads so that any downstream classifications are not "double-counted" (Hahn *et al.*, 2015).

1.3.2.4. Assembly

If a study's goal is to consider genes or genomes, assembly of short reads may be required. In assembly, short, overlapping nucleotide sequences are merged together to create longer sequences. The assumption is that these overlapping short reads are the result of alternative fragmentations of very similar DNA sequences and that merging them together recreates these pre-fragmented sequences. There are many assemblers available, and their suitability for metagenomes varies based on the downstream analyses that will be performed and the computational resources available (see (van der Walt *et al.*, 2017) for a preprint of comparison of assemblers on metagenomes). To reduce the computational memory requirements of assembly, pre-processing of metagenomes can be performed (Brown *et al.*, 2012).

In addition to the assembler used, the quality of an assembly varies based on the community being assembled. Metagenomes will be harder to assemble (i.e. result in fewer long contigs) when they have higher complexity (e.g. quantity of genes) and lower depth of coverage (the quantity of nucleotides from reads that align to a contig, which depends on the quantity and length of reads). When assembling genomes from metagenomes, those with larger genome sizes will be more difficult to assemble, because they require more reads for complete coverage. Genomes from higher abundance community members and genes that are more

common will be easier to assemble because they will be represented by more reads in the dataset.

Differing assembly performance can introduce bias in downstream analyses. If two samples are being compared and one is better assembled (e.g. due to lower complexity), then genes, ORFs, and taxa are more likely to be called from the better assembled sample's contigs. This difference in assembly quality can lead to artefactual differences in gene and taxon abundances between samples. Due to these biases, comparisons of metagenomic communities after assembly should be performed with caution.

1.3.2.5. Reference-based analysis

1.3.2.5.1. Taxonomic classification

Metagenomic sequences can be predicted as originating from a particular taxon, this is called "taxonomic classification" or "taxonomic assignment". There are four main classes of methods: similarity-based, composition-based, marker-gene-based, and hybrids, which incorporate multiple of these methods (Peabody *et al.*, 2015). Most methods will classify a read at the lowest taxonomic level they can. For example, if a read is likely from one of three genera from the same family, then the read may be assigned to that family instead of any of the genera. This approach is called assigning at the "lowest common ancestor" (LCA). This widely used approach may be improved using algorithms that reduce overly-conservative classifications (Hanson *et al.*, 2016).

Sequence similarity-based methods use the results of a sequence similarity search of reads or contigs against a database of reference sequences with known taxonomic classification. These methods can work very well when the reference database contains sequences very similar to all community members; however, this is rarely the case in natural communities. These methods can perform adequately when reference databases contain at least sequences from closely related taxa and a broader level of taxonomic resolution is used for analysis (i.e. family level classification). Similarity can be assessed at the nucleotide or protein level, with the former being faster to run and the latter better at classifying reads with more distant references. The percentage of reads left unclassified yields an indication of how well represented the community is by a reference database.

Sequence composition-based methods base their classification on nucleotide composition characteristics that are taxon-specific, e.g. tetra-nucleotide usage or codon usage

profiles. Again, these methods will perform poorly when the reference databases do not contain genomes similar to those in the community, and poor performance will be reflected in the number of reads left unclassified.

Marker-based methods identify taxa based on the occurrence of marker sequences, either using taxon-specific versions of universal genes or using taxon-specific genes. These methods can be very fast since the database of marker sequences is generally small and have fairly high precision (Peabody *et al.*, 2015). However, their accuracy greatly depends on how well represented the community is by the database and it can be difficult to assess this representation level. Unlike composition and similarity based methods, marker gene methods do not try to classify all reads, so the proportion of reads assigned is not a diagnostic of how well the method performed. Marker-based methods are best used to profile communities with excellent reference databases or to compile a high-confidence, high-resolution, but very incomplete taxonomic profile.

For more detail on these methods and a comparison of their performances, see Peabody *et al.*, 2015.

1.3.2.5.2. Gene and gene family classification

Metagenomic gene classification provides a snapshot of the genes present in a microbiome, which indicates what microbial activities are possible, but not which are actively occurring. The most common approach is to predict the functions possible based on the metagenome's similarity to genes with known functions.

The first step is to run a similarity search against a database of genes, with query sequences either being short reads, contigs, or predicted open reading frames, often from contigs. This similarity search can be performed at the nucleotide or protein level, with the former being faster and the latter being better able to detect similarity to more distant reference sequences. Generally, a fast protein similarity search algorithm is preferred (e.g. RAPSearch2 (Zhao *et al.*, 2012) or DIAMOND (Buchfink *et al.*, 2014)). The results from a similarity search can then be processed to assign the most likely gene of origin to each query sequence and account for differences in gene length in abundance calculations (using e.g. MEGAN (Huson *et al.*, 2011), HUMANN (Abubucker *et al.*, 2012)).

Gene predictions can be summarised by their function using gene family databases (e.g. KEGG (Kanehisa, 2000), COG (Tatusov *et al.*, 2000), SEED (Overbeek *et al.*, 2005), etc.).

These databases are often hierarchical, so results can be reported at higher or lower resolution, as required. Some databases also have gene interaction network information and so assignment can be followed by pathway analysis (Konwar *et al.*, 2013), where complete metabolic pathways might be given more confidence as present in a community than incomplete pathways (Abubucker *et al.*, 2012). However, pathway incompleteness may also be due to insufficient metagenome depth of coverage.

Normalisations may be required to express relative gene abundances in a meaningful way to allow comparisons between metagenomes. Simply reporting the percentage of reads classified as a particular gene is an option; however, samples with different genome sizes will skew abundances (see Chapter 3 for more detail) and longer genes will be more likely to have more reads classified as them. A more meaningful reporting metric might be gene copies per genome, which can be achieved with extra normalisations following gene family classification, often based on the relative abundance of universal single copy genes (e.g. MicrobeCensus (Nayfach and Pollard, 2015) or MUSICC (Manor and Borenstein, 2015)).

1.3.2.6. Reference-free analysis

When reference databases are poor, reference-based analysis can produce biased and potentially misleading results. In these cases, analysis of reads without depending on a reference database can provide a less biased analysis of a set of metagenomes.

1.3.2.6.1. Predicted protein clustering

Metagenome analysis based on predicted protein clustering considers metagenomes as sets of protein clusters, with each cluster either unique or present in other metagenomes. Typically, open reading frames are predicted from assembled reads and then clustered based on a threshold of percent identity to create clusters of predicted proteins (Brum *et al.*, 2015). To get abundance information for these clusters, original reads can be mapped back to the clusters' representative sequences, creating clusters of reads instead of clusters of predicted proteins.

Metagenomes can then be compared based on their quantities of clusters and the quantities of clusters that have members from two metagenomes (alpha and beta diversity, respectively, see section 1.4.2.2 for more detail). Additionally, protein clustering can be followed up with a reference-based similarity search of the cluster representative sequences. This can

contribute to a biological interpretation of the data without limiting the initial analysis by depending on a reference database.

One issue with protein clustering is that if the depth of coverage of a metagenome is low, then many reads might not assemble. Since ORF prediction from unassembled short reads is difficult, this can result in many lower abundance genes not being represented by the created protein clusters. A useful measure is to map reads back to protein clusters and report the percentage of reads that can be assigned to a cluster. If this percentage is low, then the protein clustering approach may not be suitable or at least any interpretations must consider that the analysis is based on a potentially biased subset of the data. A notable use of protein clustering was in the virome analysis in the Tara Oceans project (Brum *et al.*, 2015).

1.3.2.6.2. Unsupervised taxonomic binning

Metagenomic sequences can be grouped ("binned") according to the taxonomic classification of their source organism to recreate taxonomic "bins" (e.g. species) (Quince *et al.*, 2017). These bins can then be compared to look at what genes are present across taxa or be used in assembly of genomes from metagenomes. When this is performed by looking for sequence similarity between a read or contig and a reference sequence with known taxonomic classification, this is considered "supervised binning" (see section 1.3.2.5.1). When a comprehensive set of references are not available, contigs or reads can be binned according to their sequence composition (e.g. short k-mer frequency profiles) and/or abundance over time (Sedlar *et al.*, 2017). This is "unsupervised contig binning". Downstream applications of binning, such as comparative genomics or genome assembly, will not work well when the metagenome depth of coverage is low.

1.3.2.6.3. K-mer analysis

K-mer analysis considers a metagenome as a set of sequences that may be unique or shared among other metagenomes. In general, a k-mer analysis breaks sequences of letters (e.g. nucleotides or amino acids) down into subsequences of length "k" (e.g. 2, 10, 21 etc.) and then compares sets of sequences (e.g. metagenomes) based on the presence or abundance of shared k-mers (character sequences of length "k"). This can give a measure of the similarity between samples ("beta diversity", see 1.4.2.3). Many tools exist to break down sequences into k-mer profiles (Jellyfish (Marçais and Kingsford, 2011)) and compare samples based on k-mer

profiles (COMMET (Maillet *et al.*, 2014), simka (Benoit *et al.*, 2016), Mash (Ondov *et al.*, 2015), khmer (Crusoe *et al.*, 2015)).

The choice of length of k-mer is a balance between longer k-mers giving better resolution between more similar samples, and shorter k-mers being more likely to be shared between less similar samples. For example, very similar samples may have only minute, uniform differences at k = 3 but may have distinct k=30 profiles that separate the samples into clusters. On the other hand, very dissimilar samples might be well separated by k=3 profiles, but not have any k=30 k-mers in common, resulting in a uniform beta diversity pattern across all samples. Empirically, a k-mer length of 21 appears to work well in most instances (Ondov *et al.*, 2015).

A strength of k-mer analysis is that, unlike protein clustering, it does not rely on assembly so it is particularly useful when depth of coverage is low. One limitation of k-mer analysis is that interpretation beyond beta diversity is usually infeasible as k-mers are usually too short to have a distinct biological description and some of the faster methods (Ondov *et al.*, 2015) do not provide a list of shared k-mers for follow up analysis.

1.4. Potential applications of microbiome DNA sequencing methods in biomonitoring

As a complement to traditional microbial ecology techniques, microbiome DNA sequencing (metagenomics and phylogenetic marker gene sequencing) provides new avenues to identify biomarkers, detect and characterise changes in riverine microbial communities, and characterise community adaptation to contaminants.

1.4.1. Development of new biomarkers

1.4.1.1. Potential for metagenomics to identify improved biomarkers

Biomarkers are biological features that indicate a particular condition. For example, in water quality monitoring, the abundance of “fecal” coliform bacteria is used as an indicator of contamination of water by fecal material (Field and Samadpour, 2007). While useful, these tests are subject to false positive results and false negative results, require specialised lab procedures, and are slow (Krewski *et al.*, 2002). Part of the reason for this is that traditional *E. coli* tests are performed by counting the number of colonies formed after plating samples of

water and culturing cells under selective conditions. Culture-based tests such as this can be slow, generally requiring at least several hours and often one to two days to produce a result (Krewski *et al.*, 2002), and can only work for a subset of possible markers as many microorganisms cannot be selectively cultured.

Metagenomic and phylogenetic marker gene sequencing approaches facilitate the identification of DNA sequences as biomarkers. Using DNA sequences as biomarkers eliminates the need for culturing cells because they can be tested for presence or abundance using PCR or qPCR, respectively. This approach also allows many markers to easily be tested simultaneously, which can generate richer data and can provide more confidence in a diagnosis than a single marker. This also allows relative abundance of biomarkers to be used as an indicator, instead of only individual marker abundances.

Many biological features identified from DNA sequences can be used as biomarkers, such as taxa, OTUs, genes, gene families, and even raw DNA sequences for which the biological source or function is unknown.

1.4.1.2. Methods to identify biomarkers from DNA data

Within a general data context, biomarkers are “features” that have differential presence or abundance among sets of samples. Features such as genes and taxa can be identified in samples by the methods described in sections 1.3.1 and 1.3.2 above. Precomputed marker gene databases generated as part of marker-based taxonomic profiling pipelines (Segata *et al.*, 2012) can be especially useful sources of biomarker sequences as the sequences’ clade specificity and taxonomic annotation has been pre-computed. Raw DNA sequences for which the biological source or function is unknown can also be used as features. These can be identified using “sequence mining” techniques where sets of DNA or amino acid sequences are compared to find short sequences that are unique or most abundant in a particular subset of samples. Methods to do this include MERCI (Vens *et al.*, 2011), which allows flexibility of matching similar amino acids but is very computationally expensive, and DFI (Weese and Schulz, 2008), which is faster but typically limited to short sequences (less than 12 bp).

A typical approach to identify categorical biomarkers is to test for differential abundance of features between sets of samples. For example, if the mean abundance of a feature in group A is statistically significantly higher than in group B, that feature might be a suitable candidate biomarker to identify other samples that likely belong to group A. This can be done with simple,

general tests such as t-tests, Wilcoxon tests, and ANOVAs, or with software designed specifically for the identification of biomarkers in microbiome data, such as LEfSe (Segata *et al.*, 2011).

A challenge with metagenomic data is that there are usually many more features to test for differential abundance than samples. Testing so many features with a typical statistical test then requires adjusting the results with multiple test correction to decrease the probability of a feature appearing differential by random chance. When there is a large difference between classes the statistical results may be so strong that this is not an issue, but if detection of subtler differences is required, simple statistical tests may not be appropriate. An alternative approach to deal with this problem can be to use machine learning techniques to narrow down the features of interest. For example, Random Forests can be used to identify those features that most reliably differentiate classes of samples or predict a continuous variable. Random Forests use repeated subsampling of both samples and features to reduce the effect of overfitting.

1.4.1.3. Design of PCR-based biomarker assays

After promising biomarkers have been identified, a test needs to be designed to assess their presence or abundance in new samples. A common approach to test for the presence of a DNA-based feature is to use PCR, because it is cost effective and relatively easy to perform. One challenge in designing effective PCR-tests for biomarkers from metagenomic or marker gene data is selecting PCR primer sequences (approximately 20bp long) that are as specific as the original biomarker sequence (i.e. original differential feature). For example, if a gene is selected because it is more abundant in a subset of samples, but PCR primers are designed to amplify a region of that gene that is a common protein domain, the PCR test may not be as specific as the original gene was in the metagenomic data. One way to address this challenge and identify regions of a feature's DNA sequence that have the same abundance pattern as the sequence as a whole is to align all the metagenomic reads against the feature's full DNA sequence then examine the alignment and pick a region where all or most of the aligned reads are from the desired samples. Once specific sequences have been identified, many tools exist to design primers (Ye *et al.*, 2012; Untergasser *et al.*, 2012).

1.4.1.4. Validation of biomarkers for accuracy and generalisability

The first phase of validation is to test the accuracy of a biomarker assay (e.g. qPCR test) against the original samples from which it was designed. This stage tests the technical validity

and the accuracy of the biomarker assay. If testing many primers against many samples, the time required quickly escalates. Batch PCR approaches can make this process more feasible, such as using a BioMark system (Fluidigm Corporation, South San Francisco, CA) machine. If the assay performs as expected against the original samples, then it can be tested against new samples. This stage of validation moves towards testing a biomarker's generalisability and therefore its utility. Further validation, quantitation, and standardization is then required before a biomarker assay can be used in regulatory or health settings.

1.4.2. Detection of community shifts

1.4.2.1. Microbiome sequencing can improve community-level microbial monitoring

Traditional methods to describe and track changes in microbial communities relied on culturing; however, these methods are biased because not all microorganisms can be cultured (Amann *et al.*, 1995). In freshwater, the bacteria that can be cultured have mostly been Gammaproteobacteria, while culture-free methods have revealed that Alphaproteobacteria and Betaproteobacteria are often more abundant but not detected in culturing (Kirchman, 2012). Similar cases have been observed in marine and soil environments (Kirchman, 2012). Therefore, culture-independent metagenomic and phylogenetic marker gene approaches offer an opportunity to produce more complete profiles of microbial communities. If followed through time or space, these profiles can be used to monitor microbial community changes.

1.4.2.2. Community characterisation: composition and complexity

Communities can be characterised in terms of their composition and complexity. Composition is a description of the identity of features that are present, such as a list of genes, taxa, or k-mers and their relative or absolute abundances. Complexity, also called "alpha diversity", is a description of the diversity of features that are present. Alpha diversity was originally defined as: "The richness in species of a particular stand or community, or a given stratum or group of organisms in a stand" (Whittaker, 1960). Within microbiome research, alpha diversity typically refers to the richness and/or evenness of organism types within a locale (e.g. the number of OTUs in one sample). There are many ways to quantify alpha diversity, with metrics reflecting the number of features present ("richness") and/or the variability in feature abundances ("evenness").

A simple measure of richness is the number of features observed (“richness”); however, this metric directly depends on the number of observations taken, which in this context corresponds to the number of DNA sequence reads per sample (“depth of sequencing”). The number of features observed can be useful to compare relative richness levels between samples with the same number of observations, but is not suitable to estimate absolute richness.

An alternative approach is to use rarefaction curves (also known as collector or complexity curves). In rarefaction, observations are subsampled repeatedly at different subsample sizes and the number of features observed is plotted against each subsample size. If the resultant curve appears to asymptote at a particular number of features, then that is an estimate of the total number of features in the population (e.g. the estimated richness). If the curve does not asymptote then the number of observations taken (the sequencing depth) is insufficient to directly estimate the richness using rarefaction. When such insufficient sampling has been performed, rarefaction curves can be extrapolated to estimate richness; however multiple curves may fit the data equally well and give very different estimates (Hughes *et al.*, 2001). Further, rarefaction curve based estimation does not provide confidence measures.

There are also statistical approaches to estimate richness, such as Chao1 (Chao, 1984), where it is assumed that sampling has been incomplete and the number of observed features is increased by a value based on the ratio of features that were only observed once versus those that were observed twice (Hughes *et al.*, 2001). In this way, Chao1 attempts to estimate the number of rare features that were not observed. The variance in a Chao1 estimate can also be calculated.

Phylogenetic diversity (PD) is similar to alpha diversity in that it quantifies the number of features within a sample; however, it also takes into account the relatedness of features. It is defined and calculated as "the sum of the lengths of all those branches that are members of the corresponding minimum spanning path" (Faith, 1992). This means that if all features were plotted as leaves in a cladogram (i.e. a phylogenetic tree diagram), then the PD would be the sum of all branch lengths required to reach all the leaves. For example, a community with 10 closely related features might have a lower PD than a community with five distantly related features. While PD is often correlated with alpha diversity, it can lead to different conclusions (Grenyer *et al.*, 2007).

Two popular metrics for alpha diversity that consider richness and evenness are the Shannon Index (also known as the Shannon's diversity index, the Shannon–Wiener index, the Shannon–Weaver index and the Shannon entropy) and the Simpson Index. While both metrics consider richness and evenness, the Shannon Index is more balanced and the Simpson Index focuses more on evenness (Kirchman, 2012).

Metagenome-specific tools have been developed to estimate alpha diversity without needing to first profile genes or taxa. For example, PHACCS (Angly *et al.*, 2005) was developed for viral metagenomes. It estimates diversity by looking at the distribution of the number of contigs versus how many reads were assembled into each contig ("contig spectra"). Contig spectra give a measure of how well a sample has been assembled, with more complex samples assembling more poorly. PHACCS fails to provide a good estimate if the sample assembles very poorly and relies on the user providing an expected mean genome size and choosing an expected distribution model of the number of features versus their abundances (e.g. the user must select if the PHACCS estimate should be based on a power, exponential, or logarithmic model). Nonpareil (Rodriguez-R and Konstantinidis, 2014) is another tool that estimates the alpha diversity of a metagenome. It assesses the existing metagenome coverage in a dataset and extrapolates how many base pairs would be required to cover the whole metagenome. It estimates the number of nucleotides required to cover a metagenome by examining the number of reads that overlap in a subsample versus the number of reads that are not similar to any other reads. This ratio is used to estimate the abundance-weighted average coverage of the metagenome. Similar to rarefaction curve analysis, repeated subsamples are taken to build a trend upon which a projection line is fit. This projection gives an estimate of the number of sequenced nucleotides that would be required to cover almost all of the diversity within a sample. Nonpareil does not require the assumption of an abundance model.

1.4.2.3. Community comparisons

Communities can be compared based on their complexity and composition. Complexity-based comparisons can be made as simple numeric comparisons based on the metrics used to describe alpha diversity such as richness or Shannon Index. Composition-based community comparisons can focus on specific features (e.g. antimicrobial resistance genes, pathogenic taxa, etc.) or all features in the community. When the latter approach is used, comparison results are often expressed as pair-wise dissimilarity values between samples.

1.4.2.3.1. Dissimilarity metrics

When assessing the similarities among microbial communities, the term “beta diversity” is often used. Originally, beta diversity was defined as “The extent of change of community composition, or degree of community differentiation, in relation to a complex-gradient of environment, or a pattern of environments […].” (Whittaker, 1960). Within microbiome research, this term usually refers to the dissimilarity among communities based on a comparison of their compositions. Such community-wide dissimilarity metrics usually reflect the ratio of the number of unshared features between two samples versus the number of shared features, with some metrics taking abundance into account. These dissimilarity metrics can be referred to as “distance metrics” if they possess the triangle inequality property, which in this case means that if three samples are considered, then the sum of any two dissimilarities among them must be greater than or equal to their third dissimilarity. The dissimilarity among samples can be calculated many ways, with some popular methods described below.

The Jaccard index represents the number of shared features between two samples divided by the number of features total in both samples, without considering abundance. Because Jaccard distances do not consider abundance, they are useful for presence/absence data but can produce unintuitive results if the abundance pattern of two samples are very different. Mash distances used by the k-mer tool (see section 1.3.2.6.3) are based on Jaccard distances but account for the length of k-mer used, as this will affect the likelihood of k-mers being shared between samples by chance (Ondov *et al.*, 2015). The Bray-Curtis dissimilarity is like the Jaccard Index as it is based on the ratio of shared and unshared features but it also considers abundance.

If there is a known hierarchical relationship between features, this can be incorporated into a dissimilarity metric. For example, the UniFrac metric was developed for use with 16S rRNA OTUs and reports communities as less similar if their features share fewer branches when organised in a phylogenetic tree (Lozupone *et al.*, 2011). This metric can either ignore or consider abundances, as formulated in the unweighted or weighted versions of the metric, respectively.

1.4.2.3.2. Analysis of dissimilarities

Once dissimilarity measures have been calculated for all community pairs, the resulting matrix of values can be analysed to reveal patterns among the samples’ beta diversities.

Categorical comparisons of sets of communities can be assessed statistically, for example using a permutational multivariate analysis of variance (PERMANOVA) test (Anderson, 2005). Here, the proportion of variability among all communities explained by a particular grouping of communities can be estimated. The higher the proportion of variability explained by the grouping, the better support for these groups representing true clusters in the data. This test will also provide a significance value (p value) that reflects whether there is a higher similarity within groups than between groups than would be expected from a randomised grouping of the data.

Substructures within the data can be visualised using ordination techniques, which aim to plot a distance matrix in two or three dimensions while preserving the maximal amount of variation among samples (Buttigieg and Ramette, 2014; Legendre and Legendre, 2012). This can reveal clustering patterns in the data. Principle component analysis (PCA) may be the most widely used method and uses eigenvectors to represent the data in a smaller number of dimensions. However, PCA is limited to using a subset of dissimilarity measures (only Euclidean distance measures) and a table of features is required as input. In contrast, principle coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS) can take a pairwise dissimilarity matrix as input. These methods all have the same goal: to plot communities in a low (2 or 3) dimensional space while maximising the correlation between dissimilarity values and geometric distances between communities in the plot. NMDS is distinct from PCoA in that it is non-metric, i.e. it uses the ranks of the dissimilarity values instead of their scales. Because it uses ranks, NMDS can accept any dissimilarity metric as input while PCoA requires metric distances. The quality of an NMDS plot is expressed as a stress value, with higher values corresponding to a lower fidelity between the resultant plot and the original distance matrix. Typical guidelines suggest that stress values higher than 0.2 indicate that a plot is a poor representation of the data (Buttigieg and Ramette, 2014). The utility of a PCoA plot can be assessed using the percentage of variability among communities that each axis represents. If these numbers sum to a low value, then the plot only represents a low amount of the actual variability in the data (Buttigieg and Ramette, 2014).

If multiple measures of the same communities are taken and provide multiple dissimilarity matrices, the similarity between two matrices can be tested using a Mantel test (Mantel, 1967). This test effectively measures the correlation between dissimilarity matrices and can test for a linear relationship using Pearson correlation, or non-linear using Spearman correlation. The statistical significance of a Mantel test statistic is assessed using permutation.

For example, if a dissimilarity matrix is acquired for the bacterial communities and the viral communities for the same samples, the similarity in beta diversity patterns between these two datasets can be assessed using a Mantel test. If there is an independent source of variability that one wishes to control for, for example if samples are collected from different sites, then a partial Mantel test can be used. This approach considers third dissimilarity matrix that represents the independent variable and calculates the residuals of the regression between this third matrix and the original two matrices. The two resultant residual matrices are then used in a standard Mantel test.

1.4.2.4. Characterisation of community differences with contextual (environmental) data

When community differences are observed over a gradient of environmental conditions or over time, concurrent changes can be identified. These are correlational patterns, not causational, but can be useful for biological interpretation and hypothesis generation.

Envfit is a method in the vegan R package (Oksanen *et al.*, 2015) that fits environmental measures onto an ordination, such as an NMDS plot, and estimates the strength of the correlation between the environmental measure and the spread of the samples in the ordination. In an envfit plot, the longer the lines corresponding to environmental variables are, the stronger the relationship between the gradient in the variable's value and the spread of the samples in the ordination. Estimates of statistical significance of these correlations is measured based on permutation tests and reported as a p value. The Mantel test (Mantel, 1967) can be used to assess the correlation between a single environmental value, or a dissimilarity matrix calculated based on multiple environmental variables. The PERMANOVA test (Anderson, 2005) can also be used to test for a relationship between community similarities and continuous or categorical variables.

In addition to natural changes in environmental conditions, these methods can be used to characterise a community response to contamination.

1.5. Goals of present research

River microbial communities are important (Sigee, 2005; Wetzel, 2001; Findlay, 2010) and under threat (Meybeck, 2003). However, when I began my research, very little was known about their composition and variability over time (Zeglin, 2015), especially for viruses

(Middelboe *et al.*, 2008; Peduzzi, 2016; Jacquet *et al.*, 2010). My research goal was to improve the understanding of how river microbial communities are affected by contamination, both to further basic science and to support the future development of water quality monitoring approaches. More specifically, I had four research questions:

1. How do planktonic river bacterial metagenomes change over a one-year time frame, and vary between sites with distinct land use and sources of contamination?

2. Can biomarker assays be developed based on bacterial genes and taxa that can differentiate between riverine communities?

3. Within planktonic riverine communities, do bacteria, viruses, and microeukaryotes vary similarly temporally, geographically, and in response to contamination?

4. How does metal contamination affect a river's microbiome? Can we identify candidate microbes that may play a role in naturally bioremediating metal contamination at a proposed mining site?

To address the first question, I analysed the metagenomes and genes present in bacterial communities sampled monthly for a year from three watersheds with agricultural, urban, or minimal land use (Chapter 2). Based on these findings and complementary analyses, I developed qPCR-based biomarker assays, demonstrating that bacterial river metagenome data could directly support the development of the type of assay that could be used in water quality monitoring (Chapter 3). Because bacterial communities are only one component of microbial life in rivers, I also investigated the co-occurring viral and microeukaryotic communities, both in terms of their synchrony and relationship with environmental conditions (Chapter 4). Finally, I had the opportunity to complement these GC Watershed Project studies of planktonic organisms affected by anthropogenic contamination (Chapters 2 - 4), with a separate study of a natural case of metal contamination. In this project, I profiled bacteria in newly forming river biofilms across a natural gradient of metal contamination and performed limited metagenomic analysis (Chapter 5). Throughout these studies, the lack of existing genomic reference information for riverine microbial communities provided opportunities to identify weaknesses in popular microbiome analysis methods and present approaches better suited to microbiome

analysis in poorly characterised environments. Overall, this work aims to improve the base knowledge of planktonic microbial communities in rivers.

I designed and performed all analyses presented in this dissertation; however, due to the interdisciplinary style of this research, all projects presented were collaborative. Every chapter begins with a foreword, which describes the focus and novelty of my work, and includes an "Author Contributions" section to clearly attribute work.

Chapter 2.

Year-long metagenomic study of river bacteria across land use and water quality

2.1. Foreword

This chapter was published in the journal *Frontiers in Microbiology* and co-authored by Thea Van Rossum, Michael A. Peabody, Miguel I. Uyaguari-Diaz, Kirby I. Cronin, Michael Chan, Jared R. Slobodan, Matthew J. Nesbitt, Curtis A. Suttle, William W.L. Hsiao, Patrick K.C. Tang, Natalie A. Prystajecky, and Fiona S.L. Brinkman. (Copyright © 2015 Van Rossum et al.) For supplementary material referred to in the text, see http://journal.frontiersin.org/article/10.3389/fmicb.2015.01405/full.

This work was performed within the GC Watershed Project (see section 1.2). Design of the sampling scheme and production of the data was a large collaboration (see author contributions in section 2.7). My role in this project was to aid in DNA sequencing optimisation and perform the data analysis, interpretation, and reporting (see section 2.7 for author contributions). Specifically, I led the gene-based analysis of bacterial metagenomes and Dr. Michael Peabody led the taxonomic-based analysis, which is reported in a separate manuscript (in preparation). The novelty of my approach in the analysis of the bacterial metagenome data was in using an assembly- and reference-free analysis method (k-mer profiling), which was rare at the time of publication and is still uncommon. I also explicitly considered normalisation issues that are often ignored, including the implications of variable average genome sizes and percentages of reads assigned to gene families.

2.2. Abstract

Select bacteria, such as *Escherichia coli* or coliforms, have been widely used as sentinels of low water quality; however, there are concerns regarding their predictive accuracy for the protection of human and environmental health. To develop improved monitoring systems, a greater understanding of bacterial community structure, function and variability across time is required in the context of different pollution types, such as agricultural and urban contamination. Here, we present a year-long survey of free-living bacterial DNA collected from seven sites

along rivers in three watersheds with varying land use in Southwestern Canada. This is the first study to examine the bacterial metagenome in flowing ("lotic") freshwater environments over such a time span, providing an opportunity to describe bacterial community variability as a function of land use and environmental conditions. Characteristics of the metagenomic data, such as sequence composition and average genome size, vary with sampling site, environmental conditions, and water chemistry. For example, average genome size was correlated with hours of daylight in the agricultural watershed and, across the agriculturally and urban-affected sites, k-mer composition clustering corresponded to nutrient concentrations. In addition to indicating a community shift, this change in average genome size has implications in terms of the normalisation strategies required, and considerations surrounding such strategies in general are discussed. When comparing abundances of gene functional groups between high- and low-quality water samples collected from an agricultural area, the latter had a higher abundance of nutrient metabolism and bacteriophage groups, possibly reflecting an increase in agricultural runoff. This work presents a valuable dataset representing a year of monthly sampling across watersheds and an analysis targeted at establishing a foundational understanding of how bacterial lotic communities vary across time and land use. The results provide important context for future studies, including further analyses of watershed ecosystem health, and the identification and development of biomarkers for improved water quality monitoring systems.

2.3. Introduction

High quality freshwater is an essential natural resource and is increasingly threatened by human activities (Vörösmarty *et al.*, 2010). Microbes have long been used as sentinels of poor water quality due to their potential to be host specific, which enables source tracking, and/or sensitivity to changes in the environment (White and Wilson, 1989; White *et al.*, 1998). For example, *Escherichia coli* or coliform bacterial abundance is widely used as a proxy to monitor fecal contamination in drinking and recreational water. However, these tests rely on culturing under selective conditions and are susceptible to both false positive and false negative results (Goldstein *et al.*, 1996; McLain *et al.*, 2011). In clinical settings, pathogen-specific diagnostics have moved towards PCR-based tests due to increased sensitivity and specificity and decreased turnaround times (Espy *et al.*, 2006). A similar same methodological shift has begun in environmental monitoring for human health (e.g. United States Environmental Protection Agency Method 1609 to test for Enterococci with qPCR); however, effective biomarkers of

ecosystem health have yet to be developed. Recent studies have identified taxonomic ratios associated with fecal contamination (Garrido *et al.*, 2014; Li *et al.*, 2015) and PCR-based tests for fecal indicator bacteria have been developed, yet these still suffer the weaknesses of being either broadly distributed but susceptible to false positives or are highly specific but susceptible to false negatives (Harwood *et al.*, 2014; van der Wielen and Medema, 2010). Further, these tests focus on fecal contamination, which is valuable yet insufficient to monitor the health of an aquatic ecosystem (Palmer and Febria, 2012). Rivers are the chief source of renewable water for humans and freshwater ecosystems (Vörösmarty *et al.*, 2010), yet microbial diversity in lotic (i.e. flowing water) communities is less commonly studied than in marine or lake ecosystems, and among those studied, contaminated systems are underrepresented (Zinger *et al.*, 2012). In order to develop better tests for ecosystem health, foundational data is required to answer fundamental questions about lotic microbiota such as how they are affected by land use and how they vary over time.

Microbial communities can be described in terms of levels of diversity (e.g. richness, evenness) and composition (which taxa and genes are present). Assessing the former has been possible for decades, and the importance of characterising microbial communities is reflected in the hundreds of studies that have investigated how levels of diversity are affected by environmental changes (Zeglin, 2015). However, these studies have been limited to describing community diversity, instead of composition, due to the technologies available at the time. While important for biological understanding, metrics of diversity are difficult to translate into diagnostics. With the development of microbiome studies based on DNA sequencing, the effect of environmental conditions on microbial community composition has begun to be revealed. For example, using taxonomic characterization of riverine bacterial microbiomes, studies have shown that bacterioplankton communities vary by location and nutrient concentrations (Jackson *et al.*, 2014; Hu *et al.*, 2014; Read *et al.*, 2014; Wang *et al.*, 2015b; Savio *et al.*, 2015; Ruiz-González *et al.*, 2015a); however, these studies were limited by the duration of the sampling periods, which were less than a month, with most sites only being sampled once. In a three-year study of bacterioplankton community composition, annual community reassembly was observed and seasonal shifts were inferred, though this study focused on spring and summer, with only two samples collected in the fall and none in the winter (Fortunato *et al.*, 2013). Seasonal taxonomic shifts could be important in predicting the effect of contamination on microbial communities. For example, a study that sampled an urban river once per season for one year found variability in the recovery of bacterial community composition after exposure to sewage

effluent (García-Armisen *et al.*, 2014). Further investigating anthropogenic contamination, another study, limited to the early summer over two years, indicated that land use affected the taxonomic composition of bacterial communities in a river across forested, urban and agricultural sites (Staley *et al.*, 2014a).

Beyond taxonomic characterisation, two studies have described the metagenomes (total genetic material) present in lotic free-living microbial communities. The first study looked at one sample from the pristine upper course of the Amazon River and saw an overrepresentation of pathways involved in heterotrophic carbon processing, especially from plant material, as compared with marine samples (Ghai *et al.*, 2011). The second looked at 11 sites in the Upper Mississippi River watershed with near-by land use characterised as forested, urban or agricultural (Staley *et al.*, 2014b). In this research, the authors found that the core functional traits were largely stable across sampling sites; however, this study was based on one sample from each site and did not examine potential land use effects over time.

We hypothesized that metagenome characteristics and functional traits vary across sampling sites in lotic bacterial communities, but that this variability may require sampling across time to better observe and characterise them. Here, we present a year-long survey of bacterial genes in sampling sites across rivers in protected, agricultural, and urban watersheds. We found that metagenomic characteristics, as well as abundance of gene functional groups vary with time, land use and environmental conditions. Further, we discuss trends in average genome size (AGS), which have been shown to be important for analysis of gene functions (Nayfach and Pollard, 2015; Manor and Borenstein, 2015), and demonstrate how normalisation techniques are important to consider in these analyses. The data presented in this study is a valuable resource for the study of bacterial lotic communities. The analyses presented here aim to provide a foundational interpretation of this data that contributes to the understanding of the variability of lotic bacterial microbiomes over time and land use. This work will provide support for future developments in water quality monitoring, both directly from the sequence data presented, which could be used to identify DNA sequence biomarkers indicative of land use or water quality conditions, and indirectly by providing context aiding design of future water quality studies.

2.4. Methods

2.4.1. Sample collection, DNA sequencing, and environmental measurements

Monthly samples of flowing freshwater were collected from seven sites in three watersheds up to 130 km apart in Southwestern British Columbia, Canada. Some of the sites were pristine (protected from land use) while others were affected by agricultural or urban activity (Table 1). Watersheds were sampled monthly on different days. All sites within a watershed were sampled on the same day within two and a half hours. Twelve samples were collected from the urban watershed and 13 were collected from both the protected and agricultural watersheds. All rivers were less than 10 m wide at the sampling location except ADS, which was 30 m wide. Near-surface water was collected from rivers at all sites expect PDS, where water was collected after residence in a reservoir and passing through a pipe. PDS is the only site downstream of a lake or reservoir.

Table 1 Description of sampling sites across watersheds with varying land use.

Watershed	Site Name	Catchment land use	Description
Agricultural	AUP (Agri-Upstream)	Forest & minimal residential	Upstream of agricultural "pollution". Not affected by agricultural activity. Collected from a small rocky stream near the base of a forested hill with minimal housing nearby.
	APL (Agri-Pollution)	Agriculture	At site of agricultural "pollution". Collected from a slough in an intensely farmed and irrigated floodplain with minimal tree cover. AUP is upstream of floodplain, separated by 9 km.
	ADS (Agri-Downstream)	Agriculture & some urban	Downstream of agricultural "pollution". Collected from a river fed by an agricultural floodplain (site of APL) as well as a separate tributary from a more distant agricultural and urban area. Minimal tree cover throughout catchment. ADS is 2.5 km from APL.
Urban	UPL (Urban-Pollution)	Forest & urban	At site of urban "pollution". Collected from a stream that originates in mountainous forest then passes through 300 m of residential development.
	UDS (Urban-Downstream)	Forest, parks & urban	Downstream of urban "pollution". Collected downstream of UPL, after passing through 1 km of residential neighbourhood (half houses and half treed parks).
Protected	PUP (Protected)	Forest	Collected from river in forested, protected watershed that feeds a drinking water reservoir. Collected 1 km upstream of entry point to reservoir.
	PDS (Protected-Downstream)	Forest, reservoir & pipe	Downstream of PUP-fed reservoir, which is 1km wide by 7 km long. Sample collected after reservoir water has passed through a 9 km long pipe, 2 m in diameter. Water enters pipe from reservoir on the opposite side of reservoir from PUP. PDS is 16 km from PUP.

At each sampling event, 40L of water were collected, pre-filtered through a 105-µm pore-size spectra/mesh polypropylene filter (SpectrumLabs, Rancho Dominguez, CA) to remove larger particles then transported on ice to the lab for further processing within 24 hours. Water was filtered sequentially through 1 µm pore-size filter (Envirochek HV, Pall Corporation, Ann Harbor, MI) then a 0.2 µm pore-size filter (142 mm Supor-200 membrane disc filter, Pall Corporation, Ann Harbor, MI) to collect bacterial and archaeal sized cells. This pre-filtration did also remove larger and particle-associated bacterial cells, and of course bacteria adhering to large items like leaves will not be collected. Cells were collected off the 0.2 µm pore-size filters by vortexing with tungsten beads (i.e. bead beating) and centrifugation. DNA was extracted using the PowerLyzer Powersoil DNA Isolation Kit (MoBio, Carlsbad, CA). Cells were mainly bacterial and not archaeal (data not shown), so for simplicity, the microbial community is here referred to as bacterial.

A positive control (mock community) was prepared by spiking de-ionised water with DNA extracted (NucleoSpin Tissue, Macherey-Nagel, Düren, Germany) from 12 cultured bacterial strains (*Bacillus amyloliquefaciens* FZB42, *Bacillus cereus* ATCC 14579, *Burkholderia cenocepacia* J2315, *Escherichia coli* K-12, *Frankia* sp. CcI3, *Micrococcus luteus* NCTC 2665, *Pseudomonas aeruginosa* PAO1, *Pseudomonas aeruginosa* UCBPP-PA14, *Pseudomonas fluorescens*, Pf-5, *Pseudomonas putida* KT2440, *Rhodobacter capsulatus* SB 1003, *Streptomyces coelicolor* A3(2)). Ultrapure (Type 1) water (Milli-Q, Millipore Corporation, Billerica, MA) was used as a negative control.

Shotgun sequencing libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina, Inc., San Diego, CA). Gel-size selection was automated with Ranger Technology (Coastal Genomics Inc., Burnaby, BC) to ensure consistent and specific fragment lengths, targeting 500-800 bp (Uyaguari-Diaz *et al.*, 2015). Sequencing was performed using a MiSeq platform (Illumina, Inc., San Diego, CA) using the MiSeq Reagent Kit V2 (2x 250 bp paired-end reads, 500 cycles) at the British Columbia Public Health Microbiology and Reference Laboratory. Samples were sequenced over seven runs and a positive and negative control sample was included in each run. All raw sequences are deposited in the NCBI Sequence Read Archive under BioProject ID: 287840.

Physical water quality parameters were measured *in situ* using a YSI Professional Plus handheld multiparameter instrument (YSI Inc., Yellow Springs, 96 OH) and VWR turbidity meter (model 66120-200, VWR, Radnor, PA). Chemical parameters were measured as follows:

dissolved chloride using an automated colorimeter (SM-4500-Cl G), ammonia using phenate methods (SM-4500-NH3 G) (Eaton and Franson, 2005), orthophosphates using chemical precipitation (Murphy and Riley, 1962), and nitrites and nitrates using spectrometry (Wood *et al.*, 1967). Chlorophyll *a* was measured using fluorometric analysis (Welschmeyer, 1994). Abundances of bacterial size particles were estimated using a FACSCalibur flow cytometer (Beckton Dickinson, San Jose, CA) with a 15 mW 488 nm air-cooled argon-ion laser (Brussaard *et al.*, 1999) followed by analysis using CYTOWIN version 4.31 (2004) (Vaulot, 1989). Precipitation records and hours of daylight were collected from the Canadian Climate Data database (Environment Canada, 2013), from the closest station per watershed (station IDs: agricultural 1100031, urban 1105669, protected 1017230). Missing rainfall values were replaced with the mean value across two days before and two days after sampling. Rainfall values used for analysis are cumulative over three days prior to sampling. Minimal snowfall occurred. Other missing values were replaced with sampling site means. The Canadian Council of Ministers of the Environment's Water Quality Index was calculated using their provided spreadsheet (Canadian Council of Ministers of the Environment, 2007) based on guidelines for ammonia, chloride, DO, nitrate, pH and phosphorous.

2.4.2. Statistical analyses

All statistical analyses were performed using R 3.2.0, including methods from the Vegan package (Oksanen *et al.*, 2015). Reported means and medians are followed by the standard deviation, proceeded by the "±" symbol. Correlations were assessed using Spearman correlation coefficients, unless otherwise noted. Where applicable, *p* values were corrected for false discovery rate using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Significance test values were considered significant if *p* values were less than 0.05 and *q* values were less than 0.1.

2.4.3. Metagenome compositional analysis

Shotgun sequenced reads were trimmed to remove low quality bases using Trimmomatic (Bolger *et al.*, 2014). A sliding window of length 5 and a minimum Phred score of 20 was used at the 3' end. At the 5' end, sequences of one or more nucleotides with scores less than 20 were trimmed. Sequencing adapters were removed using Cutadapt (Martin, 2011), overlapping paired-end reads were merged using PEAR (Zhang *et al.*, 2014), and reads shorter than 100 bp were discarded. After this processing, samples had 418538 to 2165162 high

quality reads and samples were subsampled to: 418,500 reads. The January sample from site PUP was omitted from all analyses due to too few reads.

Environmental conditions associated with the samples were compared based on scaled measures using Euclidean distances, clustered using Ward's method (R: ward.D2), and visualised using non-metric multidimensional scaling (NMDS) with two axes. Hierarchical clustering was performed and evaluated using the pvclust R package (Suzuki and Shimodaira, 2006). Bootstrap *p* values were based on multiscale bootstrap resampling with 10000 repetitions.

The k-mer abundance profile of each sample was calculated by counting the frequencies all nucleotide sequences (k-mers) of length 12 in each dataset using Jellyfish (Marçais and Kingsford, 2011). These k-mer profiles were compared using Manhattan distances, as appropriate for high dimensional datasets (Aggarwal *et al.*, 2001), clustered using Ward's method (R: ward.D2) and visualised using hierarchical clustering and NMDS with two axes. Major sample clustering patterns discussed below were consistent with k-mer lengths 4, 9, and 10 (data not shown), but were most distinct with higher k-mer lengths.

Possible batch effects from collecting DNA sequence data over multiple Illumina MiSeq sequencing runs were assessed by comparing positive controls included in each run and checking for run-based clustering of samples (using NMDS visualisation based on k-mer profiles). In both cases, no indication of batch effects was observed.

2.4.4. Calculation and normalisation of functional gene group abundances

Reads were compared against NCBI's nr database (downloaded April 9, 2014) using RAPSearch2 (Zhao *et al.*, 2012). Resulting protein alignments longer than 30 amino acids were analysed using MEGAN version 5.10 (Huson *et al.*, 2011) with default parameters, including a minimum bit score of 35 and a maximum e-value of 0.01, to determine the gene families present, using the SEED (Overbeek *et al.*, 2005) and KEGG gene functional group databases (Kanehisa, 2000). Gene group abundance profiles were analysed for differential abundance using the Wilcox test after removing low abundance features (mean abundance < 0.01% in all samples). Subsamples of 100,000 reads were also analysed using MG-RAST (Meyer *et al.*, 2008) with default parameters for comparative purposes.

The quality of the MEGAN assignments of reads to gene groups was assessed using the mock community samples. These samples of known taxonomic composition were annotated using MEGAN and the KEGG database of ortholog groups. These mock community KEGG profiles were compared against reference profiles, compiled from annotated genomes. The KEGG database was used for this analysis due to the availability of annotated genomes; however, the version of the KEGG database used for this reference annotation was not the same as the version used in the MEGAN analysis. Ortholog groups missing from one of the two profiles under comparison, possibly due to differences in database version, were omitted from this analysis, leaving 1725 KEGG ortholog groups to compare. Annotation profiles were fairly well correlated between the MEGAN and reference datasets when looking at KEGG ortholog groups (r = 0.74, p < 2.2e-16, Pearson correlation used due to interest in linear relationship). This correlation improved when looking at KEGG pathways (r = 0.96, p < 2.2e-16). Of the 208 pathways, two in particular were predicted as less abundant in the MEGAN profiles relative to the reference profiles: “Ribosome” and “ABC transporter”. When these pathways were removed, the correlation rose to r=0.98. Adjusting abundance profiles by the average KEGG ortholog group gene length improved the correlations between ortholog group profiles (r = 0.86, p < 2.2e-16) but the improvement was minimal for pathway profiles (r = 0.99, p < 2.2e-16).

Normalisation of gene functional group abundance profiles by AGS was performed on subsampled reads with two approaches. The first used MicrobeCensus to estimate AGS values (Nayfach and Pollard, 2015), which were then divided by the mean AGS across samples (to avoid inconveniently large numbers) and then multiplied by group abundances; the second used MUSiCC, which adjusts group abundances directly (Manor and Borenstein, 2015). Both tools are based on the same goal: to calculate normalisation factors such that normalised universal, single copy gene abundances will be constant across samples. These tools assume that all reads are bacterial and so can be affected by the presence of eukaryotic DNA sequences. Due to the filtration strategy used during sample processing, very little eukaryotic DNA was present in the samples (median 2 ± 0.7% of domain-assigned reads). Both tools gave very similar results, with an overall Pearson correlation of 0.998 (p < 2.2e-16) between KEGG ortholog group abundance profiles across all samples, and a correlation score of 0.997 (p-values < 2.2e-16) within each sample. Currently, MUSiCC only accommodates KEGG and COG profiles and normalises assigned reads, whereas MicrobeCensus works directly on reads to estimate AGS and therefore allows the flexibility of using any downstream functional assignment tool. In the analyses that follow, MicrobeCensus normalisation is used.

2.5. Results & Discussion

2.5.1. Contamination and water chemistry is reflected in reference-free clustering of metagenomes across land use

Metagenomic shotgun sequencing of free-living bacterial communities was performed on 89 samples, collected monthly from seven sites across three watersheds under varying land use (protected, urban or agricultural) (Table 1). The agriculturally affected sites (APL & ADS) had the highest concentrations of nutrients and were the most distinct in terms of water chemistry, while the urban affected sites (UPL & UDS) and the unaffected sites (PUP & AUP) were more similar and had higher concentrations of dissolved oxygen (Table 2 and Figure 1).

Table 2 Summary of environmental variables over one year of sampling: means and standard deviations.

		Agricultural			Urban		Protected	
Measured Variable	Abbrev.	AUP	APL	ADS	UPL	UDS	PUP	PDS
Water properties								
Ammonia mg/L	NH3	0.03 ± 0.08	0.52 ± 0.36	0.16 ± 0.11	0.01 ± 0.01	0.02 ± 0.01	0.01 ± 0	0.02 ± 0.03
Dissolved chloride mg/L	Cl	1.9 ± 2.9	14.9 ± 9	10.7 ± 3.7	6.2 ± 3.3	10.0 ± 3.7	2.1 ± 0.5	2.6 ± 0.3
Dissolved oxygen mg/L	DO	15 ± 5	5 ± 3.8	10 ± 3.4	11.7 ± 2.5	11.6 ± 2	11.5 ± 1.4	10.7 ± 1.8
Nitrate mg/L	NO3	2.8 ± 0.5	4.4 ± 3.6	8.4 ± 1.3	2.8 ± 1	2.8 ± 1	0.2 ± 0.2	0.09 ± 0.04
Nitrite mg/L	NO2	0.01 ± 0	0.2 ± 0.2	0.1 ± 0.07	0.01 ± 0.01	0.02 ± 0.01	0.02 ± 0.06	0.01 ± 0.01
pH	pH	7.2 ± 0.3	6.9 ± 0.3	7.3 ± 0.2	6.4 ± 0.4	7.0 ± 0.3	6.6 ± 0.7	6.8 ± 0.4
Orthophosphate mg/L	P	ND	0.11 ± 0.11	0.13 ± 0.11	0.01 ± 0.01	0.01 ± 0	ND	ND
Specific Conductivity uS/cm	SpecCond	106 ± 23	305 ± 31	266 ± 51	75 ± 28	105 ± 26	77 ± 35	52 ± 21
Turbidity NTU	Turbid	2.3 ± 5	19 ± 11	19 ± 15	0.8 ± 0.5	1.8 ± 1.2	0.3 ± 0.1	0.4 ± 0.1
Temperature °C	TempC	8 ± 3	12 ± 5	12 ± 5	9 ± 4	9 ± 4	8 ± 5	10 ± 4
Biological measures								
Chlorophyll *a* µg/L	ChloA	0.4 ± 0.4	2 ± 2	2 ± 2.3	0.1 ± 0.1	0.6 ± 1.2	0.1 ± 0.05	1 ± 0.5
Flow cytometry million cells/mL	CellCnt	0.4 ± 0.7	2 ± 1.5	1.8 ± 1.3	0.3 ± 0.2	0.3 ± 0.2	0.2 ± 0.4	0.5 ± 0.6
Environmental measures								
Rainfall mm	Rain0to3	22 ± 24	22 ± 24	22 ± 24	19 ± 19	19 ± 19	14 ± 23	14 ± 23
Daylight hours	LightHrs	13.6 ± 2.7	13.6 ± 2.7	13.6 ± 2.7	13.5 ± 2.8	13.5 ± 2.8	13.5 ± 2.8	13.5 ± 2.8

ND (non-detect): value was below detection limit.

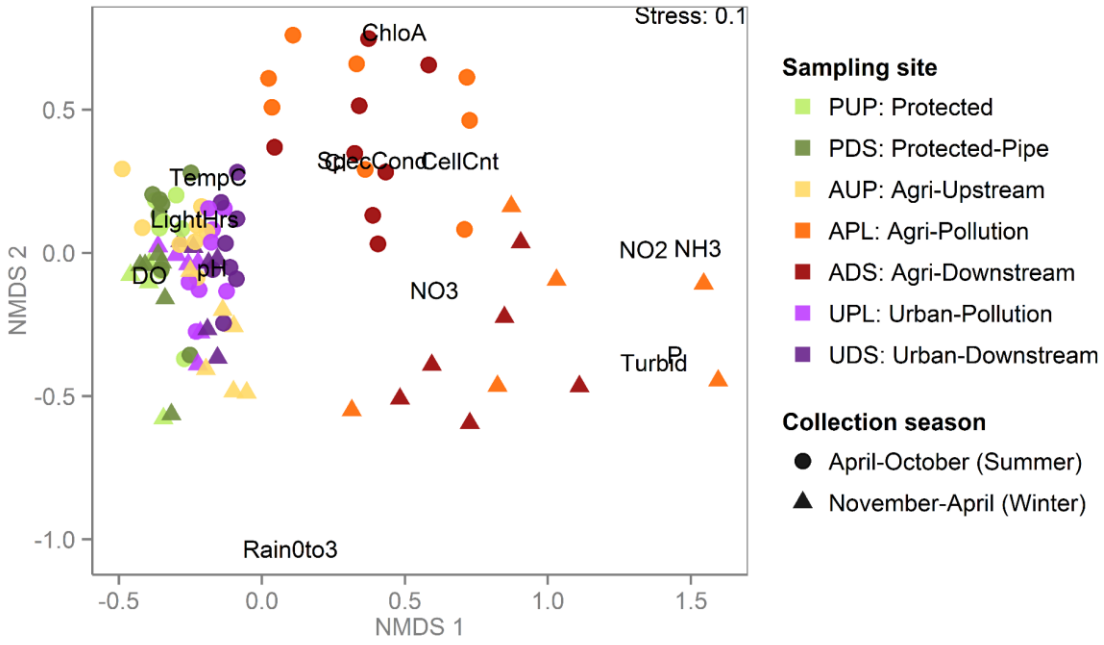

Figure 1 **Agriculturally affected sites are distinct from protected and urban sites when clustered by water chemistry and environmental variables.**

Non-metric multidimensional scaling (NMDS) plot based on environmental and chemical water measurements in which each point represents a sample, coloured by sampling site and shaped by season during which the sample was collected. Environmental measures are abbreviated as in Table 2. Samples from the agriculturally affected sites (APL & ADS) are most distinct, with summer and winter samples mostly clustering together, reflecting the winter's increased rain and consequent agricultural runoff. All other samples are more similar to each other than to the agriculturally affected sites, including the samples from the agricultural watershed that were collected upstream of agricultural activity (AUP).

Clustering metagenomes by the abundance of constant-length DNA sub-sequences (k-mers) has been shown to be an effective way to characterise microbiomes without the biases or limitations of existing microbial references (Hurwitz *et al.*, 2014; Jiang *et al.*, 2012). Here, clustering river metagenomes based on k-mer abundance resolves samples into clusters that share common sampling sites, watersheds, or environmental conditions (Figure 2A). Using hierarchical clustering, we observe 2 major clusters, I and II, that each divide into two and three smaller clusters, respectively. Cluster I is composed of two sub-clusters: Cluster 1, which is composed of 12 of 13 samples from the post-pipe (PDS) site, and Cluster 2, which composed of a subset of samples collected from the agriculturally affected sites (APL, ADS). Cluster II is composed of three sub-clusters: Clusters 4 and 5, which mostly contain samples from the unaffected (AUP, PUP) and urban sites (UPL, UDS), and Cluster 3, which contains samples both from the agriculturally affected sites and from the urban sites.

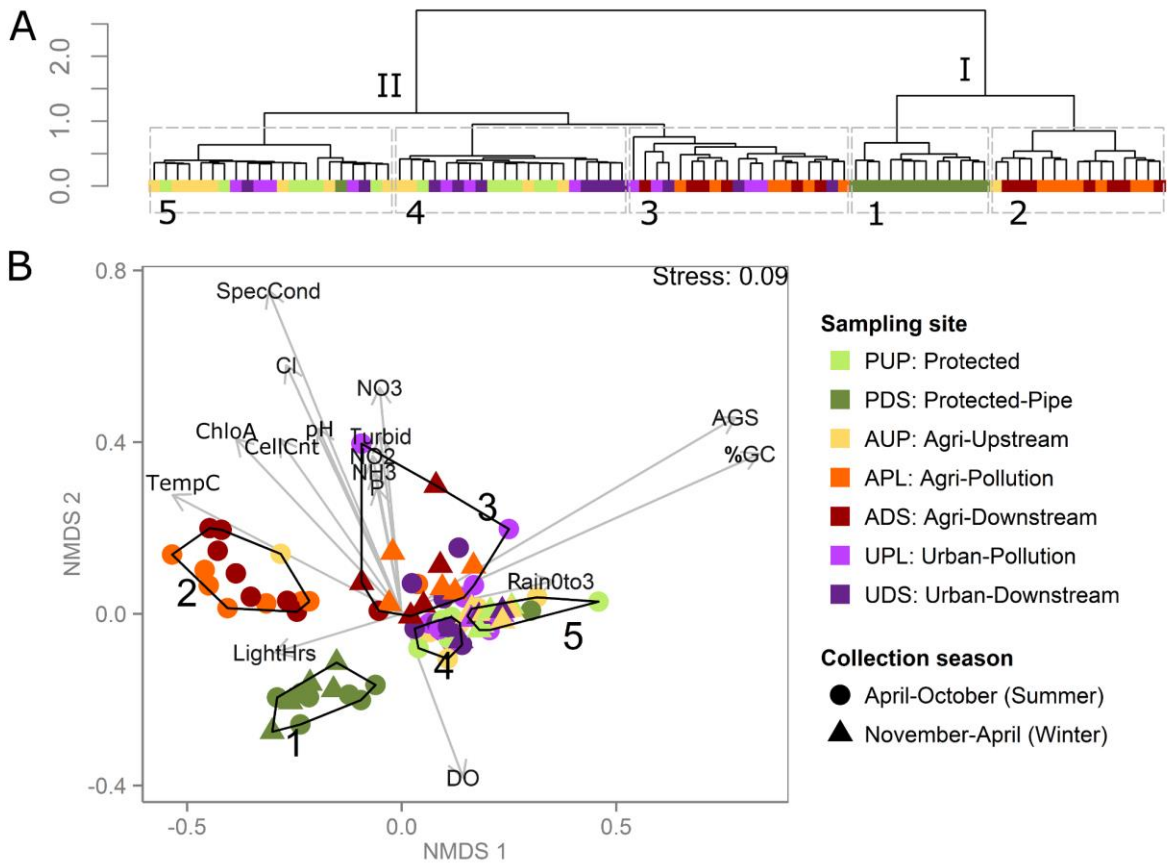

Figure 2 **Metagenomes clustered by reference-free k-mer analysis show effect of sampling site, weather conditions, and water chemistry.**

(A) Hierarchical clustering of samples based on metagenome k-mer composition. Each terminal node (leaf) is a sample, colored by sampling site. Roman numerals label major clusters, Arabic numerals label sub-clusters outlined in grey dashed lines. **(B)** NMDS plot based on k-mer abundance distributions in which each point represents a sample, coloured by sampling site. Clusters outlined and numbered in black correspond to numbered clusters in panel A. Environmental variables were correlated with ordination axes using envfit and are displayed using grey arrows, where lengths of arrows correspond to the strength of the correlation between the variable and the ordination (only variables with $p < 0.05$ displayed) and direction corresponds to increasing value (e.g. samples closer to the bottom of the plot have higher DO). Environmental measures are abbreviated as in Table 2. The percentage of nucleotides that are G or C is abbreviated as "%GC". Sampling site is the major distinction among metagenomes from natural surface water versus water collected from a pipe (PDS). Among surface water samples, clustering reflects samples' collection date, water chemistry and land use.

Though the PDS site is similar to the others in this study in terms of the water quality parameters measured (Figure 1), it stands apart in terms of its metagenomes' k-mer compositions (Figure 2A). This is likely due to how different this sampling site is from the rest in this study: PDS samples were collected from a reservoir after passing through a 9-km pipe, while all other samples were collected from near-surface water from rivers. The microbiota of these samples were likely significantly affected by both residency in the reservoir (Crump *et al.*, 2012) and transmission through the pipe (Fisher *et al.*, 2015). The clustering pattern described above shows that when metagenomes come from very distinct

lotic freshwater environments, this difference can be reflected in k-mer abundances. When samples are more similar; however, such as surface water from rivers, k-mer clustering is not simply reflective of sampling site or watershed. In this case, metagenomes cluster by water properties even when collected from unconnected watersheds 130 km apart. This suggests that there is a signal in the k-mer data that can distinguish between unaffected samples (AUP, PUP) and agriculturally affected samples (APL, ADS), but not between unaffected samples and all urban samples (UPL, UDS). This may be because of variability in the impact of urban land use on the rivers. In order to interpret the k-mer clusters in the context of the environmental data collected, we plotted the k-mer distance matrix using NMDS and fit environmental and metadata vectors onto this ordination using envfit (Figure 2B). Because k-mer abundances are based on DNA sequence, one major driver of k-mer abundance differences can be nucleotide bias. From this figure, we can see that both %GC and AGS tend to have higher values in Clusters 3, 4, and 5 than in Clusters 1 and 2. Nucleotide bias in bacteria has been associated with genome size either as a direct correlation (Musto *et al.*, 2006) or as a more complex association (Mitchell, 2007; Guo *et al.*, 2009). Nucleotide bias has also been shown to be dependent on environmental conditions, independent of shifts in phylogenetic composition (Reichenberger *et al.*, 2015). Here, we see a strong positive linear correlation between AGS and %GC (Pearson's $r = 0.8$, $p < 2.2e\text{-}16$).

All agriculturally affected samples are split between two k-mer clusters, which correspond to their sampling period. At these sites, we see two main periods in the year instead of four seasons: rainier "winters" and dryer "summers", both extending into spring and fall. This is consistent with the general climate of the region. The drier "summer" months were May to October, with an average of 4 ± 4 mm cumulative rainfall over three days before sampling, and the "winter" months were November to April, with an average of 43 ± 19 mm cumulative rainfall (t-test significant difference $p = 6e\text{-}08$). Cluster 2 contains samples collected during the drier "summer" months (May to October) and Cluster 3 contains samples collected during the rainier "winter" months (November to April), as well as seven samples from the urban sites collected from May to September. The samples in Cluster 3 tend to have higher values of specific conductivity, pH, turbidity, and cell counts and higher concentrations of orthophosphate, nitrate, nitrite, ammonia, chloride and chlorophyll *a*. Clusters 4 & 5 are associated with higher dissolved oxygen and contain almost all samples (24/25) from the unaffected sites (AUP and PUP) and no samples from the agriculturally affected sites (APL and ADS).

The urban samples are spread among clusters with unaffected (Clusters 4 and 5) and agriculturally affected samples (Cluster 3). The samples that cluster with the high-nutrient, “winter” agriculturally affected samples (Cluster 3) have significantly higher concentrations of orthophosphate than the urban samples from the other two clusters (ANOVA, p = 5e-05, q = 0.0003, means: 0.017 ± 0.005, 0.0082 ± 0.002, 0.0082 ± 0.003, for urban samples in Clusters 3, 4, and 5, respectively). These concentrations are lower than the agriculturally affected samples from this cluster (mean 0.2 ± 0.12), but are close to the water quality guideline limits (0.01-0.02 mg/L depending on season) (Canadian Council of Ministers of the Environment, 2007), indicating that this difference in concentration could affect aquatic life. This suggests that the concentration of orthophosphate may have a consistent effect on lotic bacterial communities across watersheds and land use, as previously observed within watersheds (Wang *et al.*, 2015b; Sterner *et al.*, 2004) and among photosynthetic biota across habitats (Elser *et al.*, 2007), and that this effect can be detected based on reference-free metagenome analysis.

This clustering technique also highlights potentially unusual samples that do not cluster according to the major trends described above, such as the AUP sample from September, which is the only AUP sample not in Clusters 4 or 5, and the PDS sample from October, which is the only PDS sample not in cluster 1. Because these samples are very similar to the other samples from their sites in terms of environmental conditions and cell count (Figure 1), it suggests that the samples may have been mixed-up or mislabelled during sample processing. Though we cannot be completely certain that these samples are compromised, this possibility is further supported by the unusual AGS of these samples described in the next section; hence, these samples are not included in further analyses.

These results show that in some cases, metagenomes from different sampling sites are distinct despite changing conditions, while in other cases, k-mer clustering reflects variability in samples’ water chemistry and environmental conditions across sampling sites. This analysis demonstrates that there is a bacterial signal that can distinguish between samples collected from an area with agricultural activity versus unaffected samples, but that this signal differs by time of year. Further sampling across multiple years would be required to assess whether this trend is seasonal. While k-mer profiles themselves do not directly translate to efficient water quality tests, this analysis supports that sampling over a year is important in some cases for the development of water quality tests based on bacterial communities. This k-mer analysis and is a reference-independent method that reveals that there are land-use, water chemistry, and environmental condition signals in this bacterial metagenomic data.

2.5.2. Average genome size (AGS) varies with daylight hours in the agricultural watershed – illustrating the importance of normalisation strategies

When AGS was tested for correlations with environmental variables from each site (excluding PDS), significant relationships were seen in the agricultural watershed (Table 3). At all sites in the agricultural watershed, AGS was significantly negatively correlated with daylight hours. This trend was also seen in the protected site, though was not significant after correcting for multiple testing (PUP: r = -0.59; p = 0.04; q = 0.15) and was not seen in the urban sites (Figure 3). These differences among sites may be due to bacterial community differences and/or differences in the effect of sunlight due to variability in the penetrance of light into the water and shade cover (Table 1). In the agriculturally affected sites, AGS was also significantly correlated with water temperature, rainfall, and turbidity. The increase in rainfall in both agriculturally affected sites is correlated with increases in turbidity (r = 0.68, 0.63; p = 0.007, 0.02 for APL and ADS, respectively), consistent with increased rainfall creating runoff from adjacent land, largely agricultural fields. Other potential indicators of runoff from agricultural activity, such as elevated concentrations of orthophosphate, ammonia, dissolved chloride and nitrite are also correlated with AGS (Table 3). These relationships indicate that in the agricultural watershed, as day length decreases, AGS increases, and that seasonal rain-associated water chemistry changes also correlate with this change in AGS. Due to the many possible indirect impacts of varying day length, further study would be required to identify the specific environmental drivers of AGS variation. Further sampling would also be required to determine whether this trend is seasonal across years and generalises to other geographic regions. If it does, this further demonstrates the importance of sampling across time when studying water quality, such as to develop new tests, as signals indicative of agricultural impact may vary substantially over time.

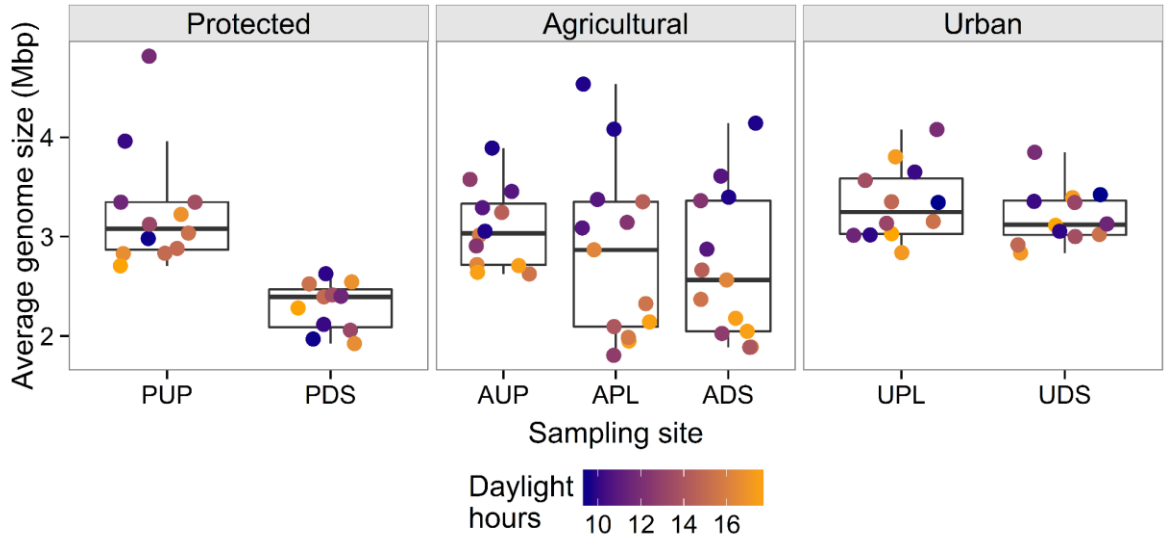

Figure 3 **Average genome size across sampling sites, coloured by daylight hours, illustrating correlations within the agricultural watershed.**

Points represent samples and are jittered on the x axis for visibility. The agriculturally affected sites (APL, ADS) have the largest ranges of values. There is a significant negative correlation between AGS and daylight hours in all sites in the agricultural watershed and a corresponding trend in PUP.

Table 3 **Environmental conditions significantly correlated with average genome size (AGS) over a year of monthly sampling within sampling sites.**

Environmental variable	Site	Spearman's *rho*	*p* value	*q* value (FDR corrected)
Ammonia	ADS	0.63	0.021	0.089
Dissolved Chloride	APL	-0.66	0.015	0.073
Dissolved Chloride	ADS	-0.75	0.0031	0.03
Nitrite	ADS	0.66	0.015	0.073
pH	APL	-0.8	9.70E-04	0.013
pH	ADS	-0.89	5.60E-05	0.0015
Orthophosphate	APL	0.84	3.20E-04	0.0062
Orthophosphate	ADS	0.64	0.019	0.088
Rainfall	APL	0.78	0.0015	0.017
Rainfall	ADS	0.69	0.009	0.058
Specific Conductivity	APL	-0.73	0.0049	0.038
Water Temperature	APL	-0.82	5.30E-04	0.0083
Water Temperature	ADS	-0.93	3.00E-06	2.30E-04
Daylight hours	AUP	-0.76	0.0041	0.036
Daylight hours	APL	-0.68	0.011	0.059
Daylight hours	ADS	-0.7	0.0082	0.058
Turbidity	APL	0.68	0.01	0.059
Turbidity	ADS	0.9	2.60E-05	0.001

Considering AGS further informs the identification of unusual samples. Both samples identified as unusual by k-mer composition clustering also have AGS values that are distinct from the other samples from their sampling sites. The September AUP sample has an AGS of 2.0 Mbp compared to the other AUP samples, which have a median of 3.0 ± 0.5 Mbp. The October sample from PDS has a high AGS of 3.7 Mbp compared to the site's median of 2.4 ± 0.4 Mbp. The other October sample from the protected watershed (from PUP) also has an unusually high AGS (4.8 Mbp) compared with the other samples from that site (median = 3.1 ± 0.5 Mbp). This shared trend between the October PUP and PDS samples is unusual because they were collected four hours apart from sites separated by a reservoir and a 9 km long pipe. This may indicate that a system-wide change occurred that introduced higher-AGS bacteria, such as an extreme runoff event, or that these samples were similarly contaminated during or after sample collection.

The relationship between AGS and environmental conditions emphasizes the importance of considering AGS when analysing metagenomic data. Beyond the direct biological relevance of AGS variability over time, this variation also indirectly affects gene functional group abundance profiles (Frank and Sørensen, 2011; Beszteri *et al.*, 2010). This is because with lower AGS, universal, single-copy genes make up a larger proportion of functional genes, and vice-versa. This causes physically real yet biologically uninteresting variability in the abundance of universal, single-copy genes among samples with varying AGS, and introduces spurious correlations (Beszteri *et al.*, 2010). In cases, such as the one seen here, where there is not only a large range of AGS values but also a strong relationship between AGS and environmental conditions, normalising by AGS is important to avoid biased data interpretation. Previous studies have shown that AGS varies among environmental sampling sites (Angly *et al.*, 2009; Frank and Sørensen, 2011) and within hosts (Nayfach and Pollard, 2015). To our knowledge, this work is the first to have tested for AGS temporal variation within sampling sites in environmental microbiomes. If not corrected for, this variation has the potential to obscure relationships between bacterial communities and environmental conditions, as described in the following section.

2.5.3. Common normalisation strategies can result in contradictory interpretations of metagenomic data

To describe bacterial gene functional groups associated with land use, translated reads were compared to protein sequences with functional annotations and assigned functions based on sequence similarity. To compare the abundance of functional gene groups across samples, abundances must be normalised by "sequencing effort" (Chistoserdovai, 2010). Common normalisation schemes involve one or more of the

following components: (1) subsampling the data such that each sample has the same number of reads before assigning reads to functional groups, (2) dividing the number of reads assigned to each functional group by the total number of reads in a sample, (3) dividing the number of reads assigned to each functional group by the total number of reads assigned to all functional groups in a sample, that is, to normalise by the percentage of reads assigned, and (4) scaling by the number of cells represented by the reads. This last approach can involve multiplying by the average genome size (AGS), which is estimated based on the abundance of multiple single copy genes (Angly *et al.*, 2009; Nayfach and Pollard, 2015; Manor and Borenstein, 2015) or dividing by the abundance of a single copy gene. Normalisation components 1 and 2 tend to give similar results except when considering low abundance groups, when they can result in the misinterpretation of zero counts, while components 3 and 4 can change results drastically (Figure 4, Appendix A, Figure A1).

In general, normalising by the percentage of reads assigned is most commonly applied; however this can lead to biases due to variation in read "mappability" (how likely a read will be similar to a reference database and classified) (Manor and Borenstein, 2015). A read's mappability to functional annotation databases can vary with technical differences, such as read-length, or with biological differences. For example, a read might not be assigned a function because it came from DNA that has an unknown function, that has diverged too much relative to reference sequences, or that is non-functional. If biological differences among metagenomes resulted in different percentages of reads assigned, then normalising by total assigned reads per sample masks a real change in gene proportions. For example, this is likely to be the case if a community undergoes a shift from more well-characterised bacteria to more poorly-characterised bacteria, when more and fewer reads will be functionally assigned, respectively. As certain phylogenetic branches of bacteria are better characterized than others, normalising by the percentage of reads assigned can introduce bias.

Another biological factor that can affect the percentage of reads assigned is a change in AGS. In general, essential, core genes make up higher proportions of smaller genomes and are more likely to have a close homolog in the reference database, while larger genomes are more likely to contain more specialised genes that are less likely to have functionally characterised reference sequences (Raes *et al.*, 2007; Nayfach and Pollard, 2015). This relationship is seen in the agriculturally affected sites (APL & ADS), where there is large variation in the percentage of reads assigned (Figure 4B) that is significantly negatively correlated with AGS (r = -0.74, -0.82; p = 0.005, 0.0009 for APL and ADS,

respectively). This indicates that a biological shift has occurred and that normalising functional profiles by the percentage of reads assigned would introduce bias. This relationship is not seen in the other sites, possibly due to the smaller ranges of AGSs or an uncharacterised confounding biological relationship.

Data normalisation can lead to contradictory results. To illustrate this effect, we compared the abundance of level-two SEED functional groups between samples from the agriculturally affected sites (APL & ADS) collected in the “summer” period versus the “winter” period (Figure 1). Data was normalised in one of four ways: (1) only by even subsampling or by even subsampling followed by normalising by: (2) the percentage of reads assigned, (3) AGS, or (4) the percentage of reads assigned and AGS. Out of 82 groups tested, 28 have differential abundances under all normalisations, 3 have differential abundances under only one normalisation scheme, and 51 have differential abundances under two or three normalisation schemes. Of those 28 functional groups with different abundances under all normalisations, 13 have opposite trends depending on the normalisation used. For example, when abundance profiles are normalised by subsampling and AGS, the “Pathogenicity islands” functional category is more abundant in the rainy “winter” samples than the dry “summer” samples (fold change between medians = 1.1, p = 0.02, q = 0.03). When normalised in any of the other ways listed above, the relationship was opposite (fold changes = -1.4, -1.1, -1.1; p = 8e-5, 0.05, 0.02, 0.05; q = 0.0003, 0.08, 0.03, 0.08, for methods 1, 2 and 4 listed above, respectively). Though the scale of these differences is small, they are statistically significant and, in this example, directly relevant to water quality assessment. If this category of “Pathogenicity island” genes was targeted as a source of biomarkers in the development of a new water quality test, the choice of normalisation scheme could directly affect whether a group of samples were considered high or low quality.

In theory, differential groups with larger effect sizes should be more robust to varying normalisation methods. We see that here, where the groups with significant differences that agree across normalisation methods have a greater average fold change than those that do not agree between normalisations (2.2 ± 0.6 versus 1.2 ± 0.2, respectively, Wilcox test p < 2.2e-16). This demonstrates that when looking for patterns in the abundances of functional groups among samples, the most conservative approach is to look for trends that are robust to normalisation method. Given the potentially extreme differences that normalisation methods can produce, the decision of which normalisation steps to take should be chosen carefully and stated explicitly.

2.5.4. Severity of contamination in an agriculturally affected watershed is reflected in gene functional group abundances across sampling sites

The Canadian Council of Ministers of the Environment (CCME) Water Quality Index (WQI) is a framework to evaluate surface water quality for the protection of aquatic life (Canadian Council of Ministers of the Environment, 2007). All samples from the sites not affected by agricultural pollution had a “good” or “excellent” water quality rating based on guidelines for ammonia, chloride, dissolved oxygen, nitrate, pH and orthophosphate. This includes the urban samples, indicating that either (1) land use did not have as large an impact on these samples as it did on the agriculturally affected samples, or (2) that this index formulation is not appropriate to assess their water quality. In the agriculturally affected sites, samples from the drier months (May to October) had CCME WQI rating of “fair” or “good” quality, while samples from the wetter months (November to April) all had a “marginal” or “poor” rating, due to high orthophosphate and/or low dissolved oxygen concentrations.

A similar classification was seen when samples from the agricultural watershed were clustered based on water chemistry and biological measures indicative of contamination (Figure 4). Three high-level clusters are significant (bootstrap value > 90%); Cluster 1 comprised samples from the unaffected site, while Clusters 2 and 3 contained a mix of samples from both affected sites. Consistent with k-mer clustering patterns and WQI ratings, Cluster 2 is mostly composed of samples collected in the drier months (May to October), while Cluster 3 is mostly from samples collected in the rainier months (November to April). The only exception to this pattern is the December sample from APL, which has very low values of ammonia, nitrate and nitrite and is grouped in Cluster 2. The two samples with lower rainfall in Cluster 3 were collected in early February, when rainfall was below seasonal averages. The rainy season samples from the agriculturally affected sites (Cluster 3) are associated with elevated turbidity and higher concentrations of nitrate, nitrite, ammonia and orthophosphate, consistent with increased runoff from the surrounding agricultural land (Figure 4).

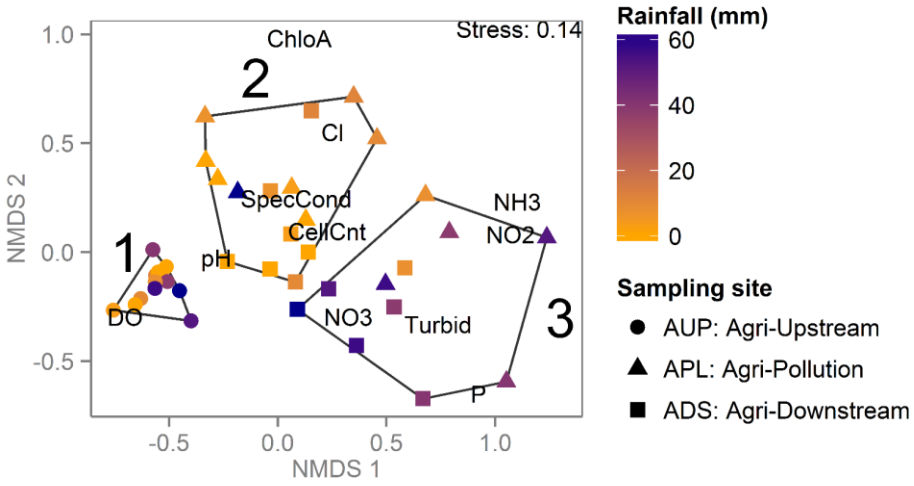

Figure 4 **Agricultural watershed samples clustered by water chemistry reveal impact of land use and rainfall.**

NMDS plot based on environmental and chemical water measurements for samples from the agricultural watershed. Each point represents a sample, coloured by cumulative rainfall over three days prior to sampling and shaped by sampling site. Significant clusters are outlined and numbered in black. Samples collected upstream of agricultural activity (Cluster 1) have higher DO levels. Samples collected in the summer from the agriculturally affected sites (Cluster 2) have higher chlorophyll *a* concentration, while the winter samples are more affected by runoff, as indicated by higher nutrient levels and turbidity (Cluster 3).

This increased runoff and agricultural impact is reflected in the metagenomic data when SEED subsystem abundances are compared between the more affected samples (Cluster 3) and the less affected samples (Clusters 1 & 2). When abundances are normalised by even subsampling and AGS, 191 groups are significantly more abundant in the more affected samples (Supplementary Table 1). When only considering the 11 groups with larger effect sizes (fold change between medians >= 2.5) (Table 4), these groups are significantly different under all normalisation conditions (normalised by subsampling only, by the percentage of reads assigned, and by the percentage of reads assigned and AGS). When considering more subtle differences (fold change < 2.5), 93 groups are only predicted as significantly different when normalised by even subsampling and AGS. In general, we see a higher abundance of subsystems associated with nutrient metabolism (carbohydrates, nitrogen, and proteins), respiration, and phage in those samples that had increased runoff from agricultural land (Cluster 3) (Table 4, Supplementary Table 1). The higher abundance of nutrient-metabolism subsystems is reasonable, given the higher concentrations of nitrate, nitrite, ammonia, and orthophosphate (Figure 4). As this data describes DNA extracted from material collected on 0.2 µm pore-size filters, most of the phage sequences are likely associated with cells as prophage, replicating phage, or phage particles attached to the surface of cells. Further study of bacterial taxonomic profiles and these specific gene sequences may provide insight into whether this SEED subsystem is more abundant due to

changes in host abundance or viral population dynamics. This analysis is limited by the biases inherent in how metagenomic reads are classified, in gene lengths, and in the SEED classification structure. However, the differential subsystems with larger effect sizes indicate that there are significant differences in the genes present between water samples with lower water quality, collected during a period of increased runoff from agricultural land, and samples with higher water quality, collected from the same agricultural area but during a dryer period.

Table 4 Differentially abundant gene functional groups between samples with higher and lower water quality in the agricultural watershed.

Differential SEED Subsystem	SEED Class (Subclass)	*q* value [1]	Fold change[2]
Malonate decarboxylase	Carbohydrates (Organic acids)	0.0045	4.6
Nitrosative stress	Nitrogen Metabolism (No subclass)	0.0045	3.7
Denitrification	Nitrogen Metabolism (No subclass)	0.0065	3.5
Phage capsid proteins	Phages, Prophages, Transposable elements (Bacteriophage structural proteins)	0.0045	3.5
Na(+)-translocating NADH-quinone oxidoreductase and rnf-like group of electron transport complexes	Respiration (Electron donating reactions)	0.0074	2.9
Lysine degradation	Amino Acids and Derivatives (Lysine, threonine, methionine, and cysteine)	0.012	2.8
Pyruvate:ferredoxin oxidoreductase	Carbohydrates (Central carbohydrate metabolism)	0.0065	2.8
Bacterial hemoglobins	Stress Response (No subclass)	0.013	2.7
D-galactarate, D-glucarate and D-glycerate catabolism	Carbohydrates (Monosaccharides)	0.0065	2.6
D-galactonate catabolism	Carbohydrates (Monosaccharides)	0.0098	2.6
Pyrimidine utilization	Nucleosides and Nucleotides (Pyrimidines)	0.0065	2.6
RNA 3' terminal phosphate cyclase	RNA Metabolism (RNA processing and modification)	0.0065	2.5

Lower-quality water samples are more affected by agricultural runoff than the higher-quality samples (Cluster 3 versus 1 and 2 in Figure 4). Differential functional groups (SEED subsystems) with fold change greater than 2.5 are listed and sorted by fold change. Abundances are normalised by AGS. Most of the subsystems that are more abundant in the lower-quality water are responsible for metabolism of nutrients, consistent with the water's higher concentrations of nitrate, ammonia and orthophosphate.

[1] Corrected for FDR

[2] Median abundance in lower quality water divided by median abundance in higher quality water

2.5.5. Gene group proportions normalised by percentage of reads assigned are stable across time and watershed but differ from previous studies

When SEED level 2 subsystem abundances are normalised across samples by the percentage of reads assigned to any subsystem, the median standard deviation of abundances is 0.2±0.1% among the 35 subsystems with at least 1% abundance, which corresponds to a 10±6% median relative standard deviation (i.e. standard deviation calculated as relative to subsystem abundance). This low variability when normalising by percentage of reads assigned was also seen in the only other metagenomic study that has looked at gene family composition in river water across land use (Staley *et al.*, 2014b). In this study, high-level KEGG pathway category abundances were remarkably stable in the Upper Mississippi across varying land use (forested, urban and agricultural). Across 11 sites they observed a maximum standard deviation of 0.5% among KEGG level 2 groups, with a median standard deviation of 0.02%. However, this study only sampled each site once (between May to July), whereas we sampled monthly over a year, and they used MG-RAST to assign reads to functional groups, whereas we used MEGAN.

In order to better compare our results against this Upper Mississippi study, we analysed our data one month at a time using their methods. Briefly, we used the MG-RAST pipeline with KEGG database annotations and normalized by the percentage of reads assigned, without adjustments for AGS. This analysis excluded the non-surface sampling site, PDS, and is limited by the normalisation choices and the accuracy of MG-RAST. To match the single time point samples of the Upper Mississippi study, we examined our May to July samples by month. In this analysis, we saw higher variability within KEGG level 2 groups, with a maximum standard deviation of 2% and a median standard deviation of 0.2%, versus 0.5% and 0.02% in the Upper Mississippi study, respectively. This higher variability in our study when looking over the same time period may indicate that the land use differences between sites in this study are greater than those in the Upper Mississippi study or may be due to inherent differences in the watersheds. In both cases, however, when looking at functional assignments as a proportion of all reads assigned a function (i.e. when normalised by the percentage of reads assigned) the variability in abundance of these high-level functional groups is low compared to their abundance values. This demonstrates how normalising by the percentage of reads assigned and failing to normalise by AGS can mask variability among samples. When the patterns of these potential normalisation factors are considered, there is clear variability among samples (Figure 5).

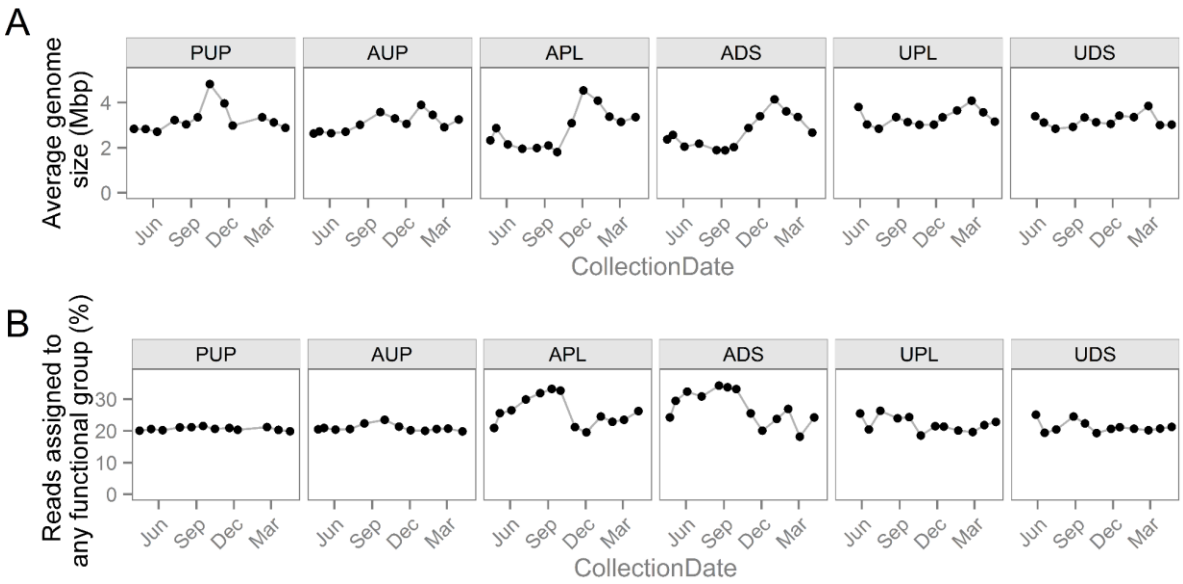

Figure 5 Normalisation factors used in different strategies to enable comparisons of gene group abundances between samples.

Notable differences exist among sites and over time in both **(A)** average genome size and **(B)** the percentage of reads assigned to any gene functional group. This variability demonstrates how normalisation schemes using these values can have a drastic effect on downstream analyses.

Though the variability among high-level functional groups analysed this way was similar between our study and the Upper Mississippi study, notable differences were observed between which groups were most abundant. According to this MG-RAST analysis performed for comparative purposes and lacking AGS normalisation, the most abundant KEGG level 2 groups across all sites in our data were “Amino Acid Metabolism” (20±2%), “Carbohydrate Metabolism” (12±0.9%) and “Membrane Transport” (11±1%), while in the Upper Mississippi study, the most abundant categories were “Membrane Transport” (21±0.3%), “Carbohydrate Metabolism” (11±0.1%) and “Signaling molecules and interaction” (11±0.3%). The differences in abundance of “Amino Acid Metabolism” and “Signaling molecules and interaction pathways” were especially pronounced, with 20% versus < 5% and < 0.06% versus 11% abundance in our study versus the Upper Mississippi study, respectively. These differences in dominant functional groups may be due to technical differences, such as our longer read lengths and different filtration system, or also may suggest that functional profiles differ over large distances. This highlights the inherent difficulty in comparing metagenomics analyses across different studies at present and at minimum the need to consider methodology variation. Regardless, these considerable inter-study differences are notable since there is so little within-study variation when analysed in this way. Our results from three watersheds that were all measured using the same methodology, but are up to 130 km apart and under differing land use suggest that, among functionally assigned reads not adjusted for AGS, general functional profiles are fairly stable

at a regional scale. Among AGS-adjusted profiles not calculated as proportions of functionally assigned reads, however, we do see more variation, both between sites and within sites over time. Further studies, in which variability in AGS and in the proportions of reads assigned are considered, are required to characterise the variability of river metagenomes at larger geographic scales.

2.6. Conclusion

This work presents the first year-long survey of bacterial metagenomes from water samples collected across protected, agricultural, and urban watersheds. It is also the first to report metagenome gene functional group differences associated with land use across time in lotic microbiomes. We have shown that fundamental metagenome characteristics such as k-mer composition and average genome size (AGS) vary with time and land use. Sampling site can be the major discriminative factor in metagenome k-mer composition when site characteristics are very different (i.e. water collected from a reservoir through a pipe versus surface water); however, among samples of surface water, metagenomes instead clustered by water chemistry, even when collected from unconnected watersheds 130 km apart. AGS is correlated with hours of daylight in all sites in the agricultural watershed. Beyond its ecological relevance, this finding also demonstrates the importance of normalising functional profiles by AGS, since this variation could confound relationships between metagenomes and environmental conditions. Further bias can be introduced by the common practice of normalising gene prediction abundances by the number of predictions made. When comparing samples with differing water quality, many gene functional groups with differential abundance were observed and those with the largest effect sizes were robust to normalisation method. However, when considering more subtle effects, normalisation strategy can have a large impact on both the interpretation of gene functional group profiles and also the identification of biomarkers. This study has shown that metagenome characteristics and gene function content change with time and land use in lotic bacterial communities. Future studies may build on the results presented here either directly, through the identification of candidate biomarkers from the sequence data presented, or indirectly, by using the findings reported as a reference in the design of future studies of freshwater ecosystem health.

2.7. Author Contributions

TV led the bioinformatics analyses and wrote the manuscript. MP contributed to analyses and revised the manuscript. MU led the sampling and DNA sequencing, with help

from MC and KC. JS and MN performed size selection of sequencing libraries. CS and WH guided the analyses and aided in interpretations. PT, NP and FB designed the project, guided the analyses and aided in interpretations, with FB also providing lead guidance in manuscript revisions. All authors contributed to final revisions of the manuscript.

2.8. Conflict of Interest Statement

JS and MN both hold shares in Coastal Genomics, a privately-owned company that provided the Ranger Technology used in this study. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

2.9. Funding

This work was funded by Genome BC and Genome Canada grant No. LSARP-165WAT, with major support from the Simon Fraser University Community Trust Endowment Fund and additional support from the Public Health Agency of Canada. TV was supported by NSERC PGSM & CGSD scholarships. MP was supported by a CIHR/MSFHR Bioinformatics Training Program fellowship and an NSERC PGSD scholarship. MU was supported by a Mitacs Accelerate fellowship.

2.10. Acknowledgements

We would like to thank Judith Isaac-Renton of the BC Public Health Microbiology and Reference Laboratory for her expertise and guidance. We also thank Matthew Croxen of the BC Public Health Microbiology and Reference Laboratory, Anamaria Crisan of the BC Centre for Disease Control, Marli Vlok and Alvin Tian of UBC, and Raymond Lo of SFU for their helpful discussions and other input. We would also like to acknowledge Jan F. Finke of UBC for conducting the flow cytometry measurements.

Chapter 3.

Identification of potential biomarkers from riverine bacterial metagenomes

3.1. Foreword

This work was performed within the GC Watershed Project (see section 1.2). Design of the sampling scheme and production of the data was a large collaboration (see author contributions in section 2.7). My role in this project was to perform the metagenomic data preparation, analysis, interpretation, and reporting and aid in optimisation of DNA sequencing. This chapter provides an example of the process for identification of biomarkers and the design of qPCR tests for their detection. The novelty of my approach was to leverage existing taxonomic marker databases, use machine learning techniques for differential feature identification, and use protein clusters as biomarkers. Miguel Uyaguari Diaz led sample collection (see Author Contributions in Author Contributions 2.7) and performed the qPCR amplifications.

Because my expertise is more in analytical methods to identify biomarkers than their biological explanation, I focus my work in this chapter on the former. This includes methods for identification from metagenomic data of features that are differentially abundant between groups of samples and then the development and initial testing of qPCR-based assays for these markers.

3.2. Abstract

A common approach used to monitor water quality is to evaluate it at the location where human use occurs, such as at the tap or at a beach. Often, these tests aim to detect fecal contamination by testing for the presence of fecal coliform bacteria using selective culturing conditions. These tests are limited by false negative results (water contaminated by other pathogens), false positive results (tests are not specific to pathogenic fecal coliforms). Further, they target human health, which fails to monitor for the broader health of aquatic ecosystems. As part of a national Genome Canada funded project to improve water quality monitoring, we investigated the use of metagenomics to identify new targets as indicators of water quality. Monthly samples were collected for a year from seven sites in three watersheds with either agricultural, urban, or minimal land use. In the agricultural watershed,

samples were collected both upstream and downstream of agricultural activity. For each sample, bacterial metagenomes were shotgun sequenced. From this data, genes, taxa and sequence clusters ("features") were identified that were associated with agricultural land use. For each feature, a quantitative polymerase chain reaction (qPCR) test was designed computationally and tested *in vitro*. In general, the qPCR tests designed using these methods performed as expected. Assessment of the generalisability of these initially successful primers to identify agriculturally affected water from other locations and other years is currently underway. This work demonstrates effective approaches to design qPCR assays for biomarkers from freshwater river metagenomes.

3.3. Introduction

"Biomarker" is a shortening of the phrase "biological marker". Typically, this refers to a biological feature, such as a gene, taxon, or metabolite, whose presence or abundance suggests some condition, such as contamination or a disease state. Biomarkers are often used to monitor for and diagnose changes in environmental and human health. For example, the abundance of fecal coliform bacteria in recreational water is often used as a biomarker for the risk to human health associated with exposure (Field and Samadpour, 2007). Here, fecal coliform bacteria are used as an indicator of the likelihood of the presence of feces-associated pathogens.

Metagenomic sequencing techniques can generate a comprehensive profile of the genes and taxa ("features") present in a sample, which can provide a rich catalogue of potential biomarkers. Because these features were identified from DNA data, metagenomic data also directly supports the design of DNA-based assays, such as qPCR. At the outset of this work, designing qPCR biomarkers from metagenomic data was a new process without well developed methods and with uncertain likelihoods for success. Since then, targeted approaches have been developed to identify differential features (e.g. LEfSe (Segata *et al.*, 2011) and a "robust PCA"-based method (Alshawaqfeh *et al.*, 2017)) and some successful qPCR assays have been developed from metagenomic data. For example, in health research, colorectal cancer biomarkers have been identified from metagenomic taxon profiles (Liang *et al.*, 2017; Yu *et al.*, 2017) and patented (Qiang *et al.*, 2016).

One of the goals of the GC Watershed Project was to identify better biomarkers of water quality. The work presented here represents a first stage of the development of such biomarkers to trial the process of developing qPCR based biomarkers from riverine metagenomic data. The goal of the biomarkers developed here is to distinguish water samples collected from sites with minimal land use versus samples collected from sites with

intense agricultural activity in their catchments. While future work may focus on developing biomarkers associated with varying conditions such as water quality index scores or nutrient concentrations, the work presented below begins with a simpler categorical classification as a first pass. The experimental design, data collection and feature profiling are described in Chapter 2. This chapter will report biomarker assay development including target feature selection, assay design, and assay technical validation (on original study samples).

3.4. Methods

3.4.1. Biomarker candidate DNA sequence identification

Primers were designed to amplify more DNA from samples collected from the agriculturally affected sites (APL & ADS, “expected high”) than from the samples collected from sites with low land use (PUP & AUP, “expected low”). Primers were not designed with consideration for the urban-affected sites (UPL & UDS, “not predicted”).

3.4.1.1. Gene family based

Shotgun sequenced reads were trimmed to remove low quality bases using Trimmomatic (Bolger *et al.*, 2014). A sliding window of length 5 and a minimum Phred score of 20 was used at the 3’ end. At the 5’ end, sequences of one or more nucleotides with scores less than 20 were trimmed. Sequencing adapters were removed using Cutadapt (Martin, 2011), overlapping paired-end reads were merged using PEAR (Zhang *et al.*, 2014), and reads shorter than 100 bp were discarded. After this processing, samples had 418538 to 2165162 high quality reads and samples were subsampled to the number of reads in the smallest sample: 418500 reads. The January sample from site PUP was omitted from all analyses due to too few reads.

Reads were compared against NCBI’s nr database (downloaded April 9, 2014) using RAPSearch2 (Zhao *et al.*, 2012). Resulting protein alignments longer than 30 amino acids were analysed using HUMAnN (Abubucker *et al.*, 2012) or MEGAN version 5.10 (Huson *et al.*, 2011) with default parameters, including a minimum bit score of 35 and a maximum e-value of 0.01, to determine the gene families present. Sequences were annotated with predicted functions using the SEED (Overbeek *et al.*, 2005), KEGG (Kanehisa, 2000), and COG (Tatusov *et al.*, 2000) gene functional group databases. Gene groups were removed if they had low abundance (less than 0.01% of reads assigned to them, averaged across samples). Gene group abundances were then normalised in three ways: (1) only by even subsampling, or by even subsampling followed by normalising by: (2) average genome size,

or (3) DNA concentration. Gene groups were tested for differential abundance between samples from agriculturally affected and unaffected sites using the Mann-Whitney-Wilcoxon test with p values corrected for false discovery rate using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Significance test values were considered significant if q values were less than 0.05. Gene groups that had significantly different abundances between sets of samples under all normalisation schemes were selected for further analysis. These gene groups were used as the input to three Random Forest analyses, one for each normalisation method. Based on the Random Forest analysis results, gene groups were ranked by their mean decrease in accuracy and selected using a cut-off chosen based on plotting the mean decrease in accuracy. The eight gene groups that were selected in all three Random Forest analyses (intersection) were chosen for biomarker development. Reads that were assigned to each of these gene groups were pulled from the original data and assembled using Velvet (Zerbino and Birney, 2008). Reads were then mapped back to the assembly consensus sequences to check for lower-coverage regions, which were removed.

3.4.1.2. Taxon based

Quality trimmed reads were tested for similarity against a database of taxonomic marker DNA sequences to create a taxonomic profile using the MetaPhlAn (Segata *et al.*, 2012) software. This approach creates a profile by aligning reads against the MetaPhlAn (Segata *et al.*, 2012) database of taxon-specific marker sequences. Resultant taxon abundances were tested for differential abundance between samples from agriculturally affected and unaffected sites using the Mann-Whitney-Wilcoxon test with p values corrected for false discovery rate using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). For each taxon of interest, marker sequences were extracted from the database and reads were aligned against these references. Alignments with the most reads were visualised using Tablet to identify sections of the reference marker sequence that were well covered and consensus sequences were generated for these alignment regions. Similarity to this consensus sequence was then tested for reads from samples where there should and should not be matches to check that this new target sequence was as differential as the original taxon identified using MetaPhlAn (Segata *et al.*, 2012). If the number of reads similar to the target sequence was higher in the desired group of samples, then this sequence was considered a potential biomarker target sequence.

3.4.1.3. Protein cluster based

Reads were trimmed to remove low quality base calls and used as queries against NCBI's nr protein database as described in Chapter 2, section 2.4.3. From every read's top

hit, the translated query sequence (protein sequence) was extracted and all translated query sequences were clustered at 70% and 90% identity using cd-hit (Fu *et al.*, 2012). To identify biomarkers that would be distinctive between two sets of samples (i.e. samples collected from sites with low versus agricultural land use), clusters were evaluated based on the number of sequences per cluster from each set of samples. For clusters with discriminative abundance between sets of samples, all the original DNA reads were pulled and common motifs were identified using MEME (Bailey *et al.*, 2010). Motifs were chosen that were at least 18 bp long, so that the sequence would be long enough to support development of a PCR primer, and were most similar across the largest number of reads. For each motif, all reads containing the motif were pulled from the original dataset, assembled to obtain a consensus sequence, and aligned against this consensus sequence. Alignments were visually examined to assess depth of coverage and verify that the consensus sequence was not chimeric. The original motif was used as the probe for the qPCR assay and primers were based on the consensus sequence.

3.4.2. Design of qPCR assay to detect abundance of candidate biomarker target sequences

After potential biomarker target sequences were identified, they were compared against nr using BLASTx to check for any conserved domains that were not specific to the source biological feature (e.g. taxon or gene group). These regions were trimmed to avoid decreased specificity if biomarker was generalised beyond the original study samples. Primers and probes for qPCR were then designed for the consensus sequences using IDT[ii], primerBLAST (Ye *et al.*, 2012), or Primer3Plus (Untergasser *et al.*, 2012).

3.4.3. High-throughput multiplex qPCR

The text in this section (3.4.3) was written by Miguel Uyaguari Diaz, who performed the qPCR assays. It is included here with the author's permission.

> "Primers were first checked for sensitivity and specificity by SYBR green-based PCR. [These results are referred to as the "pre-test" results]. […] Assays with cycle threshold values greater than 35 were not included in the HT qPCR run using the BioMark system (Fluidigm Corporation, South San Francisco, CA). All probes used a 5' 6-FAM dye with an internal ZEN quencher and 3' Iowa Black fluorescent quencher (Life Technologies, Carlsbad, CA).

[ii] http://www.idtdna.com/scitools/Applications/RealTimePCR/

DNA extracts from Watershed Project samples were diluted 10-fold and 1.25 µl of DNA from each sample was pre-amplified with low concentrated primer pairs (0.2 µM) corresponding to all assays in a 5 µl reaction volume using TaqMan Preamp Master Mix (Life Technologies, Carlsbad, CA) according to the BioMark protocol (Fluidigm Corporation, South San Francisco, CA). Unincorporated primers were removed using ExoSAP-IT High-Throughput PCR Product Clean Up (MJS BioLynx Inc., Brockville, ON) and samples were diluted 1:5 in DNA Suspension Buffer (TEKnova, Hollister, CA).

The pre-amplified products were run on the BioMark system (Fluidigm Corporation, South San Francisco, CA) using 96.96 dynamic arrays. Five µl of 10x assay mix (9 µM primers and 2 µM probes) were loaded to assay inlet, while 5 µl of sample mix (2x TaqMan Mastermix (Life Technologies, Carlsbad, CA), 20x GE Sample Loading Reagent, nuclease-free water and 2.25 µl of pre-amplified DNA were loaded to each sample inlet of the array following manufacturer's recommendations. After mixing the assays and samples into the chip by an IFC controller HX (Fluidigm Corporation, South San Francisco, CA). Quantitative PCR was performed with the following conditions: 50 °C for 2 min, 95 °C for 10 min, followed by 40 cycles of 95 °C for 15s and 60 °C for 1 min. Samples were run in quadruplicates for all 92 environmental samples and genomic DNA from *E. coli* and *P. aeruginosa* were used as positive controls for the 16S rRNA gene [data not shown]."

The sample identified as likely mislabeled or contaminated in Chapter 2 was omitted from the analysis of results (AUP 47).

3.5. Results & Discussion

Candidate biomarker assays had variable results from the initial test to assess whether a single PCR product was produced ("pre-test"), though most primers designed from the metagenomic data worked as expected (Table 6). All primer sets that passed this pre-test tended to produce more qPCR product from the intended samples (agriculturally affected) than from the unintended samples (minimal land use) (Figure 7). In some cases, the results from the urban-affected samples were more similar to samples from the sites with minimal land use (GF.07, GF.08), while in other cases they were more similar to the agriculturally affected samples (PN2.1.3). The variability in the results from the positive control primer set (16S rRNA gene, Figure 7) reflects variability in the concentration of extracted DNA.

3.5.1. Gene families

Gene family candidate markers belonged to three function groups: nitrate reductase, atrazine degradation, and xylene degradation (Table 5). The primers designed based on reads classified as a nitrate reductase gene (NarG: nitrate reductase alpha subunit,

Molybdopterin oxidoreductase) produced a single product, as expected (GF.07 & GF.08, Table 6). These primer sets also had expected amplification results (Figure 7).

> template DNA sequence from NarG reads

TCACACCACTCATGCATGACACGCCTGCCGAGCTGGCACAAGCCTTTGATGTGAAAGA
GTGGAAAAAAGGCGAGTGCGAACTGATTCCAGGCAAAACCGCACCACAAATCTCAGTG
GTTGAGCGCGATTACCCTAACACCTACAAACGCTTCACCGCGCTTGGCCCTTTGATGG
AAAAAGCGGGCAACGGAGGCAAGGGTATTGGTTGGAATACCGCCGACGAGGTGCACC
AACTCGGGGAGCTCAATGGCAAGGTGCGTGCCGAAGGCGTGACCAAGGGCATGCCCA
ACATCAGCACCGACATCGATGCTTGCGAGGTGGTGCTGCAGCTGGCCCCCGAGACCA
ACGGCCATGTAGCCGTCAAGGCTTGGGAGGCACTCTCCAAAATCACCGGACGCGACC
ACACGCACCTGGCCTTGCACCGCGAAGATGAAAAAATTCGTTTTCGCGACATCCAGGC
CCAGCCGCGCAAGATCATCAGTTCTCCCACCTGGTCGGGCCTGGAGAGCGAAAAGGT
CTCGTATAACGCGGGCTACACGAATGTGCATGAGTACATTCCGTGGCGCACGCTCACT
GGGCGCCAGCAGTTCTATCTGGACCACCCTTGGATGCTGGCTTTTGGCGAAGGCATGA
CCAGCTACCGGCCACCCGTGGACCTGAAAACGGTGGACGAGATGATCGACCGCAAGC
CTAACGGTCACAAAGAGATTTCGCTCAACTTCATTACGCCGCACCAGAAGTGGGGCAT
CCACAGCACCTACAGTGACAACCTGCACATGCTCACGCTCAACCGGGGCGGCCCGGT
GATCTGGCTGAGCGAAGACGACGCCAAGAGTGCAGGCATCGTGGACAACGACTGGGT
CGAGCTGTTCAACAGCAATGGCGCGATCGCTGCCCGCGCCGTGGTCAGCCAGCGGGT
CAATCCTGGCATGGTCATGATGTACCACGCTCAGGAAAAAACCATCAACACGCCCGGC
TCGGAAATCACCGGCATCCGCGGCGGCATCCACAACTCAGTCACGCGCATCGTGACCA
AGCCGACCCACATGATTGGCGGCTACGCCCAGTTCAGCTACGGCTTCAACTACTACGG
AACCATTGGCACCAACCGCGATGAATTTGTTGTGGTGCGCAAGATGCGCAAGATCGAT
TGGCTCGATGGTGAAGCCCCTGCCACCGTGCAGGCttgaag

Atrazine degradation and xylene degradation gene families already had validated PCR primers published (Sherchan and Bachoon, 2011; Thompson *et al.*, 2010; Devers *et al.*, 2004; Baldwin *et al.*, 2003; Cébron *et al.*, 2008) so these were used in favour of developing new primers. However, the primers used from literature either did not produce products or produced multiple products of different sizes (Table 6).

Primers designed based on gene families were successful in cases when the primers were designed based on the reads specific to this study. Primers pulled from literature were unsuccessful, either due to a failure to amplify or non-specific amplification (Table 6). Literature based primers were selected in this case due to concerns about the depth of sequencing in this study and the potential benefit due to their higher likelihood for further generalisability if successful. The trends that can be drawn from this data for the likelihood of success for published versus custom primers is limited since these categories were confounded by a difference in the target gene families, i.e. these differences in amplification could also be due to differences in the accuracy of gene family abundance prediction from the metagenomic data.

3.5.2. Protein clusters

Two clusters were chosen to design primers from, one from clustering at 70% identity and one at 90% identity. The 70% identity cluster (number 184599) had common sequence motif: AATAAAGC[T]GTGACCGTAGT (with T in brackets being most common but inconsistent across reads). All reads that contained this exact sequence (203 reads, all from APL & ADS except for a few from UPL) were aligned and the consensus sequence was used to design primers (M3.1.4, Table 5).

> Protein cluster candidate sequence from cluster 184599 – M3.1.4
CCAATGGCGCTACGACTGCGAACGACAATAAGTACTACAGCGCCGATACCGTCACCTT
AGTGAAAACCGCTGCGACTGGTAGCTACGCGAATAAGAATGTGGGCACCAATAAAGCT
GTGACCGTAGTGGGCTACACCTTAAGTGGTACCGATGCCGGTAACTACACCGTCAGCG
ATGCTTCTGGTGCGACTGCCACGATTACCGCCAAAGCAATTACCGTCACTGGCACTACT
GCCAGCGATAAGG

Instead of using MEME to identify a probe for the 90% identity cluster (number 63525), Primer3Plus was used to directly design the probe and primers. The resultant primer set (PC.02) performed mostly as expected, with the exception of two fall samples from the pristine site (PUP) and two winter samples from the downstream agricultural site (ADS) (Figure 7).

> Protein cluster template sequence from cluster 63525 – PC.02
CTTCTAAAGTGACTTACGCCGCTGATGCATTCTTCGACTTCGACAAGTCGGTGCTCAAG
CCCGCTGGCAAAGCCAAGCTGGATGACTTGGTTGGCAAAGTCAAAGGCATCAACTTGG
AAGTSATCATYGCTGTGGGTCACACTGACTCTATCGGCTCTGACGCTTACAACCAAAAA
TTGTCAGTGCGTCGCGCTGAAGCTGTGAAGGCTTACTTGGTGTCTAAAGGCATCGAGA
AAAACCGCATCTACACAGAAGGTAAAGGCGAGAAGCAACCTGTTGCTAGCAACAAGAC
CAAAGAAGGTCGCGCTAAGAACCGTCGCGTTGAGATCGAAGTTGTGGGCACCCGCAA
GAACTGATCTGACYTGATCTAAAAAAGAAACCCCGCCTCGGCGGGGTTTCTTTTTGCTT
GAACGACAATACAARCCATGCAAAACACTGCCACKCACACCAACGTTGATCCGGCCGA
ATTGGCCAAGTTTTCAGACCTCGCCCACCGYTGGTGGGACCCYGAGAGCGAATTTCGT
CCGTTGCACCAAATCAATCCATTGCGCTTGGAGTGGATCAACGGCATYKCRYCSTTACA
AGGCAAAAAAGTGTTGGACGTGGGCTGTGG

3.5.3. Taxonomic marker gene based

Based on a marker-gene taxonomic profiling tool (MetaPhlAn), *Polynucleobacter necessarius* was more abundant in the agriculturally affected sites (APL, ADS) than in the sites with low human impact (AUP, PUP). The MetaPhlAn database reference sequence with the largest number of highly similar reads from the study dataset (313 reads with a mismatch rate of 3.1%) was sequence number 100131477 (648 bp long). According to a BLASTX similarity search against NCBI's "nr" database, this sequence is a near perfect

match (99-100% identity over 88% of the sequence) to the 50S ribosomal protein L10 from *Polynucleobacter necessarius* subsp. asymbioticus QLW-P1DMWA-1, *Polynucleobacter necessarius* subsp. *necessarius* STIR1, and beta proteobacterium CB, which is in the *Polynucleobacter* genus. When reads were aligned against this database reference sequence, only a subsection of the reference sequence was covered (Figure 6).

Figure 6 Alignment of reads against reference database sequence
The coloured sequence at the very top is the reference sequence, with colours representing different nucleotides. The blue bar plot below the reference sequence summarises the number of reads aligned at each nucleotide column. In the grey rectangle blow, each contiguous set of coloured squares is a read. Some rows have multiple reads. Colours represent different nucleotides. White represents a mismatch between the aligned read and the reference sequence. Note that there are columns where study sample reads consistently mismatch the reference database sequence and that no reads matched the 3' end of the reference sequence. Visualisation created in Tablet (Milne *et al.*, 2013).

The consensus sequence from this alignment of reads is:

>100131477_concensus

CTTTGAATGTACAAGACAAAAAAGCGATTGTTGCTGATGTCGGCGCTCAATTGGCTGGA
GCTCAAACAGTCGTGCTCGCTGAATACCGTGGTATTCCAGTAGAGCAGTTGACAAAGCT
ACGTGCTAGCGCACGTGACCAAGGTGTATATCTTCGCGTTTTGAAGAACACATTGGCAC
GCCGTGCTGCACAAGGCACACAGTTTGAGCCTCTTGCTGATTCGATGGTTGGCCCCTT
GATCTACGGCATCTCTGCTGATCCGATTGCTTCGGCAAAAGTATTGCAGAACTTTGCTA
AGACTCAAGACAAGCTAGTCATTACTGCTGGCTTATATAACGGCAAGTTGTTAGACGTA

GCGGGCGTTAAATCCCTCGCGTCTATTCCAAGCCGCGACGAGTTGTTATCTCAGTTGTT
GGGTGTCATGTTGGCCCCAGTATCTGCAATGGCTCGCGTATTGGGCGCAGTAGCAGCA
CAGAAAGCCGCAGGAGCACCTGCTCCAG

Two sets of primers and probes were designed from this sequence: IDT was used to create P.N.2.1.1 and Primer3Plus was used to create P.N.2.1.3 (Table 5). In the latter, the three columns of the alignment where the reads mismatched the reference sequence were included to try to increase the specificity of the PCR test.

Both primer sets produced products as expected (Table 6) but only P.N.2.1.3 was selected for further testing. The P.N.2.1.3 primers and probe were discriminative for the agriculturally affected samples versus those from sites with minimal land use (Figure 7). These primers also had the interesting effect of amplifying products in samples from the urban affected sites (especially UPL). These urban samples were not used in the biomarker design process, but are of interest in terms of water quality assessment.

Table 5 Candidate biomarker details

Biomarker basis	Marker name	Forward primer	Probe	Reverse primer	Primer source	Gene	Gene family	Tool	Taxon specificity
Gene family	GF.01	AATTCTATGACTGGCTGTTC	NA	CGCACAATACAACCTCAC	(Sherchan and Bachoon, 2011)	atzA	Atrazine degradation	HUMAnN	Unknown
	GF.02	GCACGGGCGTCAATTCTA	ATCGGATGGACGGGCGCA	CGCATTCCTTCAACTGTC	(Thompson *et al.*, 2010)	atzA	Atrazine degradation	HUMAnN	Unknown
	GF.03	GGGTCTCGAGGTTTGATTG	NA	TCCCACCTGACATCACAAAC	(Devers *et al.*, 2004)	atzD	Atrazine degradation	HUMAnN	Unknown
	GF.04	TGAGGCTGAAACTTTACGTAGA	NA	CTCACCTGGAGTTGCGTAC	(Baldwin *et al.*, 2003)	Xylene monooxygenase	Xylene degradation	HUMAnN	Unknown
	GF.05	GAGATGCATACCACGTKGGTTGGA	NA	AGCTGTTGTTCGGGAAGAYWGTGCMGTT	(Cébron *et al.*, 2008)	PAH-RHDα	Xylene degradation	HUMAnN	Gram negative polycyclic aromatic hydrocarbons digesters (such as Pseudomonas, Ralstonia, Commamonas, Burkholderia, Sphingomonas, Alcaligenes, Polaromonas)
	GF.06	CGGCGCCGACAAYTTYGTNGG	NA	GGGGAACACGGTGCCRTGDATRAA	(Cébron *et al.*, 2008)	PAH-RHDα	Xylene degradation	HUMAnN	Gram positive polycyclic aromatic hydrocarbons digesters (such as Rhodococcus, Mycobacterium, Nocardioides and Terrabacter)
	GF.07	CGAACTGATTCC	ACAAATCTCA	AGCGTTTGTAGGT	IDT	NarG	nitrate reductase alpha subunit	MEGAN5	None

		AGGCAAAAC	GTGGTTGAGCGCGA	GTTAGGG			(Molybdopterin oxidoreductase) COG5013		
	GF.08	GCTGGCACAAGCCTTTGATGT	TCCAGGCAAAACCGCACCACA	ATCGCGCTCAACCACTGAGA	PrimerBlast	NarG	nitrate reductase alpha subunit (Molybdopterin oxidoreductase) COG5013	MEGAN5	Comamonadaceae
Protein cluster	M3.1.4	CGACTGCGAACGACAATAAG	GTGGGCACCAATAAAGCTGT	GGTAATTGCTTTGGCGGTAA	Primer3Plus	Similar to conserved domain cd04080: Carbohydrate Binding Module 6, cellulase-like	Top hits to hypothetical proteins and to filamentous hemagglutinin outer membrane proteins	Protein clustering	None
	PC.02	AAGGCATCGAGAAAAACCGC	ACACAGAAGGTAAAGGCGAGA	ATCTCAACGCGACGGTTCTT	Primer3Plus (geneious)	OmpA	OmpA/MotB containing protein	protein clustering	Burkholderiales
Taxon	P.N.2.1.1	TCCAGTAGAGCAGTTGACAAAG	CGTGACCAAGGTGTATCTTCGCGT	ACCATCGAATCAGCAAGAGG	IDT	50S rRNA L10	Ribosome	MetaPhlAn	Polynucleobacter necessarius, database sequence 100131477
	P.N.2.1.3	CTGATGTCGGCGCTCAATTG	GCTCAAACAGTCGTGCTCGCTG	GCTTTGTCAACTGCTCTACTGG	IDT	50S rRNA L10	Ribosome	MetaPhlAn	Polynucleobacter necessarius, database sequence 100131477

Table 6 Results from qPCR "pre-tests" on original study samples

Biomarker basis	Primer set name	PCR pre-test result (single product expected)
Gene family	GF.01-04	No products
	GF.05, GF.06	Multiple products
	GF.07	OK
	GF.08	OK
Protein cluster	M.3.1.4	OK
	PC.02	OK
Taxon	P.N.2.1.1	OK
	P.N.2.1.3	OK

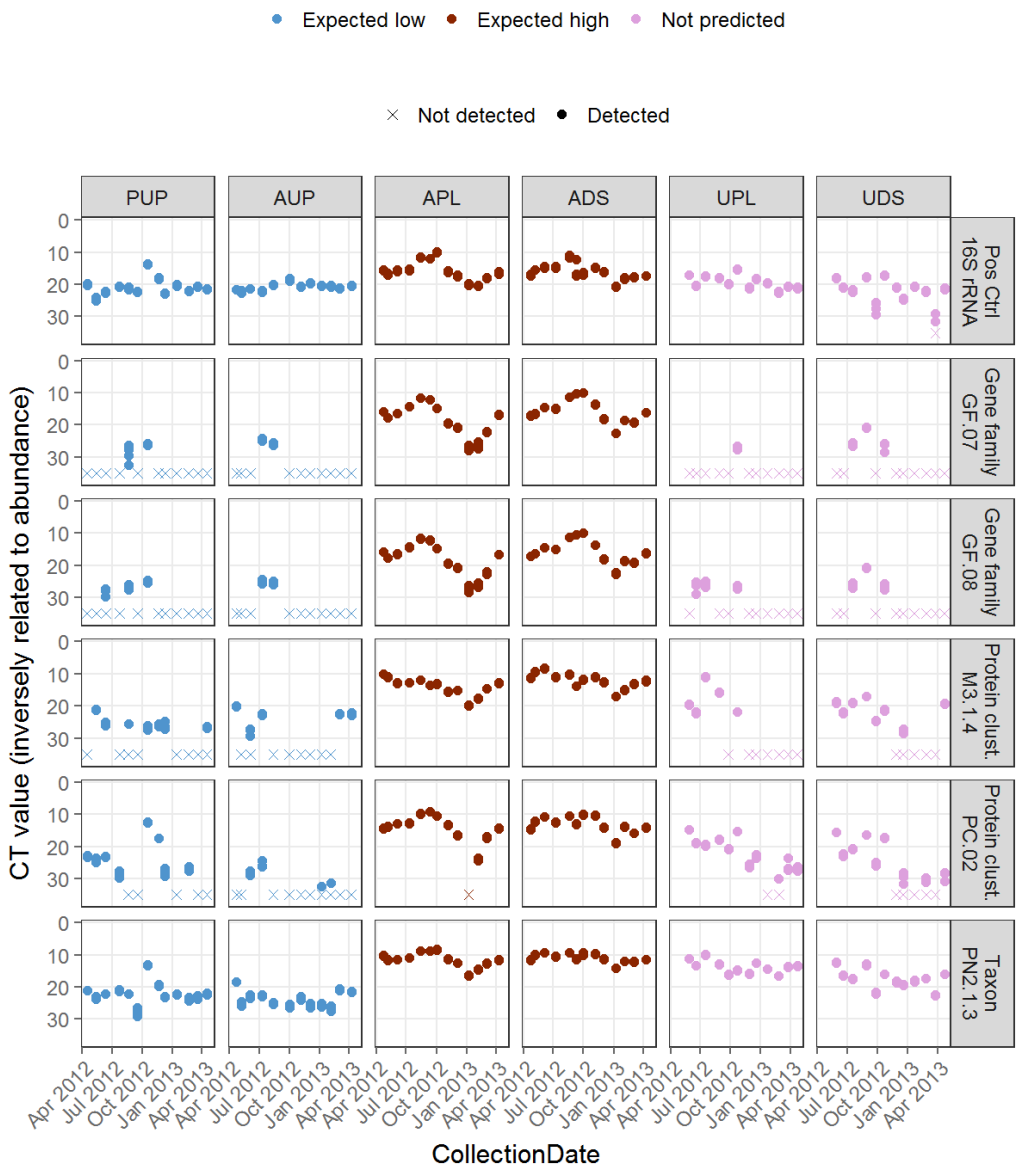

Figure 7 Results from qPCR tests of pre-checked biomarker candidates on original study samples over time

Each row of subplots displays results from a different candidate PCR marker, all of which passed initial screening (Table 6). Each column displays results from a different sampling site. Samples are represented by dots and coloured by the result expected from the qPCR. Each sample was tested four times. Values on the y axis are C_t values, where a lower value indicates a higher abundance. For legibility, low C_t values are at the top of the y axis. Samples represented by an "x" had C_t values above 35 (no signal after 35 cycles).

3.6. Conclusion

This work described methods to design qPCR biomarker assays from bacterial metagenomic data using gene families and taxa as indicators. In general, the qPCR tests designed using these methods performed as expected. However, validation against samples from different locations and years are required to assess the generalisability, and therefore utility, of these biomarkers beyond this proof-of-principle study. Overall, this work demonstrates the opportunity that metagenomic data provides to support the development of biomarker assays in poorly characterised environments.

Chapter 4.

Spatiotemporal dynamics of river viruses, bacteria and microeukaryotes

4.1. Foreword

This chapter is in preparation for publication and was authored by Thea Van Rossum, Miguel I. Uyaguari-Diaz, Marli Vlok, Michael A. Peabody, Alvin Tian, Kirby I. Cronin, Michael Chan, Jared R. Slobodan, Matthew J. Nesbitt, William W.L. Hsiao, Judy Isaac-Renton, Patrick K.C. Tang[+], Natalie A. Prystajecky[+], Curtis A. Suttle[+], and Fiona S.L. Brinkman[+]. +Senior authors contributed equally.

This work was performed within the GC Watershed Project (see section 1.2). Design of the sampling scheme and production of the data was a large collaboration (see author contributions in section 2.7). My role in this project was to aid in DNA sequencing optimisation and perform the data analysis, interpretation, and reporting. I performed the analysis of the viral and bacterial metagenomic data and built upon the amplicon analyses performed by Alvin Tian (g23) and Ana Maria Crisan (16S, 18S). The novelty of my approach in the analysis of the metagenome data was to use a k-mer based approach to analyse the variability in both the viral and bacterial data such that they could be easily compared. Since publishing a k-mer based analysis approach in Chapter 2, new tools were published to facilitate and formalise the process. Instead of using the more common analysis technique for viral metagenomes (based on clustering predicted-proteins), I used a k-mer approach to avoid limitations due to poor assembly of our data.

4.2. Abstract

Freshwater is an essential resource of increasing value, as clean sources diminish. Microorganisms in rivers, a major source of renewable freshwater, are of growing interest due to their role in drinking water safety, signaling environmental contamination (Baird and Hajibabaei, 2012), and driving global nutrient cycles (Meybeck, 2003; Findlay, 2010). Despite their importance, foundational understanding of microbial communities in rivers is lacking (Zeglin, 2015), especially across time and for viruses (Middelboe *et al.*, 2008; Peduzzi, 2016; Jacquet *et al.*, 2010). For example, no studies to date have examined the composition of free floating river viruses over time. Here we report analysis of temporal

relationships in riverine microbial communities across superkingdoms (viruses, bacteria, and microeukaryotes) and across space, using metagenomics and marker-based microbiome analysis methods. We found that many superkingdom pairs were synchronous and had consistent shifts with sudden environmental change. However, synchrony strength and relationships with environmental conditions were heterogeneous across locations and superkingdoms. Variable relationships were observed with seasonal indicators and chemical conditions previously found to be predictive of bacterial community composition (Ruiz-González *et al.*, 2015b; Staley *et al.*, 2015; Niño-García *et al.*, 2016; Zeglin, 2015), emphasizing the complexity of riverine ecosystems and raising questions around the generalisability of single-site and bacteria-only studies. In this first study of riverine viromes, distinct DNA viral communities were observed at the different geographic sites, suggesting the previously reported consistency in riverine bacteria across significant geographic distances (Niño-García *et al.*, 2016; Jackson *et al.*, 2014; Crump *et al.*, 2009), which we also observed, does not extend to viruses. This work provides foundational data for riverine microbial dynamics in the context of environmental and chemical conditions and illustrates how a bacteria-only or single-site approach would lead to an incorrect description of microbial dynamics. We show how more holistic microbial community analysis, including viruses, is necessary to gain a more accurate and deeper understanding of microbial community dynamics.

4.3. Introduction

Rivers provide critical ecosystem services such as drinking water, recreation, irrigation and nutrient cycling, yet their function is under serious threat from human activities that contaminate and shift microbial communities (Vörösmarty *et al.*, 2010). Though microorganisms are a key component of aquatic ecosystem function, fundamental knowledge of riverine microbial communities is lacking (Zeglin, 2015), in part because of the difficulty in culturing riverine microorganisms. However, progress has been made studying bacteria, especially with the rise of phylogenetic marker gene sequencing approaches (Zeglin, 2015). Riverine bacterial diversity and composition is shaped by water temperature, day length, pH (Niño-García *et al.*, 2016), nutrients (Ruiz-González *et al.*, 2015b; Staley *et al.*, 2015), water residency time (Read *et al.*, 2014; Niño-García *et al.*, 2016), and storm events (reviewed in (Zeglin, 2015)). Balancing these shaping forces, dispersal appears to play a large role both within (Staley *et al.*, 2015) and among (Niño-García *et al.*, 2016; Jackson *et al.*, 2014; Crump *et al.*, 2009) rivers, such that the relationship between community composition and spatial distance is minimal when not confounded by environmental conditions. Less is known about planktonic riverine microeukaryotes;

however, they appear to vary seasonally with light changes (Thomas *et al.*, 2012; Bradford *et al.*, 2013; Simon *et al.*, 2015), with some evidence indicating the importance of algae as an energy source (Bradford *et al.*, 2013).

In contrast to this basic characterisation of bacterial community variability, little is known about the community dynamics of free-floating viruses (virioplankton) in rivers (Middelboe *et al.*, 2008; Peduzzi, 2016; Jacquet *et al.*, 2010). River plankton viral metagenomes (viromes) have been reported in one study (Dann *et al.*, 2016), which found minimal differences in the taxa upstream versus downstream of a town; however, this analysis was limited by dependence on poor reference databases and only two samples were studied. Viral communities in lakes and oceans are better studied; however, these viromes are likely distinct from those in rivers given their differing hydrology and bacterial community compositions (Jacquet *et al.*, 2010; Aguirre de Cárcer *et al.*, 2015; Niño-García *et al.*, 2016). To date, there have been no large-scale studies of virioplankton composition in flowing (lotic) freshwater. As such, little is known about their community composition (Peduzzi, 2016; Middelboe *et al.*, 2008; Jacquet *et al.*, 2010) and basic questions, such as their variability throughout a year and the relative importance of dispersal and shaping forces in their community composition have gone unanswered.

To build fundamental knowledge of spatiotemporal variability of river plankton, we profiled viruses, bacteria, and microeukaryotes along with the context of varying environmental conditions. We sampled microorganisms monthly for a year from six sites in three watersheds in southwestern British Columbia, Canada (Figure 8a). For each sample, we performed metagenomic and/or phylogenetic marker gene sequencing (16S, 18S, g23 viral capsid) for DNA viruses, RNA viruses, bacteria (Van Rossum *et al.*, 2015), and microeukaryotes (Uyaguari-Diaz *et al.*, 2016). Environmental, chemical and biological measures were also collected (Uyaguari-Diaz *et al.*, 2016; Van Rossum *et al.*, 2015). Due to the lack of reference genomes available for freshwater viruses and the complexity of the communities, we estimated dissimilarity measures among metagenomes using a reference- and assembly-free k-mer approach (Mash (Ondov *et al.*, 2015)). To diminish any effects from potential bacterial or eukaryotic contamination in the viral data, DNA and RNA viromes are represented by two datasets. The “total” dataset includes all sequence reads. The “conservative” dataset is a subset of reads selected based on similarity to known viruses (see Methods for details). Spatiotemporal comparisons were performed within and between “superkingdoms”, viruses (DNA and RNA), bacteria, and microeukaryotes, and “environmental conditions”, catchment area weather, river water chemical concentrations, and river water physical conditions.

4.4. Methods

4.4.1. Sampling & sequencing

River water was collected monthly for 12 to 13 consecutive months from six sites in three watersheds in southwestern British Columbia, Canada. The agricultural watershed had three sampling sites, one upstream of human activity (AUP), one adjacent to intensive agriculture (APL), and one further downstream (ADS). The urban watershed had two sampling sites, one with a catchment mix of forest and residential land use (UPL), and one further downstream with mostly residential and some park land use (UDS). The pristine watershed was in a protected forest area, with no land use (PUP). Sampling sites were not downstream of any lakes or dams. Water temperatures were mild year-round, ranging from 3°C to 25°C. In the agricultural watershed, a distinct rainy period occurred from November to March, which is typical for the area. The other watersheds had more variable rainfall throughout the year. Sites from the same watershed were sampled on the same day. For full sampling and sequencing procedures see (Uyaguari-Diaz *et al.*, 2016) and (Van Rossum *et al.*, 2015); a brief overview follows.

At each sampling event, 40 L of water was collected and then filtered sequentially to concentrate particles approximating the sizes of microeukaryotes (105 to 1 µm), bacteria (1 to 0.2 µm), and viral-sized particles (Uyaguari-Diaz *et al.*, 2016). Physical and chemical water measurements were also taken (Van Rossum *et al.*, 2015). DNA was extracted from each size fraction, along with RNA from the viral-sized fraction (Uyaguari-Diaz *et al.*, 2016).

Amplicons for T4-like bacteriophages were prepared using primers targeting the myovirus g23 gene (Uyaguari-Diaz *et al.*, 2016; Filée *et al.*, 2005). Amplicons for bacteria were prepared using primers targeting the V3-V4 regions of 16S rRNA gene (Muyzer *et al.*, 1993; Caporaso *et al.*, 2011). Amplicons for microeukaryotes were prepared using primers targeting the V1-V3 regions of the 18S rRNA gene (Zhu *et al.*, 2005; Amann *et al.*, 1990). Amplicons were purified with a QIAQuick PCR Purification Kit (Qiagen Sciences, Maryland, MD) according to the manufacturer's instructions. Sequencing libraries were prepared for amplicons using NEXTflex ChIP-Seq Kit (BIOO Scientific, Austin, TX), gel size-selected as per manufacturer's instructions, and sequenced with 250-bp paired-end reads on an Illumina MiSeq platform (Illumina, Inc., San Diego, CA).

Bacterial metagenome libraries were prepared using Nextera XT DNA sample preparation kit (Illumina, Inc., San Diego, CA) and size selected using high-throughput gel-based Ranger technology (Uyaguari-Diaz *et al.*, 2015). Bacterial metagenomes were

sequenced over multiple runs with 250 bp paired-end reads on an Illumina MiSeq, with positive controls (mock communities) (Peabody *et al.*, 2015; Van Rossum *et al.*, 2015) and negative controls included in each run.

A modified adapter nonamer approach was used to synthesize viral cDNA and increase yields from the viral fraction (Uyaguari-Diaz *et al.*, 2016; Wang *et al.*, 2002). Viral metagenome libraries were prepared from randomly amplified DNA and cDNA using NEXTflex ChIP-Seq kit (BIOO Scientific, Austin, TX) by following a gel-free option provided in the manufacturer's instructions. These libraries were sequenced with 150 bp paired-end reads on an Illumina HiSeq platform (Illumina, Inc., San Diego, CA).

All raw sequences are deposited in the NCBI Sequence Read Archive under BioProject accession PRJNA287840.

4.4.2. DNA sequence pre-processing and quality control

Low quality bases were trimmed from the 3' end of reads using a sliding window with a minimum Phred score of 20 (or 15 for g23) using Trimmomatic (Bolger *et al.*, 2014). Adapters were removed using Cutadapt (Martin, 2011) with default parameters. Paired-end reads were merged using PEAR (Zhang *et al.*, 2014). Microeukaryotic 18S amplicon paired-end reads could not be merged, so Operational Taxonomic Units (OTUs) were generated from reads with the same primer sequence.

T4-like myovirus g23 amplicons reads were translated into amino acid sequences using Fraggenescan v1.16 with the Illumina 5% error model (Rho, Tang, and Ye 2010). OTUs were generated using USEARCH (Edgar, 2010) v7: sequences were dereplicated, clustered at 95% identity, then all reads were mapped back against cluster representatives to calculate abundances. Sample read totals were subsampled to 10,000 reads using the vegan package (Oksanen *et al.*, 2015) in R (R Core Team, 2013) v3.1.2. Random resampling was performed for 10,000 times and the median value of all iterations was chosen.

Bacterial 16S and microeukaryotic 18S OTUs were generated from amplicon reads using the Mothur (Schloss *et al.*, 2009) MiSeq clustering protocol (Kozich *et al.*, 2013) and rarefied to 10,000 reads.

Metagenomic reads were trimmed at the 3' end with a sliding window with a minimum Phred score of 20 using Trimmomatic (Bolger *et al.*, 2014). DNA virome reads shorter than 70 bp were discarded, resulting in a dataset of 20 Gb in 225 M reads. RNA virome reads

shorter than 100 bp were discarded and ribosomal reads were removed using meta-rRNA (Huang *et al.*, 2009), resulting in a dataset of 17 Gb across 149 M reads. Bacterial metagenome reads shorter than 100 bp were discarded, resulting in a dataset of 16 Gbp across 75 M reads.

4.4.3. Generation of high-confidence DNA & RNA virome datasets

Viromes were assembled using CLC and proteins were predicted from contigs using Prodigal in metagenomic mode with default parameters. Predicted proteins at least 26 amino acids long were clustered de novo using parallel cd-hit (Fu *et al.*, 2012), with criteria as previously used (Hurwitz and Sullivan, 2013): word length of 4 and 60% identity over 80% length of the shorter sequence. Reads were assigned to clusters with a blastx-style similarity search against cluster representative sequences using DIAMOND (Buchfink *et al.*, 2014) with minimum 60% sequence similarity over minimum 26 amino acid alignment length. While PC analysis is common in large scale marine studies (Hurwitz and Sullivan, 2013; Brum *et al.*, 2015), we did not use this dataset for primary analysis as many samples had a small proportion of reads in any protein cluster (mean 13%, range 8-30% of DNA virus reads and mean 25%, range 8-60% for RNA virus reads).

Contigs were tested for amino acid sequence similarity to reference sequences in NCBI's nr database using RAPSearch and taxonomically classified using MEGAN5. A small proportion of contigs were assigned as DNA viral (4% of contigs, 0.7% of total reads) and RNA viral (2% of contigs, 7% of total reads).

In the DNA virome dataset, 42% of contigs were assigned as bacterial, corresponding to 20% of assembled reads and 7% of total reads. To assess whether these bacterial assignments were due to miss-assignment of viral sequences (e.g. auxiliary metabolic genes, prophages) or an indication of bacterial contamination (e.g. from laboratory reagents, free-floating DNA, or host DNA packaged in viral capsid) (Hurwitz *et al.*, 2016), reads were tested for the presence of bacterial genes unlikely to occur in viruses. Across 515,000-read subsets of samples, similarity to the 16S rRNA gene was found in 1 to 156 reads (mean: 30, standard deviation: 25). Though these are small numbers, they are an indication of the number of bacterial genomes potentially present. This means that the contigs identified as bacterial in the taxonomic results cannot be ruled out as bacterial contamination. Further, the contigs that were left unassigned by the taxonomic classification also cannot be ruled out as bacterial.

To remove potential bacterial contamination from the DNA and RNA viromes, subsets of the read data were generated that only included sequences from protein clusters with at least one member that was assigned as coming from DNA or RNA viruses, respectively. This reduced the number of reads per sample from 515,000 in the "total" dataset to 10,000 in the "conservative" subset for DNA viromes and from 45,000 to 1,000 for RNA viromes. As this is a fairly small number of reads, we estimated the stability of distance matrices with low numbers of reads (see below) and used both total and conservative datasets to test trends.

4.4.4. Comparison with Tara Oceans dataset

DNA virus reads were compared at the protein similarity level against marine protein cluster representative sequences from a global survey of dsDNA viruses (Brum *et al.*, 2015) using DIAMOND (Buchfink *et al.*, 2014). Reads were considered similar to a marine sequence if they had at least 60% identity over 80% of the length of the read, which corresponds to the clustering cut-offs used to generate the marine reference dataset (Brum *et al.*, 2015). The total and conservative DNA virus data sets were used as queries, with 45,000 and 10,000 reads per sample respectively. A mean of 1.5% (standard deviation 0.5%) of reads per sample were similar to the marine data in the total dataset and 6.9% (standard deviation 3.7%) in the conservative dataset.

4.4.5. Sample similarity estimation & spatiotemporal analysis

Pairwise similarity between amplicon samples was performed using vegan (Oksanen *et al.*, 2015) in R (R Core Team, 2013) to calculate Bray-Curtis dissimilarity between OTU abundance profiles. Pairwise similarity between metagenomes was assessed using Mash (Ondov *et al.*, 2015), which tests for k-mer presence-absence. For display in heatmaps in Appendix B Figure B7, extreme values of similarities were collapsed to be represented by one color. Extreme values were defined as those values more than 2.5 times the median absolute deviation (MAD) away from the median (Leys *et al.*, 2013). Collapsed values were only used for display and not for any statistical tests.

Due to the small number of reads in the conservative RNA virus dataset, we investigated whether this depth was enough to obtain a stable representation of the communities. We randomly selected 1,000 reads ten times per sample from 68 samples that had at least 10,000 reads in the conservative RNA virus dataset. We ran Mash on these subsamples and calculated the pairwise Mantel correlations between the resultant dissimilarity matrices. All matrices had correlation scores of at least R = 0.95 with Pearson's

correlation and R=0.94 with Spearman's correlation. We decided this consistency was sufficiently high to justify confidence in high level patterns within this data.

All statistical tests were performed in R (R Core Team, 2013) v3. Permutation-based p values were calculated using 9999 permutations. Multiple test correction was performed where appropriate using the Benjamini-Hochberg procedure and adjusted p values reported as q values. Significance test values were considered statistically significant if lower than 0.05, except where indicated otherwise.

The proportion of variability among sample similarities that could be explained by sampling site was estimated using NPMANOVA as implemented in the adonis function from the vegan R package (Oksanen *et al.*, 2015). Gene family variability was based on SEED subsystem classifications (Van Rossum *et al.*, 2015) and calculated using Bray-Curtis dissimilarities. Synchrony was tested using Mantel matrix correlation tests with Spearman correlations, implemented in the vegan R package (Oksanen *et al.*, 2015). When testing samples from multiple sites for synchrony, a partial Mantel test was used to control for geographic distance between sampling sites. The NMDS plot in Figure 2 was generated using the vegan metaMDS function, with rotation and scaling of ordinations performed using the procrustes function and tested for significance using the pro.test function. Samples from April 2013 (105 and 106) were highly dissimilar and removed from this NMDS plot to enable the trends in the other 11 samples to be displayed. A similar plot for the other agriculturally affected site (ADS) is available in the supplementary material. Environmental data were tested for correlations with microbial community similarities using the envfit function. If applicable, the environmental measures to test were selected based on their magnitude and variability in the context of water quality guidelines (Canadian Council of Ministers of the Environment, 2007). Relationships among environmental measures were assessed using Spearman's correlation. Correlations within and among environmental measures and microbial community similarities were displayed in a network using the igraph R package (Csárdi and Nepusz, 2006) and the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold, 1991), followed my minor manual adjustments. Correlations that had a q value less than 0.1 were considered statistically significant. Correlations that had a p value less than 0.05 but a q value greater than 0.1 were not considered statistically significant but were included to avoid overconfidence in the absence of a relationship; however, they should be interpreted with caution.

4.5. Results & Discussion

Across superkingdoms, hours of daylight and rainfall intensity were the most commonly correlated with community composition (Figure 8 c, d), particularly in sites where these environmental measures were correlated due to large seasonal differences in rainfall intensity (Figure 8c sites AUP, APL, ADS; Appendix B: Figure B2, Figure B3, sFigure B4). In the sites where rainfall was not correlated with hours of daylight, the relationships between these environmental measures and superkingdom compositions were inconsistent (Figure 8c sites PUP, UPL, UDS; Appendix B: Figure B1, Figure B5, Figure B6). This is surprising as rainfall was hypothesized to have a particularly large and consistent impact on microbial communities since its intensity can affect the microorganisms arriving in a river by shaping both what lives in the catchment and what can be physically transported to the river. Instead, when not confounded with overall seasonal changes, rainfall was rarely significantly correlated with microbial community composition. Overall, no correlations among environmental conditions or superkingdoms were seen in all sites (Figure 8c).

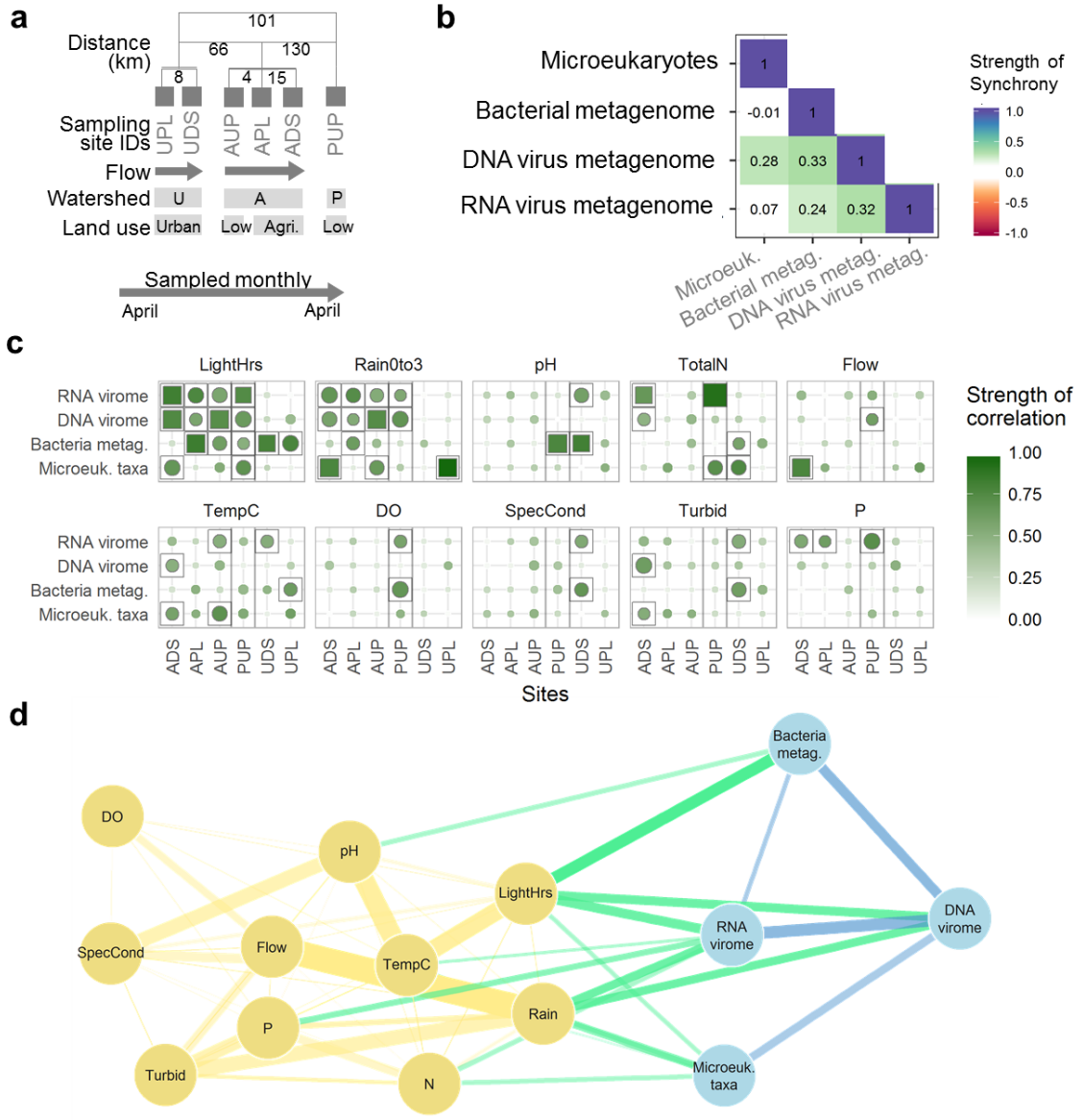

Figure 8 Temporal variation in viruses, bacteria, and microeukaryotes

a, Study design schematic of sampling sites with distances between sites, site orientation, watershed, and catchment land use. Distances are dendritic within watersheds and Euclidean between watersheds. Sites are in up- to down-stream order within watersheds. **b**, Pairwise partial Mantel tests for synchrony between viruses, bacteria and microeukaryotes, controlling for distance between sampling sites, N = 51 to 85, $q < 0.0004$. **c**, Correlations between microbial communities and environmental conditions per sampling site. Results are organised by environmental parameter into subplots where each row is a biological group and each column is a sampling site. Colour intensity reflects correlation strength. Shapes indicate the statistical significance of the correlation with squares as significant ($q < 0.1$) and circles as not statistically significant. Size of shape corresponds to the inverse of the statistical significance (q value). Grey square outlines indicate a relationship was statistically significant without multiple test correction ($p < 0.05$). Grey vertical lines separate watersheds. **d**, Network of correlations between and within environmental conditions and microbial communities, calculated per site. Nodes are environmental conditions (yellow) and microbial communities (blue). Conservative viromes were used (see methods). Edges are coloured by the nodes types they connect. Each edge represents cumulative relationships seen across the sampling sites, both those that are statistically significant ($q < 0.1$) and strong but with lower statistical confidence ($p < 0.05$). Edge width reflects the sum of the strengths (R^2) of the represented correlations. Edges are only drawn if at least one statistically significant or two lower-confidence correlations were observed, to reduce artefacts from arbitrary statistical cut-off values (see Methods).

Environmental conditions that have been previously reported as drivers of bacterial community composition were inconsistently correlated across sites and did not extend to other superkingdoms. For example, nitrogen and phosphorous concentrations were most often correlated with RNA viruses and/or microeukaryotes but not with bacteria, and pH was only correlated with bacterial composition in two sites, despite previous reports of it being a major driver (Niño-García *et al.*, 2016). Very few correlations were observed with dissolved oxygen concentration, flow intensity, specific conductivity, or turbidity. The range of correlations with environmental conditions observed across sites and superkingdoms emphasizes both the complexity and heterogeneity of riverine microbial ecosystems.

Despite inconsistent relationships with environmental conditions, viral and bacterial community compositions shifted in a similar manner over time (were “synchronous”) (Figure 9a), with the strength of synchrony varying among sampling sites (Appendix B: Figure B1 - Figure B6). Microeukaryotes had fewer synchronous relationships but were sometimes correlated with bacteria and/or DNA viruses. These cases of synchrony imply ecological relationships directly, due to interaction (e.g. predation, replication, lysis, etc.), or indirectly, due to a shared response to a varying third factor. In most cases, synchronous pairs were not significantly associated with a common third measure (Appendix B: Figure B1 - Figure B6). Synchrony between DNA viruses and bacteria is congruent with bacteriophage-host replication relationships and the theory that aquatic DNA viruses are predominantly bacteriophage. However, the observation in two sites that DNA viruses are also synchronous with microeukaryotes—without a third shared correlation pointing to an external driver in one site—raises questions about the generalisability of this hypothesis.

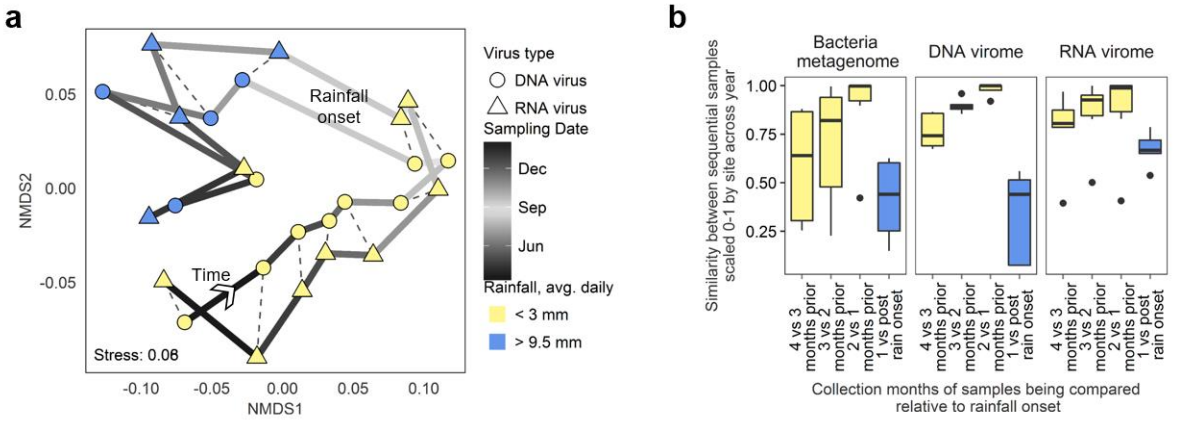

Figure 9 Onset of rainfall has consistent and large effect on riverine microplankton

a, NMDS plot of DNA & RNA viral communities from an agriculturally affected site (APL). Each point represents a viral community, solid lines connect sequential samples and are coloured by sampling date, dashed lines connect viromes extracted from the same sample. Points are coloured by the average rainfall over the three days prior to sampling. N= 13. **b**, Box plot of similarity between microbial communities collected in subsequent months, coloured by whether both sampling dates had

low rainfall (yellow) or whether the earlier date was dry but latter date had elevated rainfall (blue). N = 6 for bacteria and RNA viruses, N= 5 for DNA viruses.

Unexpectedly, DNA and RNA viral community compositions were synchronous in some sites (metagenomic and phylogenetic marker gene data, Mantel's r = 0.4 - 0.6, q = 0.02 - 0.001), even though they were not consistently synchronous with bacteria or microeukaryotes. Because few, if any, studies have profiled DNA and RNA viral community compositions concurrently over time, this synchrony has not been previously investigated. While correlational data cannot prove the drivers of synchrony, environmental data can provide context. Synchronous DNA and RNA viruses were also correlated with daylight hours (Appendix B: Figure B2, Figure B3, Figure B4) and a temporal trend is clear: sequential samples tended to be most alike and shift stepwise over time (Figure 9a, one site as example; this pattern was visible in the other sites but was less clear, see Appendix B Figure B9). This suggests that the DNA and RNA viral synchrony is not artefactual, but due to some temporal relationship, possibly with a common host group, despite the general thought that DNA and RNA viruses infect hosts with different niches.

Large shifts in DNA and RNA viromes in the agriculturally affected sites were concurrent with the onset of rainfall after a dry period. This trend was also observed in the other sampling sites and in bacterial communities (Figure 9b, microeukaryotic communities not tested due to insufficient data). These observations demonstrate the first-flush phenomenon; dry periods permit a buildup of solids, chemicals, metals, and organisms and the first significant rainfall causes an abrupt shift in the bacterial and viral communities in the receiving waters (Williamson *et al.*, 2014; Deletic, 1998). This shows that while continuous relationships with rainfall were not universal (Figure 8), response to a rainfall event was more common.

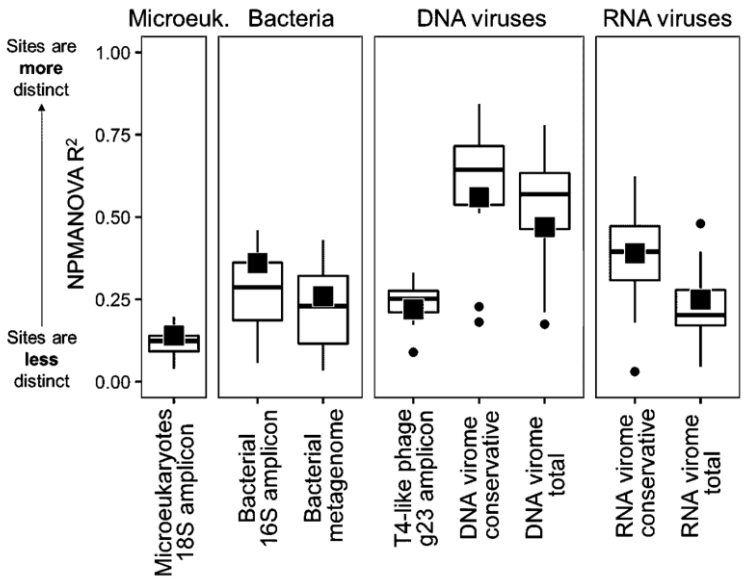

Figure 10 Geographic distinctiveness within viral, bacterial, and eukaryotic communities over 1 year of monthly samples

Proportion of variability among samples that is explained by sampling site (NPMANOVA R^2), either across all sites (black square) or pairwise between sites (boxplots). In boxplots, the lower and upper box edges correspond to the first and third quartiles, the whiskers extend to the highest and lowest values that are within 1.5 times the inter-quartile range, and data beyond this limit are plotted as points. The proportion of variability explained by sampling site in T4-like amplicon viruses was lower than viromes because though the highest similarities were always within sampling sites, many within-site sample pairs had low similarity.

While sampling site was a significant source of variation for all microbial groups, DNA viromes showed stronger geographic-based similarity than bacteria and microeukaryotes (Figure 10, Appendix B: Figure B7). This is consistent with the distinctiveness of T4-like bacteriophage seen in a study of polar lakes (De Cárcer *et al.*, 2016) but is in contrast with the similarity of DNA viruses seen in two temperate lakes (Mohiuddin and Schellhorn, 2015). Analysing bacterial amplicon data at a finer taxonomic resolution (99% identity OTUs) did not significantly increase its geographic distinctiveness (data not shown). This low geographic distinctiveness of bacteria, particularly among sites with similar land use (pairwise NPMANOVA between the two agriculturally affected sites and between the two urban-affected sites: $R^2 <= 0.23$, $q = 0.0003$, Appendix B: Figure B8), is consistent with previously shown low spatial stratification of bacteria among rivers (Niño-García *et al.*, 2016; Jackson *et al.*, 2014; Crump *et al.*, 2009). In the one case where land use varied within a watershed (Figure 8a, AUP versus APL & ADS), land use and associated water chemistry differences appeared to override geographic proximity as a predictor of microbial community similarity (Appendix B: Figure B7). These findings support a major ecological role of dispersal at this geographic scale (10 – 130 km) for riverine bacterial and microeukaryotic plankton but reveals that viruses have a distinct geographic pattern.

The higher geographic specificity of viruses observed here could reflect higher geographic specificity of host cells not sampled in this study, such as particle-associated plankton, riverbed biofilms, plants, humans, or other animals. Alternatively, viruses may be more geographically distinct because they replicate in the subset of microbial cells in the community that are active (20-50% bacterial cells (Lennon and Jones, 2011)). This subset is more likely to be geographically distinct due to their increased susceptibility to selective pressures (Lennon and Jones, 2011) and more likely to be represented by viruses due to the mechanics of the lytic cycle and host-specificity (Paez-Espino *et al.*, 2016). Thus, we hypothesise that viruses may produce a stronger geographic signal than bacteria by amplifying the effect of species sorting against the background of widely dispersed inactive cells.

4.6. Conclusion

In conclusion, temporal and spatial profiling revealed contrasting patterns between and among superkingdoms and environmental conditions in riverine microbial plankton. By examining microbial communities across time, the general—but not universal—importance of daylight hours as a correlate with composition was found for planktonic communities in this geographic region. Other expected correlates were also identified, however, by examining multiple locations, these relationships were revealed not to be universal, even within similar sites. This demonstrates the heterogeneity of riverine microbial ecosystems and the need for multi-site studies in riverine microbial ecology, as a similar study of a single site may have falsely concluded general trends. By examining multiple superkingdoms, correlations with nutrient concentrations were identified that would have been missed if only bacteria were profiled, doubt was cast on the universal predominance of bacteriophage in DNA viromes, and the strong dispersal observed in bacteria and microeukaryotes was revealed not to extend to viruses. In summary, this study provides insight into the variability of microbiomes over superkingdoms, time, and space in an understudied environment. It reveals notable differences in community dynamics across microbial groups, and demonstrates the value of collectively studying microeukaryotes, bacteria and viruses across multiple time points and locations in microbiome studies.

4.7. Author contributions

JIR, PKCT, NAP, CAS, and FSL designed the study, guided the analyses, aided in interpretations, and acquired funding. MIU led the sampling and sequencing, with assistance from MC, KIC, and TV. JRS and MJN performed size selection of sequencing libraries. TV

performed all bioinformatics and data analysis and wrote the manuscript, in consultation with FSL. AT compiled OTU tables for the g23 data. MV, MAP, and WWLH guided analyses. All authors contributed to final revisions of the manuscript.

4.8. Funding

This work was funded by Genome BC and Genome Canada grant No. LSARP-165WAT, with major support from the Simon Fraser University Community Trust Endowment Fund and additional support from the Public Health Agency of Canada. TV was supported by NSERC PGSM & CGSD scholarships. MP was supported by a CIHR/MSFHR Bioinformatics Training Program fellowship and an NSERC PGSD scholarship. MU was supported by a Mitacs Accelerate fellowship.

Chapter 5.

Microbiome analysis across a natural copper gradient in a freshwater stream at a proposed Northern Canadian mine site

5.1. Foreword

This chapter was published with minor changes in the journal *Frontiers in Environmental Science*, in an article of the same title, co-authored by Thea Van Rossum[#], Melanie M Pylatuk[#], Heather L Osachoff, Emma J Griffiths, Raymond Lo, May Quach, Richard Palmer, Nicola Lower, Fiona SL Brinkman and Christopher J Kennedy. Thea Van Rossum and Melanie M Pylatuk contributed equally ([#]). For supplementary material referred to in the text, see http://journal.frontiersin.org/article/10.3389/fenvs.2015.00084/full .

This chapter describes work from a partnership between Genome BC, the Casino Mining Corporation, Palmer Environmental Consulting Group, and the labs of Fiona SL Brinkman and Christopher J Kennedy at SFU. In this project, we had the opportunity to investigate microbial communities across a small number of locations along gradient of naturally metal-rich and acidic environments at a single time point. This investigation included a limited metagenomic analysis, which enabled some study of eukaryotes and viruses. However, due to the limited nature of the project, this study focused on the bacterial community.

Design of the experiment and production of the sequencing data was a collaboration led by CJ Kennedy, FSL Brinkman and members of the Casino Mining Corporation (see author contributions in section 5.8). My role in this project was to consult on sequencing strategies and lead the data analysis and reporting. My approach in the analysis of this data was notable in using complementary phylogenetic and metagenomic analyses and the incorporation of a database of metal-specific genes.

5.2. Abstract

Due to the environmental persistence, bioaccumulation, and toxicity of metals released by mining activities, mitigation methods are crucial to minimize impacts on aquatic

environments. Bioremediation is one mitigation strategy used to reduce the potential for metal accumulation and toxicity in aquatic organisms. At a potential mine site in Yukon, Canada, elevated copper (Cu) concentrations and low pH are found in a water course near a naturally mineralized area; however, Cu concentrations and acidity are greatly reduced downstream. Physicochemical processes do not appear to explain this natural remediation and it is suggested that unique microbial communities may be responsible through Cu immobilization**.** To investigate the role of microbes in sequestering or transforming Cu in the water, biofilm samples were collected from 5 sites along a natural copper gradient: upstream of Cu introduction, on a Cu-rich tributary, 30 m downstream of Cu introduction where Cu levels were reduced, and 2 and 7 km further downstream where Cu concentrations were low. Taxonomic profiles of microbial communities (microbiomes) were compiled using DNA sequencing of 16S rRNA gene amplicons. Clear relationships between total Cu concentrations, pH and the microbiomes were evident. In the most Cu-affected samples, communities were dominated by bacteria from the *Gallionellaceae* family. Metagenomic sequencing profiled the genes present in microbiomes from the most Cu-contaminated sampling location and the area immediately upstream and showed that microbes in this area have genes thought to be involved in heavy metal tolerance. This study provides fundamental knowledge of microbial communities at a potential mine site and characterizes the genes possibly involved in providing tolerance to an acidic and metals-rich environment. These results inform hypotheses for future experiments to support the development of bioremediation approaches that incorporate native microorganisms from mining sites.

5.3. Introduction

Mining is an important natural resource extraction method that typically requires mitigation strategies to reduce or eliminate resulting adverse environmental impacts. If impacts to aquatic environments are predicted during the design and assessment phases of mine development, water treatment is often considered. However, current water treatment strategies and the management of resulting sludges can be extremely costly and time consuming (Perales-Vela *et al.*, 2006). Bioremediation, an alternative to conventional physicochemical treatment processes, uses biological organisms to metabolize, alter or capture contaminants of concern in an engineered project or at a contaminated site. Bioremediation examples include: toluene-degrading bacteria in a Massachusetts, United States of America (USA) watershed (Tay *et al.*, 2001); mercury-accumulating periphyton in Boreal Canadian Shield Lakes (Desrosiers *et al.*, 2006); and naphthenic acid-removing bacteria in oil sands processing water, Alberta, Canada (Islam *et al.*, 2015). Bioremediation may be the preferred option for reducing contaminant concentrations in circumstances

where access is difficult, habitat will be destroyed, species at risk will be disrupted, or the area is simply too large to be feasibly physically remediated. Hence, bioremediation, along with geochemical processes, can contribute to the attenuation of environmental impacts and may be an innovative option to solve some contamination issues.

In Northern Canada (Yukon Territory), a copper-gold-silver-molybdenum mine has been proposed near Casino Creek by the Casino Mining Corporation (CMC)[iii]. Baseline environmental assessments of the aquatic environment have shown that the background concentrations of metals in some areas are elevated above Canadian Council of Ministers of the Environment (CCME) water quality guideline values for the protection of aquatic life[iv], which are concentrations adopted by the Yukon Government to protect the environment. More specifically, preliminary studies found that copper (Cu) concentrations were naturally elevated in some parts of Casino Creek but reduced immediately downstream. Physical and chemical evaluations (e.g., hydrogeological modeling of dilution and calculations of solubility limits for common minerals [e.g., tenorite (CuO)]) offered no indication as to the cause of reductions in Cu concentrations, suggesting that this may be a case of natural Cu bioremediation.

Metals bioremediation is performed by many types of organisms, including bacteria, fungi and algae (Tay *et al.*, 1998; Arini *et al.*, 2012; Desrosiers *et al.*, 2006). Across a wide range of environments, including aquatic ecosystems, these organisms naturally organize into biofilm communities (Malik, 2004) in which cells adhere to each other on a surface and produce a matrix of extracellular polymeric substance (EPS). Some biofilms can withstand high concentrations of metals (Orell *et al.*, 2010 and references therein) and acidic conditions (Arini *et al.*, 2012; Baker and Banfield, 2003), which make them a useful tool in developing new biotechnologies for mining-related areas of research. The science of strategically using biofilms is continuously developing (ITRC, 2008), including their use for bioremediation purposes.

Biofilms are complex and difficult to cultivate artificially, and new tools to study microbiomes *in situ* have enabled the profiling of biofilm community structures based on their DNA sequences (Besemer *et al.*, 2012), greatly enhancing the knowledge base regarding their natural compositions. The objective of this study was to use new approaches to characterize native microorganisms from a proposed mine site and investigate their potential involvement in Cu bioremediation. This study used microbiome profiling and metagenomic

[iii] http://www.casinomining.com/project/proposed_mine/

[iv] http://www.ccme.ca/en/resources/canadian_environmental_quality_guidelines/index.html

techniques (i.e. new DNA tools) to investigate biofilms collected from an artificial substrate placed *in situ* at five sites in the Casino Creek watershed to describe how the microbiome varies concurrently with Cu concentrations and acidic conditions. Microbial taxonomic profiles were compiled using amplicon sequencing of the 16S rRNA genes present in extracted DNA and community profiles were compared between sites. Genes from select samples were profiled using shotgun methods for metagenomic DNA sequencing to characterize the entire microbial community and identify metal-associated genes that could possibly explain elevated Cu-tolerance and the hypothesized ability to reduce Cu concentrations in the water column. The results of this study may be useful to inform future development of bioremediation strategies using native microbes to mitigate potential increases in environmental metals concentrations from mining operations. This is one of the first studies to investigate the microbial community structure and gene content of microorganisms from a naturally mineralized area prior to the initiation of mining operations and therefore represents an example of microbial investigations that can support natural resource extraction processes.

5.4. Materials and Methods

5.4.1. Sampling locations

Five locations for biofilm sampling were selected in the Casino Mine watershed, Yukon, Canada (Figure 11; Table 7). Site A is located furthest upstream on Casino Creek, just prior to the introduction of Cu to the system (via the Proctor Gulch tributary). Site B is on the copper-rich tributary in Proctor Gulch, immediately upstream of the confluence with Casino Creek, and has the highest Cu concentrations with visible dark orange staining on rocks. Site C is on the Casino Creek mainstem and is 30 m downstream of Site A and Cu concentrations were slightly elevated and some orange staining on rocks was visible. Site D is approximately 2 km downstream of Site C and is located in the proposed mine tailings pond area. Site E is furthest downstream, approximately 5 km south of Site D, and sits at the toe of the proposed tailings pond.

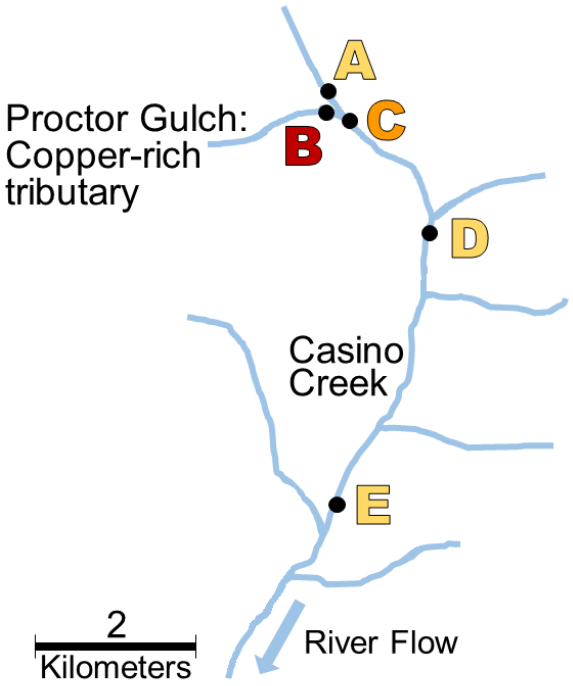

Figure 11 Casino Creek sampling site locations
A map of biofilm sampling locations (Sites A to E) along Casino Creek, Yukon, Canada.

Table 7 Sampling site descriptions and conditions.

Site Name	Site ID	Location Description	Substrate	Water Temperature (°C)	Dissolved oxygen saturation (%)	Specific Conductivity (µS/cm²)	pH	Total Copper (mg/L)	Dissolved Copper (mg/L)	Notes
A	W101	Casino Creek, upstream of Proctor Gulch	boulder	1.5	87.7	327	6.19	0.12	0.10	
B	W12	Proctor Gulch	embedded substrate	0.6	88.8	1018	3.75	1.02	0.96	dark orange staining on rocks
C	W102	Casino Creek, downstream of Proctor Gulch	boulder / cobble	1.3	87.8	511	4.19	0.37	0.35	orange staining on rocks
D	W8	Casino Creek in proposed tailings	cobble / boulder	1.3	95.2	244	7.19	0.07	0.02	
E	W11	Casino Creek at toe of proposed Tailing dam	cobble / boulder	1.7	93.5	246	7.26	0.03	0.01	

Sites B and C have higher acidity and concentrations of copper.

5.4.2. Casino Creek biofilm sample collection

Biofilm samplers were deployed in triplicate at each sampling location on the same day. Samplers consisted of PVC pipes, 20 cm in length and 5 cm in diameter (approximate volume 400 ml), containing 5 mm glass beads inside a nylon bag as the substrate for biofilm colonization (Besemer *et al.*, 2012). Plastic netting was placed over the ends of the PVC pipes to exclude large debris while allowing for water flow through the sampler. Samplers were tethered and submerged at depths of 0.3 to 0.4 m. Median water hardness measured during baseline environmental assessments in Casino Creek was 111 mg/L $CaCO_3$.

The biofilm samplers were left *in situ* for 18 d during late summer / early fall. On the final day, water quality parameters (dissolved oxygen, pH, conductivity) and substrate information were noted (Table 7). Copper concentration in water samples was measured by atomic absorption spectroscopy. Samplers were retrieved and placed in sterile plastic bags with site water and kept cool on ice until processed. To harvest the biofilms (colonized organisms), glass beads were removed from the nylon bag into a glass beaker containing 0.5 volumes of corresponding site water and sonicated approximately 10 min. To collect and concentrate the biofilm organisms, the water was filtered through a Super200 0.22 µm filter (Pall Corporation, VWR, Mississauga, ON, Canada). This process was repeated using the remaining 0.5 volume of corresponding site water. Filters containing the organisms were stored in 15 ml tubes and kept frozen during shipping (via air from Whitehorse, Yukon to Burnaby, BC) until transfer to -80 ºC.

Whole, dry filters were cut using sterilized scissors into 10 - 15 mm wide strips. Filter segments were placed into a Powerlyzer® glass bead tube containing 0.75 ml bead solution and DNA was extracted using the Powerlyzer® PowerSoil® DNA isolation kit (Mo Bio Laboratories Inc., Carlsbad, CA, USA) following the manufacturer's instructions. The elution step was performed twice for a total final volume of 0.12 ml of DNA. The DNA samples were stored at -20 ºC. DNA concentrations were evaluated using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). A sample was considered acceptable if it had a DNA concentration greater than 5 ng/µl.

5.4.3. Amplicon library preparation and sequencing

DNA concentrations were adjusted to 5 ng/µl. Each Polymerase Chain Reaction (PCR) contained: 5 ng of biofilm DNA, 13 µL of molecular biology grade water, 10 µL of 5 Prime Hot Master Mix (5 Prime, Gaithersburg, MD, USA) and 0.5 µL of each forward and reverse primer (10 µM final concentration) as described in Caporaso *et al.* (2011). PCR was

performed in triplicate on a TGradient Thermocycler (Biometra, Horsham, PA, USA) with the following program: 94 °C for 3 min; 35 cycles of 94°C for 45 s, 50 °C for 60 s, and 72 °C for 90 s; and finally an extension at 72 °C for 10 min. Primers that amplify the V4 to V5 region of the 16S rRNA gene were obtained from (Klindworth *et al.*, 2013), forward S-D-Bact-0564-a-S-15: 5'- AYT GGG YDT AAA GNG -3' (520F) and reverse S-D-Bact-0785-b-A-18: 5'- TAC NVG GGT ATC TAA TCC -3' (802R). A 2% agarose gel was run on each PCR sample to confirm amplification occurred. PCR cleanup of the 16S amplified samples was performed as described in the 16S Metagenomic Sequencing Library Preparation instructions (Illumina, San Diego, CA, USA).

Indices were incorporated into the 16S amplified samples as per the Illumina 16S Metagenomic Sequencing Library Preparation instructions. Each index reaction used 10 µL of DNA, 5 µL of Index primer 1 (N701-N705), 5 µL of index primer 2 (S502-S504 or S517), 25 µL of 5 Prime Hot Master Mix, and 10 µL of molecular biology grade water. PCR was performed to anneal indices to the 16S rRNA amplicon with the following conditions: 95 °C for 3 min; 8 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s; a final extension at 72 °C for 5 min. Several control samples were included at this stage: (1) a MilliQ water-only negative control, (2) a MilliQ water and DNase negative control, (3) a filter-only negative control, and (4) a spiked positive control. Negative controls showed undetectable (< 0.4 ng/µL) amounts of DNA prior to amplification. The spiked positive control contained an in-house laboratory mixture of DNA extracted from cultures of: *Bacillus amyloliquefaciens* (FZB4Z), *Escherichia coli* (K12), *Pseudomonas aeruginosa* (PAO1), *Pseudomonas putida* (KTZ440) and *Rhodobacter capsulatus* (SB1003).

A second PCR cleanup of the indexed 16S amplified samples occurred as described in the Illumina 16S Metagenomic Sequencing Library Preparation instructions. Prior to input on the Illumina MiSeq cartridge, DNA concentrations were assessed using a Qubit® 2.0 Fluorometer (Life Technologies, Grand Island, NY, USA), and DNA quality and size (to confirm 390 bp products) were assessed using a DNA 1000 chip on a Bioanalyzer 2100 instrument (Agilent Technologies, Santa Clara, CA, USA). DNA libraries were normalized to 4 nM with 10 mM Tris, pH 8.5. All samples and controls were pooled and run on the Illumina MiSeq Sequencer, following the Illumina 16S Metagenomic Sequencing Library Preparation instructions. All raw sequences are deposited in the NCBI Sequence Read Archive under BioProject ID: PRJNA297682.

5.4.4. Metagenomic library preparation and sequencing

Two samples were selected for further analysis using shotgun sequencing: one from Site B, where the Cu concentration is highest, and one from Site A, upstream of Cu introduction (Table 7). The same in-house produced positive control as described above was also processed with these samples.

Prior to input on the Illumina MiSeq cartridge, DNA concentrations were assessed using a Qubit® 2.0 Fluorometer (Life Technologies) to check for appropriate concentration (0.2 ng/µL) and cluster density for the Illumina instrument. Tagmentation was performed according to the Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA, USA) instructions. The amount of DNA for each sample varied slightly but was at least 1 ng. Indices were incorporated into the samples for metagenomic sequencing as per the Illumina Nextera XT DNA Sample Preparation Kit instructions. Each index reaction contained 25 µL DNA, 5 µL of Index primer 1 (N701 or N702), 5 µL of index primer 2 (S502 or S504), and 15 µL of NPM master mix. To anneal index primers to tagmented DNA, PCR was performed as follows: 72 ^{o}C for 3 min; 95 ^{o}C for 30 s; 12 cycles of 95 ^{o}C for 10 s, 55 ^{o}C for 30 s, and 72 ^{o}C for 30 s; a final extension at 72^{o}C for 5 min.

PCR cleanup was performed for the metagenomic libraries following the procedure from the Illumina Nextera XT DNA Sample Preparation Guide. For the metagenomic libraries, DNA quality and size (to verify the appropriate range of 100 - 1000 bp) were assessed using a DNA 1000 chip on a Bioanalyzer 2100 instrument (Agilent Technologies). DNA concentrations were assessed using a Qubit® 2.0 Fluoromoter (Life Technologies). Each library was diluted to 4 nM with 10 mM Tris, pH 8.5, and pooled with 0.1X volume of positive control before being run on the Illumina MiSeq Sequencer, following Illumina Nextera XT DNA Sample Preparation Guide. All raw sequences are deposited in the NCBI Sequence Read Archive under BioProject ID: PRJNA297682.

5.4.5. Amplicon data analysis

Amplicon sequence reads were preprocessed according to best practices (Schirmer *et al.*, 2015). Low confidence bases were trimmed by Phred score using Trimmomatic (Bolger *et al.*, 2014), with a sliding window of length 3 and a minimum Phred score of 20. Error correction was performed using BayesHammer (Nikolenko *et al.*, 2013), followed by merging of overlapping paired reads using PEAR (Zhang *et al.*, 2014), both with default settings. Reads containing ambiguous bases and reads that were > 5% longer or shorter than expected were discarded. All samples were rarefied to 100,000 reads before using

QIIME v 1.9 (Caporaso *et al.*, 2010b) to perform open-reference OTU picking. Reads were clustered using uclust (Edgar, 2010) at 97% identity against the Greengenes v13_8 16S rRNA database (DeSantis *et al.*, 2006). Reads that did not match to a reference sequence were clustered *de novo* at 97% identity. Where possible, OTUs were taxonomically annotated using uclust (Edgar, 2010) against Greengenes reference sequences (DeSantis *et al.*, 2006). OTUs with less than 3 reads were removed to avoid noise. A PyNAST (Caporaso *et al.*, 2010a) alignment of OTUs was used to build a phylogenetic tree using FastTree (Price *et al.*, 2009). OTU diversity and abundances were analysed in R 3.0 using PhyloSeq (McMurdie and Holmes, 2013) and vegan (Oksanen *et al.*, 2015).

In the positive control, 40,095 reads were assigned to 1,087 OTUs, with 99.7% of reads assigned to OTUs from the four expected families. The most abundant OTU assigned to the incorrect family contained 0.015% of the sample's reads and was assigned to the most abundant family from the experimental samples (*Gallionellaceae*). This indicates that OTUs with extremely low abundance may be due to a small amount of sample cross-contamination. In the experimental samples, 299,356 reads were assigned to 6,767 OTUs. After OTUs with abundances lower than 0.015% were discarded, 251,783 reads were assigned to 1,579 OTUs. To compare across samples, all samples were sub-sampled to the lowest per-sample read count (2,432 reads), such that 36,480 reads were assigned to 1,128 OTUs across all samples.

Alpha(α)-diversity was measured using Shannon-Weiner index and β-diversity was measured using weighted Unifrac distances. Correlations between OTU abundances and Cu concentration were measured by first discarding low-variance OTUs (variance in proportional abundance across samples < 1E-7). The remaining OTUs were tested for correlation using Pearson correlation and p values were adjusted for false discovery rate using Benjamini-Hochberg method. Statistical significance was determined when q values were less than 0.05.

5.4.6. Metagenomic data analysis

Metagenome reads were trimmed to remove low confidence bases using Trimmomatic (Bolger *et al.*, 2014), with a sliding window of length 5 and a minimum Phred score of 20. Sequencing adapters were removed using cutadapt (Martin, 2011), overlapping paired-end reads were merged using PEAR (Zhang *et al.*, 2014), and reads shorter than 100 bp were discarded. After this processing, 3 million and 10 million reads remained in samples from Sites A and B, respectively. Reads were then compared against BacMet, a database of metal-associated genes using BlastX with a 60% identity threshold (Pal *et al.*, 2014).

Assembly of reads from Site B was performed using SPAdes with default parameters (Bankevich *et al.*, 2012). Experimental samples were subsampled down to 1.8 million reads to be compared against nr (downloaded January 22, 2015) using RAPSearch2 (Zhao *et al.*, 2012). Protein alignments with an e-value less than 0.01 and length greater than 30 amino acids were analysed using MEGAN version 5.10 (Huson *et al.*, 2011) to determine the taxa and gene families present. MEGAN was run using default parameters to assign reads using the March 2015 taxonomic reference file and the SEED database (Overbeek *et al.*, 2005) mapping file (most recent version: January 2014).

5.5. Results

5.5.1. Cu-rich samples were dominated by *Gallionellaceae*

Total Cu concentrations at sampling sites ranged from 0.03 to 1.02 mg/L (Table 7) and were highly correlated with specific conductivity and pH (Pearson's Rho: 1.0 and 0.91, respectively) and not correlated with temperature (data not shown) or percent saturation of dissolved oxygen (Table 8). Taxonomic compositions of sampled microbiomes (the community of microorganisms) were predicted from DNA sequences of the 16S rRNA gene. Overall, *Proteobacteria* was the most abundant phylum across samples, with the remainder of the communities being composed of *Bacteroidetes* in Sites D and E and a combination of phyla in Site A (Figure 12; class and order data presented in supplementary material S1A and S1B). Sites B and C were dominated entirely by *Proteobacteria,* except for a small amount of *Cyanobacteria* in Site B. The dominant taxonomic family (87 - 93%) present in the Cu-exposed microbiomes at Sites B and C was *Gallionellaceae* (Figure 12). Across all other samples, 0.2 – 23% of reads were assigned to *Gallionellaceae*. Site A was the least characterized at the family level, with the majority of the community belonging to an unknown family, while the most abundant family at Sites D and E was *Comamonadaceae* (Figure 12).

Table 8 Correlations among metadata variables.

	Cu	Cond.	pH	DO %sat.
Total copper concentration (mg/L) (Cu)	1			
Specific conductivity µS/cm (Cond.)	1*	1		
pH	-0.91*	-0.92*	1	
Dissolved oxygen % saturation (DO %sat.)	-0.49	-0.53	0.77	1

Total copper (mg/L) and dissolved copper (mg/L; data not shown) values are correlated at Pearson's Rho = 1. Conductivity (µS/cm; data not shown) and Specific Conductivity (µS/cm^2) values are correlated at Pearson's Rho = 1. Abbreviations

used: Cu = total copper concentration (mg/L), Cond. = specific conductivity (µS/cm), DO %sat. = dissolved oxygen % saturation. Asterisks (*) indicate statistically significant Pearson correlations ($p < 0.05$).

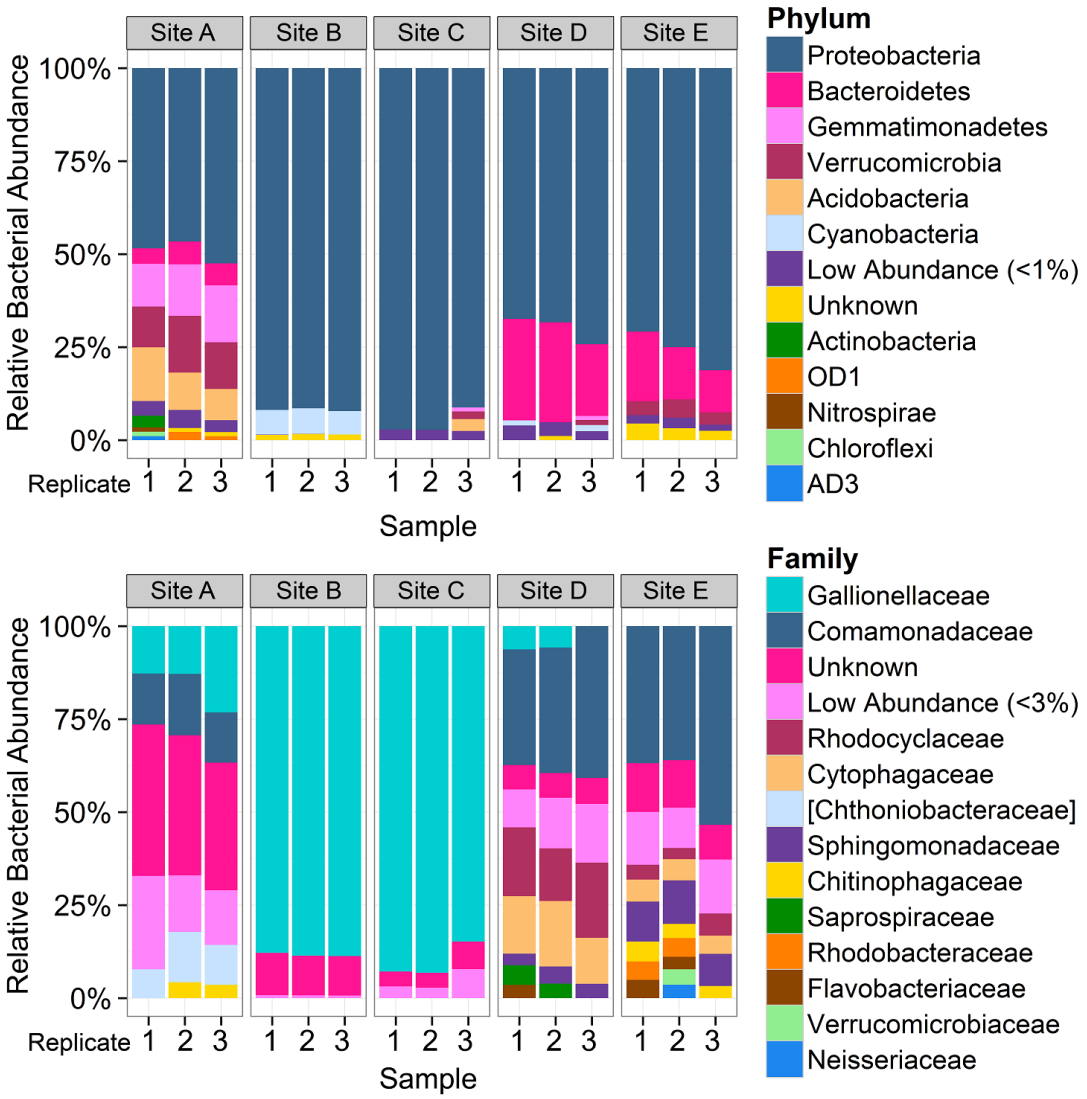

Figure 12 Overview of taxonomic composition of each sample by phylum and family.

Bar plots represent the proportion of reads assigned to the taxonomic composition of each sample by phylum (top) and by family (bottom), grouped by sampling site (n = 3 per site). Colours do not correspond between figures. Low abundance taxa (less than 3% at the phylum level and 1% at the family level) were collapsed to improve readability. Square brackets indicate proposed taxonomy based on Greengenes phylogenetic analysis. Phylum level plot shows that Site A was distinct from other low-copper sites and that cyanobacteria were present in Site B. Family level analysis shows dominance of *Gallionellaceae* at Sites B and C. The higher number of OTUs that could not be assigned to a family at Site A indicates that those bacteria were more distantly related to known bacteria than those present in the other sites.

Of all the reads assigned to the 593 *Gallionellaceae* operational taxonomic units (OTUs), 99.99% were assigned to 591 OTUs in the *Gallionella* genus. While the high number of *Gallionella* OTUs indicates that there was some species diversity present in these samples, most of these were present in very low abundance and could also have been due

to noise from sequencing errors. One of the *Gallionella* OTUs was by far the most abundant, comprising 81% of all *Gallionella*-assigned reads: OTU 830064 (Figure 13). This OTU was most abundant in Cu-rich Sites B and C, where it accounted for 76 - 83% of reads, but it was also seen at low levels in all other samples (0.2 - 4%). In contrast, the reads assigned to *Gallionella* in Site A were mostly from a different OTU: OTU 4342654, which was the second most abundant *Gallionella* OTU overall. This OTU was also present across all other sampling sites but in very low abundance (0.02% - 0.1%).

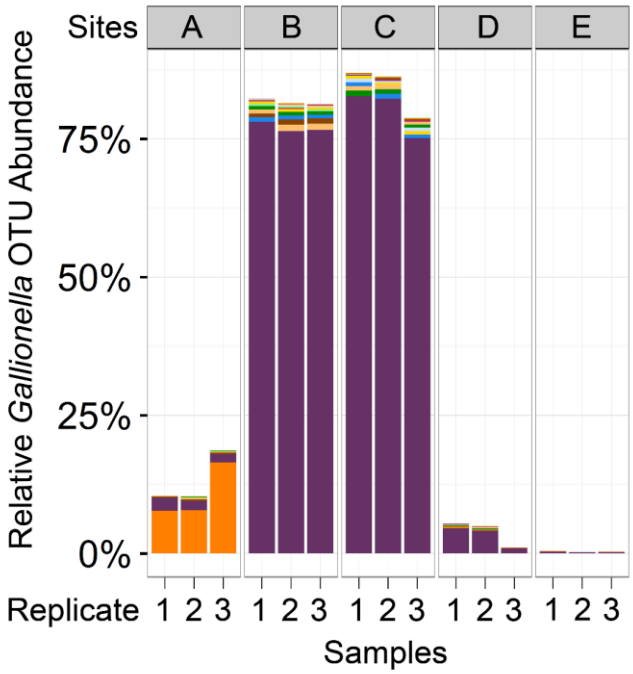

Figure 13 Abundance of Gallionella OTUs across samples
Bar plots represent the number of reads assigned to all OTUs classified as *Gallionella*. Each color represents a different OTU. The two most abundant *Gallionella* OTUs (brown and peach) represented 87% of all reads assigned to any *Gallionella* OTU. Purple = OTU 830064; orange = OTU 4342654 (Greengenes IDs).

To identify OTUs with abundances correlated with Cu concentration, OTUs with low variance in abundance across Sites A - E were first discarded, leaving 227 OTUs to be tested. Six of these OTUs were significantly correlated with Cu concentration and had at least 1% abundance in any sample: four belong to the *Gallionella* genus, one was unclassifiable, although similar to a sequence found in an arctic stream (Larouche *et al.*, 2012), and one was from the order *Stramenopiles* (Table 9).

Table 9 OTUs positively correlated with copper concentration are from Gallionellaceae and Stramenopiles.

OTU Name	Pearson's Rho	Abund-ance %	q-value	Phylum	Class	Order	Family	Genus
830064	0.81	82.7	3.70E-03	Proteobacteria	Betaproteobacteria	Gallionellales	Gallionellaceae	Gallionella
808810	0.93	6.1	1.30E-05	Cyanobacteria / Stramenopiles (Chloroplast)				
NCR.508055	0.96	1.3	1.40E-06	Unassigned. 88% identical to two uncultured bacterium clones from an arctic stream: EpiUMB50 (GenBank: FJ849304.1) & EpiUMB1 (GenBank: FJ849261.1).				
NCR.467936	0.89	1.1	1.90E-04	Proteobacteria	Betaproteobacteria	Gallionellales	Gallionellaceae	Gallionella
587098	0.69	1	3.80E-02	Proteobacteria	Betaproteobacteria	Gallionellales	Gallionellaceae	Gallionella
259765	0.93	1	1.40E-05	Proteobacteria	Betaproteobacteria	Gallionellales	Gallionellaceae	Gallionella

List of OTUs with abundance levels significantly correlated with copper concentration according to Pearson's correlation with Benjamini-Hochberg false discovery rate correction (significant: $q \leq 0.05$) and with at least 1% abundance. OTUs are sorted by abundance, measured as the maximum percent of reads that belong to this OTU within any sample. OTU names are Greengenes database identifiers unless beginning with NCR, which are novel OTU clusters. All OTUs, except NCR.508055, from Kingdom: Bacteria.

While 16S data indicates that one OTU belonging to the genus *Gallionella* dominates Site B, metagenomic sequencing results from Site B indicate that DNA present is most closely related to sequences from multiple genera in the *Gallionellaceae* family (Figure 14). Reads assigned in this family at the species level are split between three genera: *Gallionella*, with 62.5% of reads, *Sideroxydans*, with 34% of reads, and *Ferriphaselus*, with 3.5% of reads. *De novo* assembly of the metagenome from Site B resulted in an N50 of 835. Although this is low, many long contigs were constructed: 4 longer than 100 kbp and 558 longer than 10 kbp. Further sequencing that targets genome completion may be required to improve this assembly.

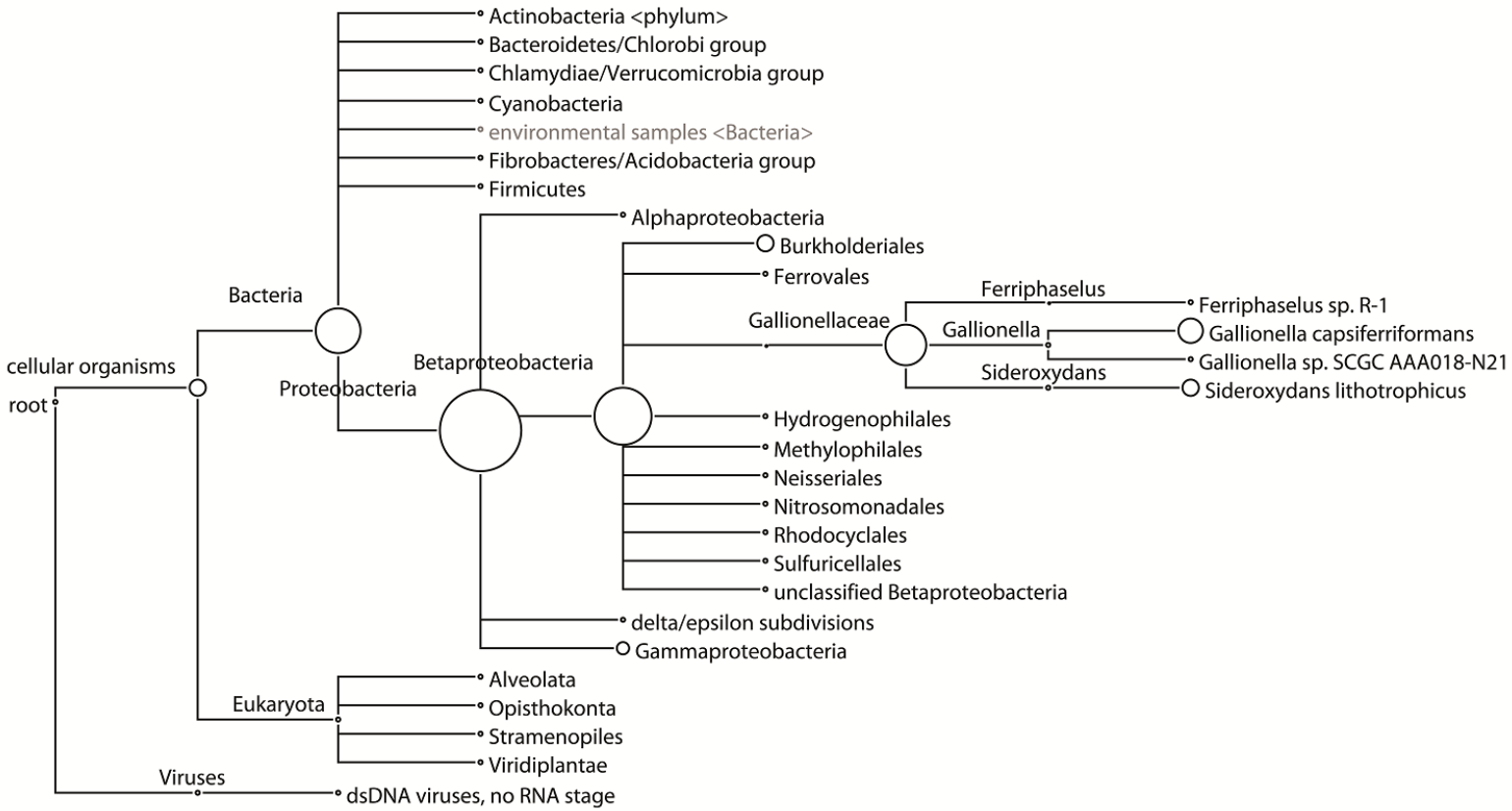

Figure 14 Phylogenetic tree of taxa at Site B predicted from shotgun sequencing analysis.
Circle size represents the number of reads assigned to each taxon, showing predominance of sequences classified as multiple genera in the Gallionellaceae family. Taxa only shown when at least 0.1% of reads are assigned.

5.5.2. Bacterial communities recover in phylogenetic diversity but not composition after Cu is depleted

Within-sample bacterial diversity was measured on OTU counts using the Shannon–Wiener index, which takes into account richness and abundance. Biofilm samples from Sites B and C (high total Cu concentrations: 1.02 mg/L and 0.37 mg/L, respectively) had the lowest diversity, while biofilm samples from Sites A, D and E (low total Cu concentrations: 0.12 mg/L, 0.03 mg/L and 0.07 mg/L, respectively) had higher diversity (Figure 15A). Between-sample bacterial diversity was calculated using weighted-UniFrac distances to measure diversity between samples. This method is phylogenetically sensitive, such that it measures the distance between samples based both on the abundance and relatedness of shared and unshared OTUs. The pair-wise distance matrix is represented in Figure 15B using principle coordinates analysis (PCoA). Samples clustered closer together had more similar microbiomes. Samples from the Cu-affected Sites B and C clustered together, as did samples from the two downstream Sites D and E where Cu had been depleted, while the samples from the site before Cu was introduced (Site A) were distinct from both groups (Sites B and C, or D and E). This suggested that while diversity levels recovered after Cu concentrations returned to baseline values, the composition of the community did not return to a pre-Cu exposure state. However, other factors that were not evaluated in detail in this study (e.g. stream topography or substrate type) may account for the microbiome differences between sampling locations A and D or E.

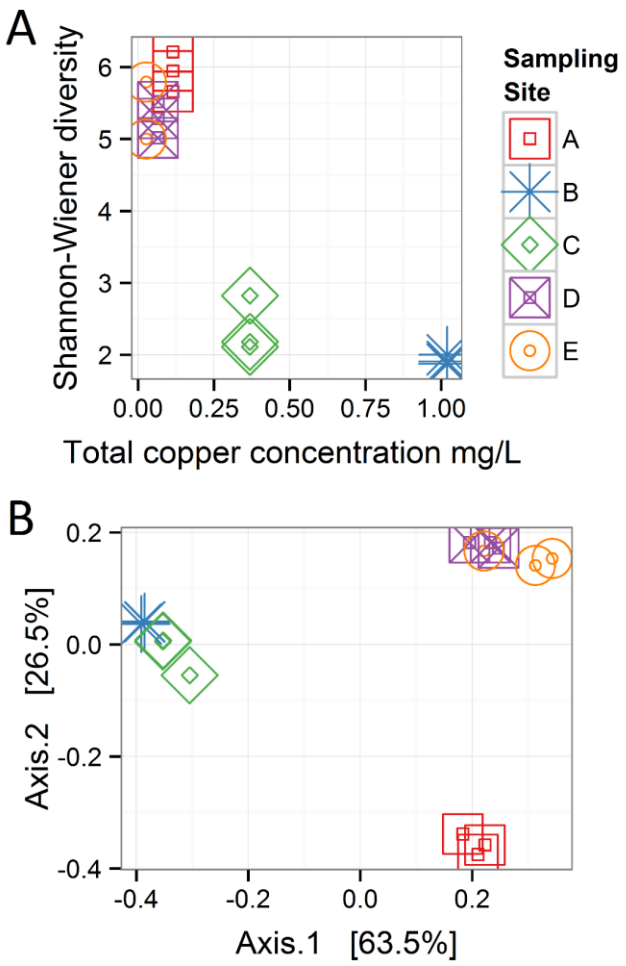

Figure 15 Bacterial diversity across sampling sites.
(A) Within-sample diversity (α-diversity) is measured with the Shannon-Wiener index. (B) Between-sample diversity (β-diversity) is measured with weighted UniFrac distances and represented with a principle coordinates analysis plot, in which 63.5% and 26.5% of the variability among samples is projected on Axis 1 and 2, respectively; n = 3 for all sampling sites. These plots show when copper concentrations were high (Sites B and C), α diversity was low, and bacterial communities recover in phylogenetic diversity after copper concentrations decline, but community composition does not return to the pre-copper state.

5.5.3. Eukaryotic community also differs between sites with low and high Cu concentrations

The taxonomic analysis of metagenomic sequences showed that bacteria were the main component of the biofilms examined and that Sites A and B differed in their eukaryotic and archaeal compositions (Figure 14; list of taxa abundances in Suppl. File 1). In Site A, 52% of reads were classified taxonomically. Reads may not have been classified due to low sequence complexity or high evolutionary distance from reference genomes. Of those reads that were classified, 99% were classified as bacterial, 1% as

eukaryotic, and 0.5% as archaeal. The eukaryotes were mostly *Opisthokonta* (0.6%), including *Metazoa* (0.3%) and Fungi (0.2%). The archaea were mostly *Euryarchaeota* (0.2%), with few reads classified down to the class level; the most abundant classes were *Methanomicrobia* (0.08%) and *Halobacteria* (0.04%). Viruses were detected at very low abundance (0.07%), as expected due to the pore-size of the filters used. In Site B, 61% of reads were classified taxonomically, of which 96% were bacterial, 4% were eukaryotic, and very few were archaeal (0.06%). The most abundant eukaryotic taxa identified in Site B were *Opisthokonta* (1.7%), including Fungi (1%) and *Metazoa* (0.5%), and two types of algae: *Stramenopiles* (0.5%) and *Viridiplantae* (0.4%). Viruses were again detected at low abundance (0.2%), half of which were *Phycodnaviridae*, which infect aquatic eukaryotic algae. These results show that Sites A and B differed beyond their bacterial communities, with archaea and *Metazoa* more abundant in Site A, while Fungi, algae, and algal viruses were more abundant in Site B.

5.5.4. Metal-associated genes and pathway-level differences occur between sites with high and low Cu concentrations

More reads from the Cu-rich Site B than from Site A were assigned to metal-associated genes, with 0.7% and 0.09% assigned, respectively (Suppl. File 3). Of the 22 metal-associated genes with at least 10 reads per million assigned in either sample, all were more abundant in Site B compared to Site A, with 9 having at least a 20-fold higher abundance (Table 10). Of the remaining 13 lower abundance genes, 12 had at least a 2-fold higher abundance in Site B compared to Site A (Suppl. File 3).

Table 10 Metal-associated genes with higher predicted abundance in metal-rich Site B over Site A.

Gene name	BacMet Annotation	
	Associated Metal	**Description**
chrB	Chromium (Cr)	Part of *chrBACF* operon from the transposable element TnOtChr. Has a regulatory role for expression of the *ChrA* transporter.
dpsA	Iron (Fe)	*dpsA* responses to stress, iron-binding protein; has role in iron ion homeostasis.
cznA	Cadmium (Cd), Zinc (Zn), Nickel (Ni)	Cadmium-zinc-nickel resistance protein *cznA*.
pstA	Arsenic (As)	An integral membrane protein. Part of the binding-protein-dependent transport system for phosphate; probably responsible for the translocation of the substrate across the membrane.
chrA	Chromium (Cr)	Part of *chrBACF* operon from the transposable element TnOtChr. It pumps chromate out of the cell.
mgtA	Cobalt (Co), Magnesium (Mg)	Magnesium-transporting ATPase, P-type 1; Mediates magnesium influx to the cytosol.
mexI	Vanadium (V)	RND Multidrug efflux transporter *mexI*; Part of *mexGHI-ompD* efflux pump operon. The efflux pump also involved in N-acyl-homoserine lactones (AHLs) homeostsis in *Pseudomonas aeruginosa*.
cusA / ybdE	Copper (Cu), Silver (Ag)	Part of a cation efflux system (*CusA, CusB, CusC* and *CusF*) that mediates resistance to copper and silver; located in cell inner membrane; Belongs to the *AcrB/AcrD/AcrF* family.
pstS	Arsenic (As)	A periplasmic phosphate binding protein; Part of the ABC transporter complex *PstSACB* involved in phosphate import.

Abundances were normalized by sequencing depth. Genes listed have at least 10 reads per million assigned in Site B and a 20-fold increase at Site B over Site A. Gene *chrB* had over 250-fold increase.

More reads from Site B were assigned to a SEED subsystem functional group than from Site A, with 27% and 18% of reads assigned, respectively (Suppl. File 2). This was likely due to fewer closely-related reference genomes being available for the upstream site. Table 11 lists the 20 SEED subgroups that were abundant, with at least 0.1% reads assigned, and differential, with at least a four-fold difference in abundance between Sites A and B. Most of these subgroups were more abundant in Site B than Site A, with a larger ratio than could be accounted for by the difference in number of reads classified (1.5-fold difference). The subgroup with the greatest fold change between sites was CRISPRs, followed by 'Type III, Type IV, Type VI, ESAT secretion systems'.

Table 11 SEED subgroups that differed in abundance between Sites A and B.

SEED Subgroup (level 2)	Belongs to SEED Group (level 1)	More abundant in Site
CRISPRs	DNA Metabolism	B
Type III, Type IV, Type VI, ESAT secretion systems	Virulence	B
Polyhydroxybutyrate metabolism	Fatty Acids, Lipids, and Isoprenoids	A
Denitrification	Nitrogen Metabolism	B
Bacterial hemoglobins	Stress Response	B
Glutathione-regulated potassium-efflux system and associated functions	Potassium Metabolism	B
Restriction-Modification System	DNA Metabolism	B
Orphan regulatory proteins	Regulation and Cell Signalling	B

All SEED level 2 subgroups that had at least 0.1% abundance and at least a 4-fold change between Sites A and B, sorted by fold change. Numbers omitted due to lack of statistical power.

5.6. Discussion

Microbial communities are being intensively studied for their environmental applications, including bioremediation to reduce undesirable contaminant concentrations that result from anthropogenic activities. During the environmental assessment of a proposed metals mine in Northern Canada, an unusual feature of the area was discovered: aqueous Cu concentrations that were high in one reach of a creek were quickly reduced downstream with data suggesting that the causative factor was biological. In this study, microbiome samples were collected from several reaches of the creek, including those high in metal concentrations and low in pH, for metagenomic analysis in order to characterize their taxonomic and genetic compositions. Microbial profiles at sampling locations with high Cu concentrations were highly similar and dominated by the *Gallionellaceae* bacterial family, suggesting that one or more species of *Gallionellaceae* may be involved in the observed reduction in aqueous Cu concentrations. The findings of this study provides further evidence that *Gallionella*-like species can thrive in environments that are acidic and high in metal concentrations, as previously described (Fabisch *et al.*, 2013). This study contributes to the characterization of native microorganisms that can potentially be cultivated for *in situ* bioremediation.

Metagenomic sequencing of DNA from biofilm samples from Sites A and B revealed that these communities were vastly dominated by bacteria, with > 99% and > 96% of taxonomically classified reads assigned to bacteria, respectively; thus, eukaryotic and archaeal proportions were minimal (Suppl. File 1). Fungi are known to sequester or uptake metals by biosorption but since fungi were proportionally quite low in the Site A and Site B biofilms (0.2% and 1%, respectively), their contribution to Cu sequestration/transformation was assumed to be low. The same assumption may be true for archaea and algae found at Site B (0.06% and 0.9%, respectively). Thus, due to the dominance of bacteria in the biofilms obtained, these results suggest that the bacteria present at Site B and/or Site C (which was not evaluated for its proportions of eukaryotes, archaea and bacteria) are the most likely candidates for the removal of aqueous Cu from the water column, if bioremediation was occurring. A low proportion of eukaryotes in an acidic, alkaline, and/or metals-rich environment does not appear to be unusual, although studies of these environments often focus on microbial community structures and thus the eukaryotic proportion is not evaluated or reported (Bier *et al.*, 2015; Tsitko *et al.*, 2014; Liljeqvist *et al.*, 2015). It is possible that the biofilm collection period of 18 days may have affected the final composition of the communities through succession limitations (i.e. limited colonization of eukaryotes), or the structure of the biofilm sampling devices may also have had an effect on colonization, although a similar apparatus has been used successfully elsewhere (21 day collection) to study biofilm development in freshwater lotic environments (Besemer *et al.*, 2012). Further investigation of the riverbed biofilms at Sites B and C will be needed to confirm the bacterial dominance and microbial contributions to Cu sequestration/transformation.

The bacterial communities present at Cu-affected Sites B and C were dominated by one OTU belonging to the genus *Gallionella* (Figure 13). However, the metagenomic sequencing results from Site B showed reads assigned to *Gallionella capsiferriformans* ES-2, *Sideroxydans lithotrophicus* ES-1, and *Ferriphaselus sp.* R-1 (Figure 14). *G. capsiferriformans* ES-2 and *S. lithotrophicus* ES-1 are the only complete genomes in the *Gallionellaceae* family and were recently sequenced (Emerson *et al.*, 2013). The 16S genes for these two species are quite different (93% sequence identity, which is well below the 97% OTU threshold used in this study) but the genomes are fairly similar, with 40% of genes homologous at 60% identity (Emerson *et al.*, 2013). This supports assigning the bacteria at Sites B and C as belonging to the *Gallionellaceae* family;

however, the *Gallionella*-specific label may be incorrect. The contradiction between the 16S rRNA results and the metagenomic sequencing results was informative, suggesting that either the dominant OTU observed from Site B does belong to *Gallionella* and that *G. capsiferriformans* ES-2 has undergone considerable gene-loss since *Gallionella* diverged from the other *Gallionellaceae* genera, or that the 16S rRNA-driven genus classification is unreliable in this clade, in which case this OTU may represent a new genus in *Gallionellaceae* that is related to both *G. capsiferriformans* ES-2 and *S. lithotrophicus* ES-1. Future studies could focus on identifying and isolating this member of *Gallionellaceae*, as well as the distinctly different member of the *Gallionellaceae* family found upstream at Site A (Figure 13). Comparisons of the genomes of these two *Gallionellaceae* populations may provide insights into microbial adaptations for extreme environments. Furthermore, isolating and culturing the *Gallionellaceae* from Site B would allow for exploration of bioremediative potential, which could result in the development of a new product for remediation activities.

From the data collected in this study, hypotheses can be drawn about potential bioremediation mechanisms that may be occurring. In general, bacteria can manage metals through active entrapment (biosorption or sequestration) and/or metabolic transformation into precipitates (Malik, 2004; Andreazza *et al.*, 2010; Ghosh and Saha, 2013), some of which bind to microbial surfaces via adsorption (Gadd, 2010). Biosorption and metal precipitation can even co-occur, making it difficult to determine the contribution of each process to metal immobilization (Glasauer *et al.*, 2001). Microbial transformational processes that alter the speciation of metals (e.g. from Fe^{2+} to Fe^{3+}) may achieve detoxification, enabling survival (Perales-Vela *et al.*, 2006; Orell *et al.*, 2010). Such transformations can result in the production of insoluble products, such as the orange-coloured iron (III) oxide precipitate. When iron oxide production appears to be the dominant detoxification process occurring, studies have shown that heavy metals can be co-precipitated by bacteria, including *Gallionellaceae* (Cu, cadmium, nickel and zinc (Fabisch *et al.*, 2013); manganese (Akob *et al.*, 2014)). The presence of *Gallionellaceae* and orange staining at Sites B and C indicate that copper could be co-precipitating with iron oxides. While sediments and precipitates (i.e. crystalline structures of minerals) were not examined for the presence of other metal oxides in this study, this could be performed in the future. In addition, riverbed biofilm samples could be evaluated in future studies to determine if Cu co-locates with biofilm mass.

Examining the gene content of the biofilm collected from Site B suggests alternative hypotheses to explain the lowered aqueous Cu concentration downstream. One of the most well-known natural precipitation mechanisms for dissolved Cu is performed by sulphate-reducing bacteria, which precipitate metals as highly insoluble sulfides. For example, sulphate-reducing bacterial biofilms have been shown to accumulate Cu on their surfaces as metal-sulfides (White and Gadd, 2000). More specifically, *Desulfovibrio* isolates have been shown to reduce thiosulphate (Bang *et al.*, 2000) and to precipitate Cu^{2+} as covellite (CuS) and chalcocite (Cu_2S) minerals (Karnachuk *et al.*, 2008). Within the *Gallionellaceae* family, *S. lithotrophicus* ES-1 has also been shown to reduce thiosulphate, while *G. capsiferriformans* ES-2 and all other known species in the *Gallionellales* order cannot (Emerson *et al.*, 2013). Sulphate reduction in *S. lithotrophicus* ES-1 is thought to be associated with the *soxXYZAB* operon; genes from this operon were predicted in DNA sequences from Sites A and B. Sulphate-reducing bacteria are typically anaerobic and Cu-accumulating biofilms have been cultured anaerobically (White and Gadd, 2000). While sulphate is present at elevated concentrations in the Casino Creek system, this system as a whole is not anaerobic (dissolved oxygen saturation > 87%; Table 7). However, the complex inner-structure of biofilms has been shown to support sulphate-reducing bacteria even in aerobic water (Okabe *et al.*, 1999). This suggests that the *Gallionellaceae* bacteria identified in this study could be precipitating copper through sulphate reduction. The presence of the *soxXYZAB* operon further supports that the most abundant OTU observed at Site B is a novel species in *Gallionellaceae* that has a closer 16S sequence similarity to *G. capsiferriformans* ES-2 but key genetic and thus metabolic similarities to *S. lithotrophicus* ES-1. Another major mechanism by which bacteria can immobilize metals is through intracellular binding with polyphosphate bodies, a process that has been observed in cyanobacteria (Malik, 2004). Although this has not been observed in *Gallionellaceae*, a high abundance of two genes from the *pstSACB* operon (*pstA* and *pstS*), an ATP binding cassette (ABC) phosphate transporter, were found in the metals-rich Site B (Table 8). However, this operon has also been associated with arsenate transport (Li *et al.*, 2013) and so these bacteria could be equipped with these genes either to import phosphate, tolerate arsenate, or for some other unknown reason. The mechanistic hypotheses described above are only suggestive of possible processes at this point in time. Additional sequencing of the microbial community transcriptome

(metatranscriptomics) under controlled conditions could give insight into the specific immobilization or transformational mechanisms involved.

Metagenome sequencing of the biofilm sample from Site B gave insight into the metals-tolerance genes in these organisms. The most abundant metal-associated gene identified in the Site B biofilm was *cusA*/*ybdB*, which is involved in a cation efflux system that mediates resistance to Cu and silver (Table 10). The second most abundant metal-associated gene observed was *mdtC*, which is involved in zinc efflux. Zinc was not identified during the mine development process as a current or future metal of environmental concern but it is present in the water column at relatively low levels. Potentially, the high abundance of this gene relates to its involvement as an efflux pump for more than one cationic metal. Since genes were counted as present with a 60% identity threshold, these results more generally suggest that heavy−metal efflux ATPases were present in this metagenome. One of the best characterized P-type ATPases related to Cu resistance, *copA* (Orell *et al.*, 2010), was detected only at very low abundance in this sample, which was expected due to the lack of annotated *copA* genes in the two available sequenced *Gallionellaceae* genomes. The metal-associated gene with the largest fold change between Site B and Site A was *chrB* (Table 10), which belongs to the *chrBACF* operon and has been shown to be involved in resistance to high concentrations of chromium (Branco *et al.*, 2008). *ChrA*, which also belongs to this operon, also had a high fold change at Site B over Site A (Table 9). It is possible that the abundance of these two *chr* genes was not indicating chromium or chromate resistance, but rather a resistance against harmful reactive oxygen species that can be produced during the detoxification of other metals (e.g. Cu[I]) because genes in the *chrBACF* operon have also been shown to enable resistance to superoxide anions (Branco *et al.*, 2008). Genes from the *chrBACF* operon have not been annotated in *G. capsiferriformans* or *S. lithotrophicus* but *chrA* is present in *Ferriphaselus* sp. R-1, which is an incomplete genome from the *Gallionellaceae* family. Predicted protein sequences similar to *chrA* from *F.* sp. R-1 using BlastP were searched in the *G. capsiferriformans* and *S. lithotrophicus* genomes, but good quality matches were not found (data not shown, best hit: 46% protein sequence identity over 11% of the query amino acid sequence). This further supports our hypothesis that this study has identified a new genus within *Gallionellaceae*, one that has a unique profile of metals tolerance genes, enabling adaptation to this environment. Overall, the metagenomic sequencing results

identified numerous well-known metal-related tolerance genes; thus, the Site B biofilm was resistant to the heavy metal cations in the water and is likely well situated to tolerate conditions in a mining operations environment.

This study represents a first stage investigation at a future mining site into the native microbial communities present in a stream affected by acidic and high metals concentration conditions. It is an important step in characterizing the diversity of microbes adapted to extreme environments and the naturally-evolved microbial potential for beneficial and effective future uses, before mining activities begin and disturb the natural environment. These findings provide important baseline data regarding microbes present and the effect of metals on the community composition, as well as supporting future sampling and investigations of their potential utility in bioremediation.

5.7. Conclusions

A native microbial community from a main watercourse in a proposed mining operation in Yukon, Canada is dominated by bacteria in the family *Gallionellaceae*. This microbiome tolerates elevated heavy metals concentrations and may be providing beneficial bioremedial actions by reducing aqueous Cu concentrations, a metal of environmental concern. The 16S rRNA gene and metagenomic investigations in this study have identified the taxonomic and functional profiles of key microbial communities and may enable the assembly of one or more genomes in the *Gallionellaceae* family from this unique site. The shotgun sequencing of DNA from biofilm samples provides insight into the mechanisms that may be providing resistance to heavy metals, by the identification of genes that encode for efflux pumps, cell wall components, and metabolic processes for metals tolerance. This study provides important baseline data that indicates a potential microbial component in a case of natural Cu depletion. It provides support for future studies to specifically investigate the biofilm community (or the most abundant bacteria within) that may be responsible for this transformation. The methods and findings are relevant for future investigations into the potential use of a native biofilm to bioremediate contaminants and reduce adverse effects in aquatic organisms. This study is one of the first investigations to profile the taxa and genes in a stream microbiome from the Yukon, Canada, an area rich in valuable metals, and therefore provides support for investigations on novel organisms useful to the mining industry and for environmental protection.

5.8. Author contributions

TV analyzed results with bioinformatics techniques, wrote parts of Materials and Methods and Discussion and majority of Results Sections of the manuscript, and prepared figures and tables. HO aided with method of sample collection, contributed to interpretation of results, and wrote portions of the manuscript. MP, EG, and RL performed laboratory work: DNA extractions, library preparations, and sequencing runs. MP did background research and wrote some of the manuscript. NL, MQ, and RP aided with study design and performed field collection of samples. FB supervised TV, EG, and RL, contributed to interpretation of results, and edited the manuscript. CK supervised HO and MP, designed study, contributed to interpretation of results, and edited the manuscript.

5.9. Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Casino Mining Corporation (CMC) provided supportive co-funding for the GenomeBC study. The findings and conclusions presented by the authors are their own and do not necessarily reflect the view or position of CMC.

5.10. Acknowledgements

This study was funded by a GenomeBC grant #UPP004 to CJK in partnership with the CMC and Palmer Environmental Consulting Group (PECG). We acknowledge additional funding to TVR from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Molecular Biology and Biochemistry Department of Simon Fraser University (SFU; Burnaby, BC, Canada), and to MMP from the Biological Sciences Department of SFU. Help from Sydney Love and other staff and students in the Brinkman and Kennedy laboratories at SFU was greatly appreciated, as well as from Dr. P. Tang and Dr. M. Uyaguari-Diaz from the BC Centre for Disease Control (Vancouver, BC, Canada) and Dr. K. Besemer from the University of Vienna (Vienna, Austria).

Chapter 6.

Discussion and concluding remarks

6.1. Strengths of DNA-sequencing-based approaches for advances in environmental monitoring

In the last decade, microbiome research has exploded in popularity. The important role of microbial communities has been recognised in fields across human health, agriculture, fisheries, food production, industry, and environmental monitoring. Traditional culture-based methods are well established in environmental monitoring and contribute many valuable findings in the field. However, microbiome DNA-sequencing based approaches (phylogenetic marker gene and metagenomic sequencing) have important advantages that address some limitations of traditional culture-based methods. For example, microbiome DNA-sequencing based approaches produce less biased profiles of microbial communities, allowing a more complete characterisation of the effect of contaminants and changing environmental conditions on microbial communities as a whole. Compared to gel-based phylogenetic methods, sequencing-based methods can provide a higher resolution picture of the microbial community profile, since OTUs can be made that are more taxonomically specific than gel bands. While gel-based methods can be followed up with band excision and DNA sequencing, amplicon sequencing methods provide this data from the start for the full sample. Research with complementary culture-based and metagenomic approaches hold promise to increase our understanding of microbial communities.

The richer set of taxa and genes (“features”) that sequencing-based methods provide is useful for identifying novel targets in environmental monitoring. Instead of relying on established biological knowledge, which may be lacking in poorly characterised environments, these more complete feature profiles allow the use of data-driven methods for discovery (e.g. machine learning). These methods can identify features that may be previously unknown or would not have been hypothesised to be associated with the conditions under study. Because the DNA sequences are available for these newly identified features, PCR primers can then be designed for their detection. In theory, the specificity of these features can also allow for “source tracking”,

in which genes or taxa are identified that should only be present in a particular source (Harwood *et al.*, 2014). For example, if a gene is only known to occur in cow feces and is then observed in a river, the inference can be made that the water has been contaminated with cow feces. While this is an attractive opportunity to take advantage of DNA sequences from microbial communities and has had some success, it is difficult to prove that a feature only exists in one microbial source when many microbial sources are uncharacterised and difficult to reach a sufficient level of detection effort to reasonably "prove" absence (Harwood *et al.*, 2014; Gomi *et al.*, 2014; Panasiuk *et al.*, 2015).

6.2. Limitations of DNA-sequencing-based approaches for advances in environmental monitoring

Despite the advantages of DNA-sequencing-based approaches for microbiome research in environmental monitoring, there are some important limitations to consider. These limitations exist in the sample collection, preparation, sequencing and analysis stages. Since my research has focused on the analysis phase, I will limit my discussion below accordingly.

Many microbiome studies are concerned with the bacteria associated with human stool, and many analysis tools have been developed to be used on these microbiomes. Bacteria associated with human stool are unlike many other microbial communities because they are fairly well represented by reference genomic databases (Goodrich *et al.*, 2014; Langille *et al.*, 2013). As such, many analysis strategies developed for human stool microbiome analysis assume that many of the microbial community taxa or genes will be similar to something in a database. When these approaches are applied in environments that lack good representation in reference databases, the inappropriateness of this assumption can produce misleading results, for example through inappropriate normalisations (Chapter 2) or through poorly supported taxonomic inferences (Langille *et al.*, 2013). To address these issues, careful use of tools is required and further effort to produce more reference genome sequences will be advantageous.

Most metagenomic projects use Illumina sequencing, which yields DNA reads of limited length that decrease in accuracy towards the end of each read. Though this

limited length has increased from 75 bp to now 300 bp and can be effectively near 600 bp with paired-end sequencing (Illumina), this is still is shorter than the median length of bacterial genes (780 bp (Brocchieri and Karlin, 2005)) and reads are unlikely to coincide with a full gene sequence region. This short read length means that if full gene, operon, or genomic sequences are of interest, it can be necessary to assemble reads. Assembly of metagenomic data can have variable success due to the complexity of the metagenome and depth of coverage. This variable success rate can introduce bias into metagenomic comparative analyses. While recently there is an effort to quantify and compare the quality of metagenomic assemblies (Vollmers *et al.*, 2017), in many analysis pipelines this is not done or not reported. Traditional measures of assembly quality were established for single-genome assemblies and focus on reporting the number and length of contigs (e.g. N50). While these metrics are relevant in metagenomes, they do not consider an important metric: the proportion of the metagenome's reads that are included in the assembly. If the proportion of reads assembled into contigs is small, then the contigs produced only represent a small, biased proportion of the data. This can lead to downstream analyses of these contigs producing biased results. This can be the case, for example, in approaches based on clustering genes predicted from contigs. Though these methods are useful due to their lack of reliance on reference data, their reliance on assembled data can be problematic. When assembled contigs are poor representations of the original data, then assembly free approaches, such as k-mer analysis, are more appropriate.

One major limitation of the common DNA sequencing approaches to study microbiomes (phylogenetic marker gene and metagenomic sequencing) is that they are genome-centric and so do not consider microbial activity. Metagenomic profiles will describe the genes present and their abundances, but not which genes are being transcribed. The most straightforward cause for an increase in the abundance of a gene in a metagenomic profile is due to an increase in the relative abundance of the microbes that have this gene. Whether this change in microbial abundance is related to the function of the gene in question, however, is not certain. A lack of activity information can also affect the interpretation of taxonomic profiles. Many microbes in the environment are dormant (e.g. approximately 80% on average in soil, 40% in freshwater lakes) (Lennon and Jones, 2011) and microbial taxa profiles cannot distinguish between active and dormant cells. Because dormant cells are less affected by local conditions,

their presence in a profile can dilute the signal from a microbial response to local conditions. This can lead to a consistent taxonomic profile across samples, even while the microbial communities are different in terms of the taxa that are active.

Some of these limitations may be addressed with emerging developments in metagenomics and complementary approaches. For example, the limitations caused by short reads requiring assembly may be improved by new technologies that produce longer reads, with lengths routinely from 10 to 100 kbp and reportedly to almost 1Mbp (e.g. Pacific Biosciences ("PacBio"), MinION from Oxford Nanopore) (Mardis, 2017). These longer reads are promising as they may cover entire viral genomes and bacterial genes and operons. Typically, the number of reads produced by these technologies is lower than with Illumina short-read sequencing, meaning that the diversity of a microbial community will be less deeply sequenced. Due to this lower yield, long reads do not necessarily provide a universal improvement for microbiome DNA sequencing. In cases where high abundance genes or taxa are of interest, long reads can be beneficial for metagenomic studies. In phylogenetic marker gene sequencing studies, when high resolution of low complexity communities is desired, there may be a benefit to using long sequences. Full length 16S rRNA gene sequences may provide better resolution than the smaller highly variable regions (Singer *et al.*, 2016); however, the benefit of full length sequences is still under investigation (Wagner *et al.*, 2016). Long reads could also allow the use of longer phylogenetic marker gene sequences. A challenge with current "long read methods" is that they require more DNA starting material than Illumina sequencing methods.

The limitations caused by the lack of information about activity can be lessened using complementary approaches that profile the molecules produced by microbial activity (e.g. transcriptomics, proteomics, and metabolomics). There are also techniques that can estimate activity of bacterial species by comparing the depth of coverage of a genome at the origin of replication versus the opposite end or overall coverage (Korem *et al.*, 2015; Brown *et al.*, 2016). However, these approaches require high depth of coverage and/or mostly complete reference genomes and so cannot be used in many cases.

6.3. How this work could inform development of biomonitoring strategies

The findings presented in this thesis contribute to basic microbial ecology, which may support future research and development of new biomonitoring strategies. In addition to demonstrating taxon- and gene-family-based methods to develop potential biomarkers from metagenomic data, this work also provides some general findings that inform future development strategies.

I showed that k-mer profiles can be a useful approach in the analysis of communities with limited coverage and limited representation in reference databases. This approach was useful in initial analyses to identify outliers and in more specific analyses of community similarities (Chapter 2, Chapter 4). However, a k-mer based analysis translates less directly to the identification of specific biomarker DNA sequences than a gene-based approach. Despite this limitation, the k-mer approach provided potentially useful contextual findings for biomarker research, revealing that while bacteria, viruses, and microeukaryotes were temporally synchronous, they had different relationships with environmental conditions and with geography. For example, while viral biomarkers are promising due to their host-specificity and higher likelihood to represent active bacteria, the strong geographic clustering of viromes observed in Chapter 4 suggests the geographic generalisability of viral sequences as biomarkers may be limited. In summary, these results indicate that viral riverine biomarkers may have high sensitivity but not be generalizable across watersheds.

6.4. Future directions

This thesis presents foundational descriptions of riverine microbial community variability over time and water quality, including the first data describing the variability in composition of planktonic viral communities in rivers across time and space. Descriptive microbiome studies, like the ones presented above, can be built upon with controlled experiments or more targeted descriptive sampling. Controlled experiments at the ecosystem level are difficult, but can be done, for example with controlled exposures in natural yet relatively closed systems such as seasonal ponds (Carrino-Kyker *et al.*, 2013). Alternatively, the next phase could include following up with a more targeted sampling strategy. This approach could address the challenges in defining water quality

by focusing on a specific type of contamination of interest with well defined and characterised sources. Based on our findings above, after identifying a specific type of contamination to study, sampling across watersheds and seasons would be important.

References

Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, *et al.* (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* **8**: e1002358.

Aggarwal CC, Hinneburg A, Keim DA. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. 420–434.

Aguirre de Cárcer D, López-Bueno A, Pearce DA, Alcamí A. (2015). Biodiversity and distribution of polar freshwater DNA viruses. *Sci Adv* **1**: e1400127.

Akob DM, Bohu T, Beyer A, Schaffner F, Handel M, Johnson CA, *et al.* (2014). Identification of Mn(II)-Oxidizing Bacteria from a Low-pH Contaminated Former Uranium Mine. *Appl Environ Microbiol* **80**: 5086–5097.

Allan IJ, Vrana B, Greenwood R, Mills GA, Roig B, Gonzalez C. (2006). A 'toolbox' for biological and chemical monitoring requirements for the European Union's Water Framework Directive. *Talanta* **69**: 302–322.

Alshawaqfeh M, Bashaireh A, Serpedin E, Suchodolski J. (2017). Consistent metagenomic biomarker detection via robust PCA. *Biol Direct* **12**. e-pub ahead of print, doi: 10.1186/s13062-017-0175-4.

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

Amann RI, Ludwig W, Schleifer KH. (1995). Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiol Rev* **59**: 143–169.

Anderson MJ. (2005). PERMANOVA Permutational multivariate analysis of variance. *Austral Ecol* 1–24.

Andreazza R, Pieniz S, Benedict CO, Camargo F a O, Bento FM. (2010). Characterization of copper biosorption and bioreduction by copper resistant bacteria isolated from a vineyard soil. *Sci Total Environ* 27–30.

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, Salamon P, *et al.* (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* **6**: 41.

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, *et al.* (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5**: e1000593.

Arini A, Feurtet-Mazel A, Morin S, Maury-Brachet R, Coste M, Delmas F. (2012). Remediation of a watershed contaminated by heavy metals: A 2-year field biomonitoring of periphytic biofilms. *Sci Total Environ* **425**: 242–253.

Bailey TL, Bodén M, Whitington T, Machanick P. (2010). The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* **11**: 179.

Baird DJ, Hajibabaei M. (2012). Biomonitoring 2.0: A new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Mol Ecol* **21**: 2039–2044.

Baker BJ, Banfield JF. (2003). Microbial communities in acid mine drainage. *FEMS Microbiol Ecol* **44**: 139–152.

Baldwin BR, Nakatsu CH, Nies L. (2003). Detection and enumeration of aromatic oxygenase genes by multiplex and real-time PCR. *Appl Environ Microbiol* **69**: 3350–8.

Bang SW, Clark DS, Keasling JD. (2000). Engineering hydrogen sulfide production and cadmium removal by expression of the thiosulfate reductase gene (phsABC) from Salmonella enterica serovar typhimurium in Escherichia coli. *Appl Environ Microbiol* **66**: 3939–3944.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19**: 455–477.

Di Bella J.M. BY. GGB. BJP. RG. (2013). High throughput sequencing methods and analysis for microbiome. *J Microbiol Methods* **95**: 401–414.

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* **57**: 289–300.

Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, *et al.* (2016). Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Comput Sci* **2**: e94.

Besemer K, Peter H, Logue JB, Langenheder S, Lindström ES, Tranvik LJ, *et al.* (2012). Unraveling assembly of stream biofilm communities. *ISME J* **6**: 1459–1468.

Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ. (2010). Average genome size: a potential source of bias in comparative metagenomics. *ISME J* **4**: 1075–7.

Bier RL, Voss KA, Bernhardt ES. (2015). Bacterial community responses to a gradient of alkaline mountaintop mine drainage in Central Appalachian streams. *ISME J* **9**: 1378–1390.

Bikel S, Valdez-Lara A, Cornejo-Granados F, Rico K, Canizales-Quinteros S, Soberón X, *et al.* (2015). Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. *Comput Struct Biotechnol J* **13**: 390–401.

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–20.

Bradford TM, Morgan MJ, Lorenz Z, Hartley DM, Hardy CM, Oliver RL. (2013).

Microeukaryote community composition assessed by pyrosequencing is associated with light availability and phytoplankton primary production along a lowland river. *Freshw Biol* **58**: 2401–2413.

Branco R, Chung AP, Johnston T, Gurel V, Morais P, Zhitkovich A. (2008). The chromate-inducible chrBACF operon from the transposable element TnOtChr confers resistance to chromium(VI) and superoxide. *J Bacteriol* **190**: 6996–7003.

Brocchieri L, Karlin S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res* **33**: 3390–400.

Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. http://arxiv.org/abs/1203.4802 (Accessed May 11, 2017).

Brown CT, Olm MR, Thomas BC, Banfield JF. (2016). Measurement of bacterial replication rates in microbial communities. *Nat Biotech* **34**: 1256–1263.

Brum JR, Ignacio-espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, *et al.* (2015). Patterns and Ecological Drivers of Ocean Viral Communities. *Science (80-)* **348**: 1261498-1–11.

Brussaard CPD, Thyrhaug R, Marie D, Bratbak G. (1999). Flow cytometric analyses of viral infection in two marine phytoplankton species, micromonas pusilla (prasinophyceae) and phaeocystis pouchetii (prymnesiophyceae). *J Phycol* **35**: 941–948.

Buchfink B, Xie C, Huson DH. (2014). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Buttigieg PL, Ramette A. (2014). A guide to statistical analysis in microbial ecology: a community-focused, living review of multivariate data analyses. *FEMS Microbiol Ecol* **90**: 543–550.

Canadian Council of Ministers of the Environment. (2007). Canadian Environmental Quality Guidelines and Summary Table. http://ceqg-rcqe.ccme.ca/.

Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266–7.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–6.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A* **108 Suppl**: 4516–22.

De Cárcer DA, Pedrós-Alió C, Pearce DA, Alcamí A. (2016). Composition and interactions among bacterial, microeukaryotic, and T4-like viral assemblages in lakes

from both polar zones. *Front Microbiol* **7**. e-pub ahead of print, doi: 10.3389/fmicb.2016.00337.

Carrino-Kyker SR, Smemo KA, Burke DJ. (2013). Shotgun metagenomic analysis of metabolic diversity and microbial community structure in experimental vernal pools subjected to nitrate pulse. *BMC Microbiol* **13**: 78.

Cébron A, Norini M-P, Beguiristain T, Leyval C. (2008). Real-Time PCR quantification of PAH-ring hydroxylating dioxygenase (PAH-RHDα) genes from Gram positive and Gram negative bacteria in soil and sediment samples. *J Microbiol Methods* **73**: 148–159.

Chao A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scanadinavian J Stat* **11**: 265–270.

Chistoserdovai L. (2010). Functional metagenomics: recent advances and future challenges. *Biotechnol Genet Eng Rev* **26**: 335–52.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, *et al.* (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res* **42**: D633-42.

Crump BC, Amaral-Zettler LA, Kling GW. (2012). Microbial diversity in arctic freshwaters is structured by inoculation of microbes from soils. *ISME J* **6**: 1629–39.

Crump BC, Peterson BJ, Raymond PA, Amon RMW, Rinehart A, McClelland JW, *et al.* (2009). Circumpolar synchrony in big river bacterioplankton. *Proc Natl Acad Sci U S A* **106**: 21208–12.

Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, *et al.* (2015). The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research* **4**. e-pub ahead of print, doi: 10.12688/f1000research.6924.1.

Csárdi G, Nepusz T. (2006). The igraph software package for complex network research. *InterJournal, Complex Syst* **1695**: 1–9.

Dann LM, Rosales S, McKerral J, Paterson JS, Smith RJ, Jeffries TC, *et al.* (2016). Marine and giant viruses as indicators of a marine microbial community in a riverine system. *Microbiologyopen*. e-pub ahead of print, doi: 10.1002/mbo3.392.

Danovaro R, Luna GM, Dell'anno A, Pietrangeli B. (2006). Comparison of two fingerprinting techniques, terminal restriction fragment length polymorphism and automated ribosomal intergenic spacer analysis, for determination of bacterial diversity in aquatic environments. *Appl Environ Microbiol* **72**: 5982–9.

Deletic A. (1998). The first flush load of urban surface runoff. *Water Res* **32**: 2462–2470.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, *et al.* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–72.

Desrosiers M, Planas D, Mucci A. (2006). Total mercury and methylmercury

accumulation in periphyton of Boreal Shield Lakes: Influence of watershed physiographic characteristics. *Sci Total Environ* **355**: 247–258.

Devers M, Soulas G, Martin-Laurent F. (2004). Real-time reverse transcription PCR analysis of expression of atrazine catabolism genes in two bacterial strains isolated from soil. *J Microbiol Methods* **56**: 3–15.

Dunn G, Harris L, Cook C, Prystajecky N. (2014). A comparative analysis of current microbial water quality risk assessment and management practices in British Columbia and Ontario, Canada. *Sci Total Environ* **468**–**469**: 544–552.

Eaton AD, Franson MAH. (2005). Standard Methods for the Examination of Water & Wastewater. American Public Health Association.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–1.

Edgar RC. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**: 996–8.

Elser JJ, Bracken MES, Cleland EE, Gruner DS, Harpole WS, Hillebrand H, *et al.* (2007). Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol Lett* **10**: 1135–42.

Emerson D, Field EK, Chertkov O, Davenport KW, Goodwin L, Munk C, *et al.* (2013). Comparative genomics of freshwater Fe-oxidizing bacteria: Implications for physiology, ecology, and systematics. *Front Microbiol* **4**: 254.

Environment Canada. (2013). Historical climate data. climate.weather.gc.ca/ (Accessed January 1, 2014).

Espy MJ, Uhl JR, Sloan LM, Buckwalter SP, Jones MF, Vetter EA, *et al.* (2006). Real-time PCR in clinical microbiology: applications for routine laboratory testing. *Clin Microbiol Rev* **19**: 165–256.

Fabisch M, Beulig F, Akob DM, Küsel K. (2013). Surprising abundance of Gallionella-related iron oxidizers in creek sediments at pH 4.4 or at high heavy metal concentrations. *Front Microbiol* **4**: 1–12.

Faith DP. (1992). Conservation evaluation and phylogenetic diversity. *Biol Conserv* **61**: 1–10.

Field KG, Samadpour M. (2007). Fecal source tracking, the indicator paradigm, and managing water quality. *Water Res* **41**: 3517–3538.

Filée J, Tétart F, Suttle CA, Krisch HM. (2005). Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc Natl Acad Sci U S A* **102**: 12471–6.

Findlay S. (2010). Stream microbial ecology. *J North Am Benthol Soc* **29**: 170–181.

Fisher JC, Newton RJ, Dila DK, McLellan SL. (2015). Urban microbial ecology of a freshwater estuary of Lake Michigan. *Elem Sci Anthr* **3**: 64.

Fisher MM, Triplett EW. (1999). Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. *Appl Environ Microbiol* **65**: 4630–4636.

Fondriest Environmental I. (2013). pH of Water. *Fundam Environ Meas*. http://www.fondriest.com/environmental-measurements/parameters/water-quality/ph/#p7 (Accessed April 11, 2017).

Fondriest Environmental I. (2014). Water Temperature. *Fundam Environ Meas*. http://www.fondriest.com/environmental-measurements/parameters/water-quality/water-temperature/#watertemp1 (Accessed April 11, 2017).

Fortunato CS, Eiler A, Herfort L, Needoba JA, Peterson TD, Crump BC. (2013). Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. *ISME J* **7**: 1899–911.

Frank JA, Sørensen SJ. (2011). Quantitative metagenomic analyses based on average genome size normalization. *Appl Environ Microbiol* **77**: 2513–21.

Fruchterman TMJ, Reingold EM. (1991). Graph drawing by force-directed placement. *Softw Pract Exp* **21**: 1129–1164.

Fu L, Niu B, Zhu Z, Wu S, Li W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–2.

Gadd GM. (2010). Metals, minerals and microbes: Geomicrobiology and bioremediation. *Microbiology* **156**: 609–643.

García-Armisen T, İnceoğlu Ö, Ouattara NK, Anzil A, Verbanck MA, Brion N, *et al.* (2014). Seasonal variations and resilience of bacterial communities in a sewage polluted urban river. *PLoS One* **9**: e92579.

Garrido L, Sánchez O, Ferrera I, Tomàs N, Mas J. (2014). Dynamics of microbial diversity profiles in waters of different qualities. Approximation to an ecological quality indicator. *Sci Total Environ* **468**–**469**: 1154–61.

van der Gast CJ. (2015). Microbial biogeography: The end of the ubiquitous dispersal hypothesis? *Environ Microbiol* **17**: 544–546.

Ghai R, Rodrıguez-Valera F, McMahon KDK, Rodriguez-Valera F, McMahon KDK, Toyama D, *et al.* (2011). Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River Lopez-Garcia P (ed). *PLoS One* **6**: e23785.

Ghosh A, Saha P Das. (2013). Optimization of copper bioremediation by Stenotrophomonas maltophilia PD2. *J Environ Chem Eng* **1**: 159–163.

Glasauer S, Langley S, Beveridge TJ. (2001). Sorption of Fe (Hydr)Oxides to the Surface of Shewanella putrefaciens: Cell-Bound Fine-Grained Minerals Are Not Always

Formed de Novo. *Appl Environ Microbiol* **67**: 5544–5550.

Goldstein ST, Juranek DD, Ravenholt O, Hightower AW, Martin DG, Mesnik JL, *et al.* (1996). Cryptosporidiosis: an outbreak associated with drinking water despite state-of-the-art water treatment. *Ann Intern Med* **124**: 459–68.

Gomi R, Matsuda T, Matsui Y, Yoneda M. (2014). Fecal Source Tracking in Water by Next-Generation Sequencing Technologies Using Host-Specific *Escherichia coli* Genetic Markers. *Environ Sci Technol* **48**: 9616–9623.

Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, *et al.* (2014). Conducting a Microbiome Study. *Cell* **158**: 250–262.

Grenyer R, Rouget M, Davies TJJ, Cowling RM, Faith DP, Bank M Van Der, *et al.* (2007). Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* **445**: 757–760.

Griffith JF, Weisberg SB, Arnold BF, Cao Y, Schiff KC, Colford JM. (2016). Epidemiologic evaluation of multiple alternate microbial water quality monitoring indicators at three California beaches. *Water Res* **94**: 371–381.

Guo F-B, Lin H, Huang J. (2009). A plot of G + C content against sequence length of 640 bacterial chromosomes shows the points are widely scattered in the upper triangular area. *Chromosome Res* **17**: 359–64.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward D V, Giannoukos G, *et al.* (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**: 494–504.

Hahn AS, Hanson NW, Kim D, Konwar KM, Hallam SJ. (2015). Assembly independent functional annotation of short-read data using SOFA: Short-ORF functional annotation. In: *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, pp 1–6.

Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JBH. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**: 497–506.

Hanson NW, Konwar KM, Hallam SJ. (2016). LCA∗: An entropy-based measure for taxonomic assignment within assembled metagenomes. *Bioinformatics* **32**: 3535–3542.

Harwood VJ, Staley C, Badgley BD, Borges K, Korajkic A. (2014). Microbial source tracking markers for detection of fecal contamination in environmental waters: Relationships between pathogens and human health outcomes. *FEMS Microbiol Rev* **38**: 1–40.

Henderson G, Cox F, Kittelmann S, Miri VH, Zethof M, Noel SJ, *et al.* (2013). Effect of DNA Extraction Methods and Sampling Techniques on the Apparent Structure of Cow and Sheep Rumen Microbial Communities Bertilsson S (ed). *PLoS One* **8**: e74787.

Hu A, Yang X, Chen N, Hou L, Ma Y, Yu C-P. (2014). Response of bacterial

communities to environmental changes in a mesoscale subtropical watershed, Southeast China. *Sci Total Environ* **472**: 746–56.

Huang Y, Gilna P, Li W. (2009). Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**: 1338–1340.

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**: 4399–406.

Hurwitz BL, Sullivan MB. (2013). The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology Thompson F (ed). *PLoS One* **8**: e57355.

Hurwitz BL, U'Ren JM, Youens-Clark K. (2016). Computational prospecting the great viral unknown. *FEMS Microbiol Lett* **363**. e-pub ahead of print, doi: 10.1093/femsle/fnw077.

Hurwitz BL, Westveld AH, Brum JR, Sullivan MB. (2014). Modeling ecological drivers in marine viral communities using comparative metagenomics and network analyses. *Proc Natl Acad Sci* **111**: 10714–10719.

Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res* **21**: 1552–60.

Ibekwe AM, Ma J, Murinda SE. (2016). Bacterial community composition and structure in an Urban River impacted by different pollutant sources. *Sci Total Environ* **566**: 1176–1185.

Illumina. Sequencing Platforms. https://www.illumina.com/systems/sequencing-platforms.html (Accessed May 11, 2017).

Islam MS, Zhang Y, McPhedran KN, Liu Y, Gamal El-Din M. (2015). Granular activated carbon for simultaneous adsorption and biodegradation of toxic oil sands process-affected water organic compounds. *J Environ Manage* **152**: 49–57.

ITRC. (2008). In Situ Bioremediation of Chlorinated Ethene: DNAPL Source Zones. Washington, D.C. www.itrcweb.org.

Jackson CR, Millar JJ, Payne JT, Ochs CA. (2014). Free-living and particle-associated bacterioplankton in large rivers of the Mississippi River basin demonstrate biogeographic patterns. *Appl Environ Microbiol* **80**: 7186–7195.

Jackson MA, Bell JT, Spector TD, Steves CJ. (2016). A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. *PeerJ* **4**: e2341.

Jacquet S, Miki T, Noble R, Peduzzi P, Wilhelm S. (2010). Viruses in aquatic ecosystems: important advancements of the last 20 years and prospects for the future in the field of microbial oceanography and limnology. *Adv Oceanogr Limnol* **1**: 97–141.

Jiang B, Song K, Ren J, Deng M, Sun F, Zhang X. (2012). Comparison of metagenomic samples using sequence signatures. *BMC Genomics* **13**: 730.

Kanehisa M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**: 27–30.

Karnachuk O V., Sasaki K, Gerasimchuk AL, Sukhanova O, Ivasenko DA, Kaksonen AH, *et al.* (2008). Precipitation of Cu-Sulfides by Copper-Tolerant *Desulfovibrio* Isolates. *Geomicrobiol J* **25**: 219–227.

Kembel SW, Wu M, Eisen JA, Green JL, Huse S. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance Von Mering C (ed). *PLoS Comput Biol* **8**: e1002743.

Kim K-H, Bae J-W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol* **77**: 7663–8.

Kirchman DL. (2012). Processes in Microbial Ecology. Oxford University Press: Oxford;New York;

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, *et al.* (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**: 1.

Konwar KM, Hanson NW, Pagé AP, Hallam SJ. (2013). MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* **14**: 202.

Kopylova E, Noe L, Touzet H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**: 3211–3217.

Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, *et al.* (2015). Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science (80-)* **349**. http://science.sciencemag.org/content/349/6252/1101 (Accessed May 14, 2017).

Koskella B, Meaden S. (2013). Understanding Bacteriophage Specificity in Natural Microbial Communities. *Viruses* **5**: 806–823.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Appl Environ Microbiol* **79**: 5112–5120.

Krewski D, Balbus J, Butler-Jones D, Haas C, Isaac-Renton J, Roberts KJ, *et al.* (2002). Managing health risks from drinking water--a report to the Walkerton inquiry. *J Toxicol Environ Heal Part A* **65**: 1635–1823.

Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, *et al.* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814–21.

Larouche JR, Bowden WB, Giordano R, Flinn MB, Crump BC. (2012). Microbial biogeography of arctic streams: Exploring influences of lithology and habitat. *Front Microbiol* **3**: 1.

Legendre P, Legendre LF. (2012). Numerical Ecology. Elsevier.

Lennon JT, Jones SE. (2011). Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat Rev Microbiol* **9**: 119–130.

Leys C, Ley C, Klein O, Bernard P, Licata L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* **49**: 764–766.

Li H, Li M, Huang Y, Rensing C, Wang G. (2013). In silico analysis of bacterial arsenic islands reveals remarkable synteny and functional relatedness between arsenate and phosphate. *Front Microbiol* **4**: 1.

Li X, Harwood VJ, Nayak B, Staley C, Sadowsky MJ, Weidhaas J. (2015). A Novel Microbial Source Tracking Microarray for Pathogen Detection and Fecal Source Identification in Environmental Systems. *Environ Sci Technol* **49**: 7319–29.

Liang Q, Chiu J, Chen Y, Huang Y, Higashimori A, Fang J, *et al.* (2017). Fecal Bacteria Act as Novel Biomarkers for Noninvasive Diagnosis of Colorectal Cancer. *Clin Cancer Res* **23**. http://clincancerres.aacrjournals.org/content/23/8/2061 (Accessed May 20, 2017).

Liljeqvist M, Ossandon FJ, González C, Rajan S, Stell A, Valdes J, *et al.* (2015). Metagenomic analysis reveals adaptations to a cold-adapted lifestyle in a low-temperature acid mine drainage stream. *FEMS Microbiol Ecol* **91**: 1.

Liu L, Yang J, Yu X, Chen G, Yu Z. (2013). Patterns in the composition of microbial communities from a subtropical river: Effects of environmental, spatial and temporal factors. *PLoS One* **8**: 1.

Liu WT, Marsh TL, Cheng H, Forney LJ. (1997). Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Appl Environ Microbiol* **63**: 4516–22.

Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. (2011). UniFrac: an effective distance metric for microbial community comparison. *ISME J* **5**: 169–172.

Lu Y, Hugenholtz P, Batstone DJ, Akgun S, Boyaci H, Hugenhotz P. (2015). Evaluating DNA Extraction Methods for Community Profiling of Pig Hindgut Microbial Community Kellogg CA (ed). *PLoS One* **10**: e0142720.

Mahé F, Rognes T, Quince C, De Vargas C, Dunthorn M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* **3**: e1420.

Maillet N, Collet G, Vannier T, Lavenier D, Peterlongo P. (2014). Commet: Comparing and combining multiple metagenomic datasets. In: *Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014*. pp 94–

98.

Malik A. (2004). Metal bioremediation through growing cells. *Environ Int* **30**: 261–278.

Manor O, Borenstein E. (2015). MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol* **16**: 53.

Mantel N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**: 209–220.

Marçais G, Kingsford C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–70.

Mardis ER. (2017). DNA sequencing technologies: 2006-2016. *Nat Protoc* **12**: 213–218.

Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW, *et al.* (2014). Caught in the middle with multiple displacement amplification: the myth of pooling for avoiding multiple displacement amplification bias in a metagenome. *Microbiome* **2**: 3.

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

McLain JET, Rock CM, Lohse K, Walworth J. (2011). False-positive identification of Escherichia coli in treated municipal wastewater and wastewater-irrigated soils. *Can J Microbiol* **57**: 775–84.

McMurdie PJ, Holmes S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**: e61217.

Mendes Silva D, Domingues L. (2015). On the track for an efficient detection of Escherichia coli in water: A review on PCR-based methods. *Ecotoxicol Environ Saf* **113**: 400–411.

Mercier C, Boyer F, Bonin A, Coissac E. (2013). SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences. In: *Programs and Abstracts of the SeqBio 2013 workshop (Abstract), GdRBIM and gdrIM, Montpellier, France*. pp 27–29.

Meybeck M. (2003). Global analysis of river systems: from Earth system controls to Anthropocene syndromes. *Philos Trans R Soc Lond B Biol Sci* **358**: 1935–55.

Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, *et al.* (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**: 386.

Middelboe M, Jacquet S, Weinbauer M. (2008). Viruses in freshwater ecosystems: An introduction to the exploration of viruses in new aquatic habitats. *Freshw Biol* **53**: 1069–1075.

Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, *et al.* (2013). Using tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* **14**: 193–

202.

Mitchell D. (2007). GC content and genome length in Chargaff compliant genomes. *Biochem Biophys Res Commun* **353**: 207–10.

Mohiuddin M, Schellhorn H. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* **6**. e-pub ahead of print, doi: 10.3389/fmicb.2015.00960.

Murphy J, Riley JP. (1962). A modified single solution method for the determination of phosphate in natural waters. *Anal Chim Acta* **27**: 31–36.

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. (2006). Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochem Biophys Res Commun* **347**: 1–3.

Muyzer G. (1999). DGGE/TGGE a method for identifying genes from natural ecosystems. *Curr Opin Microbiol* **2**: 317–322.

Muyzer G, de Waal EC, Uitterlinden AG. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol* **59**: 695–700.

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

Nikolenko SI, Korobeynikov AI, Alekseyev MA. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14 Suppl 1**: S7.

Niño-García JP, Ruiz-González C, del Giorgio PA. (2016). Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *ISME J* 1–12.

Nocker A, Burr M, Camper AK. (2007). Genotypic Microbial Community Profiling: A Critical Technical Review. *Microb Ecol* **54**: 276–289.

Okabe S, Itoh T, Satoh H, Watanabe Y. (1999). Analyses of spatial distributions of sulfate-reducing bacteria and their activity in aerobic wastewater biofilms. *Appl Environ Microbiol* **65**: 5107–5116.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, *et al.* (2015). Package 'vegan': Community Ecology Package. *Community Ecol Packag version* **2**: 280.

Ondov BD, Treangen TJ, Mallonee AB, Bergman NH, Koren S, Phillippy AM. (2015). Mash: Fast genome and metagenome distance estimation using MinHash. *Genome Biol* 29827.

Orell A, Navarro CA, Arancibia R, Mobarec JC, Jerez CA. (2010). Life in blue: Copper resistance mechanisms of bacteria and Archaea used in industrial biomining of minerals. *Biotechnol Adv* **28**: 839–848.

Overbeek R, Begley T, Butler RM, Choudhuri J V, Chuang H-Y, Cohoon M, *et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* **33**: 5691–702.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, *et al.* (2016). Uncovering Earth's virome. *Nature* **536**: 425–430.

Pal C, Bengtsson-Palme J, Rensing C, Kristiansson E, Larsson DGJ. (2014). BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res* **42**: D737-43.

Palmer MA, Febria CM. (2012). Ecology. The heartbeat of ecosystems. *Science* **336**: 1393–4.

Panasiuk O, Hedström A, Marsalek J, Ashley RM, Viklander M. (2015). Contamination of stormwater by wastewater: A review of detection methods. *J Environ Manage* **152**: 241–250.

Peabody MA, Van Rossum T, Lo R, Brinkman FSL. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics* **16**: 363.

Peduzzi P. (2016). Virus ecology of fluvial systems: a blank spot on the map? *Biol Rev* **91**: 937–949.

Perales-Vela HV, Peña-Castro JM, Cañizares-Villanueva RO. (2006). Heavy metal detoxification in eukaryotic microalgae. *Chemosphere* **64**: 1–10.

Plummer JD, Long SC. (2007). Monitoring source water for microbial contamination: Evaluation of water quality measures. *Water Res* **41**: 3716–3728.

Price MN, Dehal PS, Arkin AP. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**: 1641–50.

Qiang F, Dongya Z, Youwen Q. (2016). Biomarkers for colorectal cancer.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-6.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**: 38.

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. (2017). Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35**: 833–844.

R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria. http://www.r-project.org/ (Accessed November 26, 2014).

Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.

Read DS, Gweon HS, Bowes MJ, Newbold LK, Field D, Bailey MJ, *et al.* (2014). Catchment-scale biogeography of riverine bacterioplankton. *ISME J* **9**: 516–526.

Reichenberger ER, Rosen G, Hershberg U, Hershberg R. (2015). Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. *Genome Biol Evol* evv063-.

Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, *et al.* (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**: e545.

Rodriguez-R LM, Konstantinidis KT. (2014). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**: 629–635.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ Prepr* **4**: e2409v1.

Rosseel T, Van Borm S, Vandenbussche F, Hoffmann B, van den Berg T, Beer M, *et al.* (2013). The Origin of Biased Sequence Depth in Sequence-Independent Nucleic Acid Amplification and Optimization for Efficient Massive Parallel Sequencing Kapoor A (ed). *PLoS One* **8**: e76144.

Van Rossum T, Peabody MA, Uyaguari-Diaz MI, Cronin KI, Chan M, Slobodan JR, *et al.* (2015). Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. *Front Microbiol* **6**: 1405.

Ruiz-González C, Niño-García JP, del Giorgio PA. (2015a). Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecol Lett* **18**: 1198–1206.

Ruiz-González C, Niño-García JP, Lapierre J-F, del Giorgio PA. (2015b). The quality of organic matter shapes the functional biogeography of bacterioplankton across boreal freshwater ecosystems. *Glob Ecol Biogeogr* n/a-n/a.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, *et al.* (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**: 87.

Santamaria M, Fosso B, Consiglio A, De Caro G, Grillo G, Licciulli F, *et al.* (2012). Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform* **13**: 682–695.

Savio D, Sinclair L, Ijaz UZ, Parajka J, Reischer GH, Stadler P, *et al.* (2015). Bacterial diversity along a 2600 km river continuum. *Environ Microbiol* **17**: 4994–5007.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* gku1341-.

Schloss PD. (2015). No, greengenes' reference alignment hasn't improved. *mothur blog*. http://blog.mothur.org/2015/08/04/No-greengenes-hasnt-improved/ (Accessed February 24, 2017).

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–41.

Sedlar K, Kupkova K, Provaznik I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput Struct Biotechnol J* **15**: 48–55.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, *et al.* (2011). Metagenomic biomarker discovery and explanation. *Genome Biol* **12**: R60.

Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**: 811–4.

Sherchan SP, Bachoon DS. (2011). The presence of atrazine and atrazine-degrading bacteria in the residential, cattle farming, forested and golf course regions of Lake Oconee. *J Appl Microbiol* **111**: 293–299.

Sigee DC. (2005). Freshwater Microbiology. John Wiley & Sons, Ltd: Chichester, UK.

Simon M, López-García P, Deschamps P, Moreira D, Restoux G, Bertolino P, *et al.* (2015). Marked seasonality and high spatial variability of protist communities in shallow freshwater systems. *ISME J*. e-pub ahead of print, doi: 10.1038/ismej.2015.6.

Singer E, Bushnell B, Coleman-Derr D, Bowman B, Bowers RM, Levy A, *et al.* (2016). High-resolution phylogenetic microbial community profiling. *ISME J* **10**: 2020–2032.

Staley C, Gould TJ, Wang P, Phillips J, Cotner JB, Sadowsky MJ. (2014a). Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. *Front Microbiol* **5**: 524.

Staley C, Gould TJ, Wang P, Phillips J, Cotner JB, Sadowsky MJ. (2014b). Core functional traits of bacterial communities in the Upper Mississippi River show limited variation in response to land cover. *Front Microbiol* **5**: 414.

Staley C, Gould TJ, Wang P, Phillips J, Cotner JB, Sadowsky MJ. (2015). Species sorting and seasonal dynamics primarily shape bacterial communities in the Upper Mississippi River. *Sci Total Environ* **505**: 435–445.

Sterner RW, Smutka TM, McKay RML, Xiaoming Q, Brown ET, Sherrell RM. (2004). Phosphorus and trace metal limitation of algae and bacteria in Lake Superior. *Limnol Oceanogr* **49**: 495–507.

Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. (2015). rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res* **43**: D593–D598.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, *et al.* (2015). Structure and function of the global ocean microbiome. *Science (80-)* **348**: 1261359–

1261359.

Suzuki R, Shimodaira H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**: 1540–2.

Tatusov RL, Galperin MY, Natale DA, Koonin E V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**: 33–36.

Tay ST-L, Hemond FH, Krumholz LR, Cavanaugh CM, Polz MF. (2001). Population Dynamics of Two Toluene Degrading Bacterial Species in a Contaminated Stream. *Microb Ecol* **41**: 124–131.

Tay STL, Hemond HF, Polz MF, Cavanaugh CM, Dejesus I, Krumholz LR. (1998). Two new Mycobacterium strains and their role in toluene degradation in a contaminated stream. *Appl Environ Microbiol* **64**: 1715–1720.

Thingstad TF. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* **45**: 1320–1328.

Thomas MC, Selinger LB, Inglis GD. (2012). Seasonal diversity of planktonic protists in Southwestern Alberta rivers over a 1-year period as revealed by terminal restriction fragment length polymorphism and 18S rRNA gene library analyses. *Appl Environ Microbiol* **78**: 5653–5660.

Thompson BM, Lin C-H, Hsieh H-Y, Kremer RJ, Lerch RN, Garrett HE. (2010). Evaluation of PCR-based Quantification Techniques to Estimate the Abundance of Atrazine Chlorohydrolase Gene in Rhizosphere Soils. *J Environ Qual* **39**: 1999.

Tsitko I, Lusa M, Lehto J, Parviainen L, Ikonen ATK, Lahdenperä A-M, *et al.* (2014). The Variation of Microbial Communities in a Depth Profile of an Acidic, Nutrient-Poor Boreal Bog in Southwestern Finland. *Open J Ecol* **4**: 832–859.

Tyagi S, Sharma B, Singh P, Dobhal R. (2013). Water Quality Assessment in Terms of Water Quality Index. *Am J Water Resour* **1**: 34–38.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, *et al.* (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res* **40**: e115–e115.

Uyaguari-Diaz MI, Chan M, Chaban BL, Croxen MA, Finke JF, Hill JE, *et al.* (2016). A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome* **4**: 20.

Uyaguari-Diaz MI, Slobodan JR, Nesbitt MJ, Croxen MA, Isaac-Renton J, Prystajecky NA, *et al.* (2015). Automated Gel Size Selection to Improve the Quality of Next-generation Sequencing Libraries Prepared from Environmental Water Samples. *J Vis Exp* e52685.

Vaulot D. (1989). CYTOPC, Processing software for flow cytometric data. **2**: 8.

Vellend M. (2010). Conceptual synthesis in community ecology. *Q Rev Biol* **85**: 183–

206.

Vens C, Rosso M-N, Danchin EGJ. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **27**: 1231–8.

Vollmers J, Wiegand S, Kaster A-K. (2017). Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLoS One* **12**: 1–31.

Vörösmarty CJ, McIntyre PB, Gessner MO, Dudgeon D, Prusevich A, Green P, *et al.* (2010). Global threats to human water security and river biodiversity. *Nature* **467**: 555–61.

Wagner J, Coupland P, Browne HP, Lawley TD, Francis SC, Parkhill J. (2016). Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol* **16**: 274.

van der Walt AJ, Van Goethem MW, Ramond J-B, Makhalanyane TP, Reva O, Cowan DA. (2017). Assembling Metagenomes, One Community At A Time. *bioRxiv*. http://biorxiv.org/content/early/2017/03/24/120154 (Accessed May 11, 2017).

Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey H a, Ganem D, *et al.* (2002). Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A* **99**: 15687–92.

Wang Y, Liu L, Chen H, Yang J. (2015a). Spatiotemporal dynamics and determinants of planktonic bacterial and microeukaryotic communities in a Chinese subtropical river. *Appl Microbiol Biotechnol* **99**: 9255–9266.

Wang Y, Yang J, Liu L, Yu Z. (2015b). Quantifying the effects of geographical and environmental factors on distribution of stream bacterioplankton within nature reserves of Fujian, China. *Environ Sci Pollut Res* **22**: 11010–11021.

Weese D, Schulz MH. (2008). Efficient String Mining under Constraints Via the Deferred Frequency Index. In: Perner P (ed) Lecture Notes in Computer Science Vol. 5077. *Proceedings of the 8th Industrial Conference on Data Mining: LNAI 5077.* Springer: Berlin, Heidelberg, pp 374–388.

Welschmeyer NA. (1994). Fluorometric analysis of chlorophyll a in the presence of chlorophyll b and pheopigments. *Limnol Oceanogr* **39**: 1985–1992.

Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, *et al.* (2014). Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**: 19.

Westcott SL, Schloss PD. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**: e1487.

Wetzel RG. (2001). Limnology : lake and river ecosystems. Academic Press.

Weynberg KD, Wood-Charlson EM, Suttle CA, van Oppen MJH. (2014). Generating viral metagenomes from the coral holobiont. *Front Microbiol* **5**: 206.

White C, Gadd GM. (2000). Copper accumulation by sulfate-reducing bacterial biofilms. *FEMS Microbiol Lett* **183**: 313–318.

White DCD, Flemming CAC, Leung KT, Macnaughton SJ. (1998). In situ microbial ecology for quantitative appraisal, monitoring, and risk assessment of pollution remediation in soils, the subsurface, the rhizosphere and in biofilms. *J Microbiol Methods* **32**: 93–105.

White DC, Wilson JW. (1989). Subsurface microbiota as monitors of contaminant migration and mitigation. *Symp New F Tech Quantifying Phys Chem Prop Heterog Aquifers*. http://www.davidcwhite.org/fulltext/242.pdf (Accessed November 27, 2014).

Whittaker RH. (1960). Vegetation of the Siskiyou Mountains, Oregon and California. *Ecol Monogr* **30**: 279–338.

van der Wielen PWJJ, Medema G. (2010). Unsuitability of quantitative Bacteroidales 16S rRNA gene assays for discerning fecal contamination of drinking water. *Appl Environ Microbiol* **76**: 4876–81.

Williamson KE, Harris J V., Green JC, Rahman F, Chambers RM. (2014). Stormwater runoff drives viral community composition changes in inland freshwaters. *Front Microbiol* **5**. e-pub ahead of print, doi: 10.3389/fmicb.2014.00105.

Wood ED, Armstrong FAJ, Richards FA. (1967). Determination of nitrate in sea water by cadmium-copper reduction to nitrite. *J Mar Biol Assoc United Kingdom* **47**: 23–31.

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**: 134.

Yilmaz S, Allgaier M, Hugenholtz P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods* **7**: 943–944.

Yu J, Feng Q, Wong SH, Zhang D, Liang Q yi, Qin Y, *et al.* (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**: 70–78.

Zeglin LH. (2015). Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Front Microbiol* **6**: 454.

Zerbino DR, Birney E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

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

Zhao Y, Tang H, Ye Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**: 125–6.

Zhu F, Massana R, Not F, Marie D, Vaulot D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79–92.

Zinger L, Gobet A, Pommier T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21**: 1878–96.

Appendix A.

Extended figures for Chapter 2

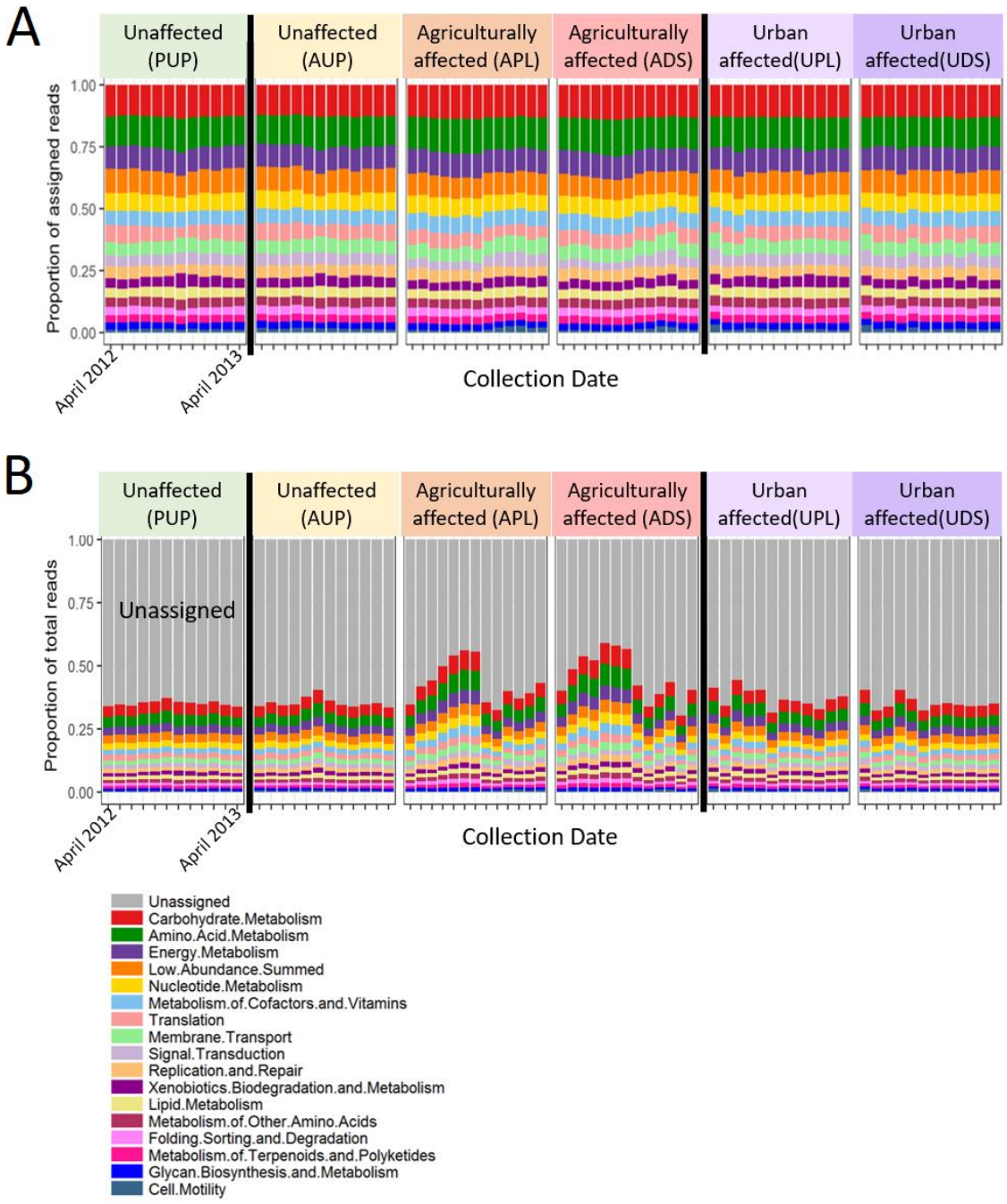

Figure A1 Normalisation of gene family abundances by proportion of reads assigned masks variability

Stacked bar plot showing abundance of high level gene families from the KEGG database. Each column is a sample. Samples are arranged by sampling site, indicated by colours in the top title row. Black bars separate watersheds. Within each sampling site, samples are arranged chronologically. Gene family abundances are represented as (A) proportions of the reads assigned or (B) proportion of subsampled reads. Variability in the proportion of reads unassigned is not seen when function profiles are only described as proportions of reads assigned.

Appendix B.

Extended figures for Chapter 4

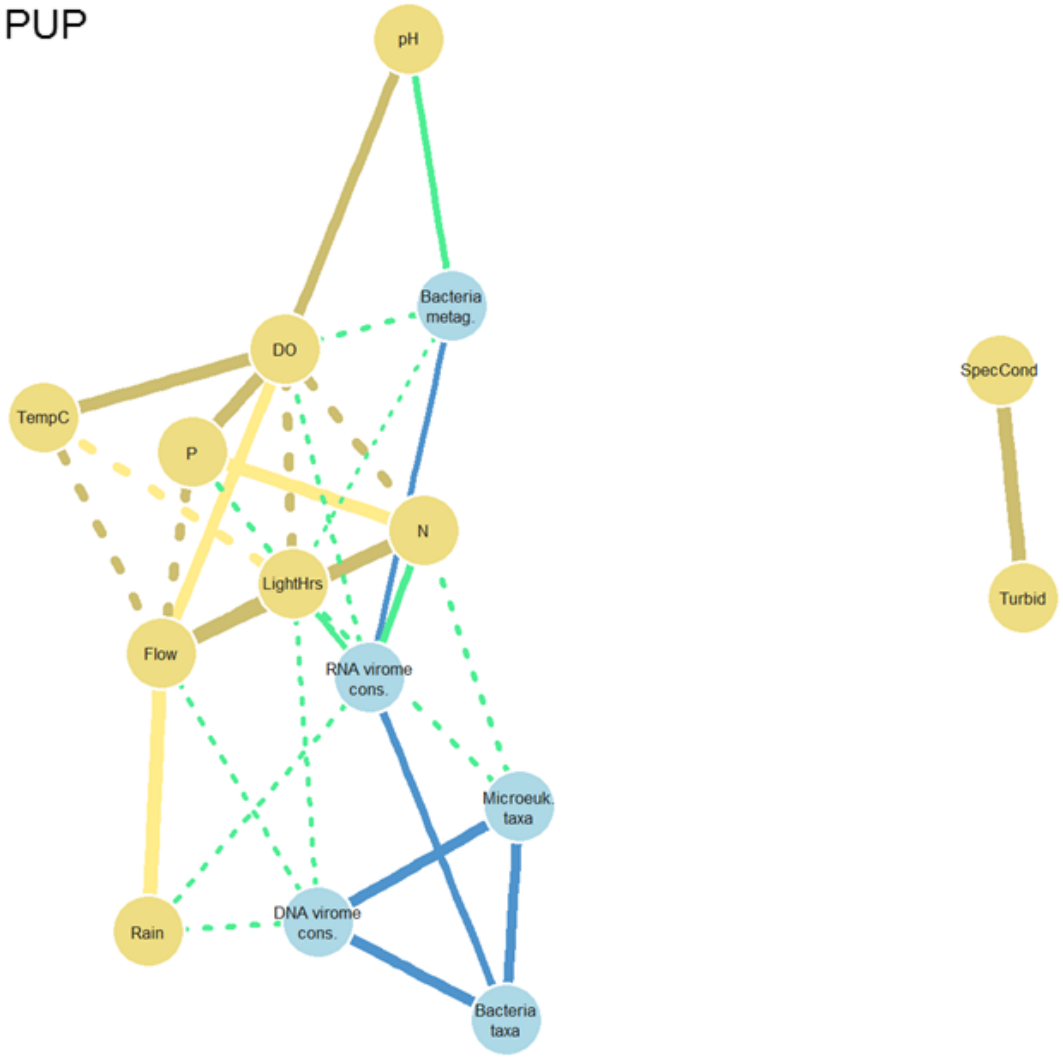

Figure B1 Temporal variation in viruses, bacteria, and microeukaryotes in site PUP (Watershed "P", protected catchment, limited land use)

Network of correlations between and within environmental conditions and microbial communities in site PUP. Nodes are environmental conditions (yellow) and microbial communities (blue). Conservative viromes were used (see methods). Edges are coloured by the nodes types they connect. Light yellow edges are positive correlations between environmental measures, dark yellow edges are negative correlations. Solid edges represent statistically significant relationships ($q < 0.1$), dashed edges represent relationships that are strong but have lower statistical confidence ($p < 0.05$). Edge width reflects the strength (R^2) of the represented correlation.

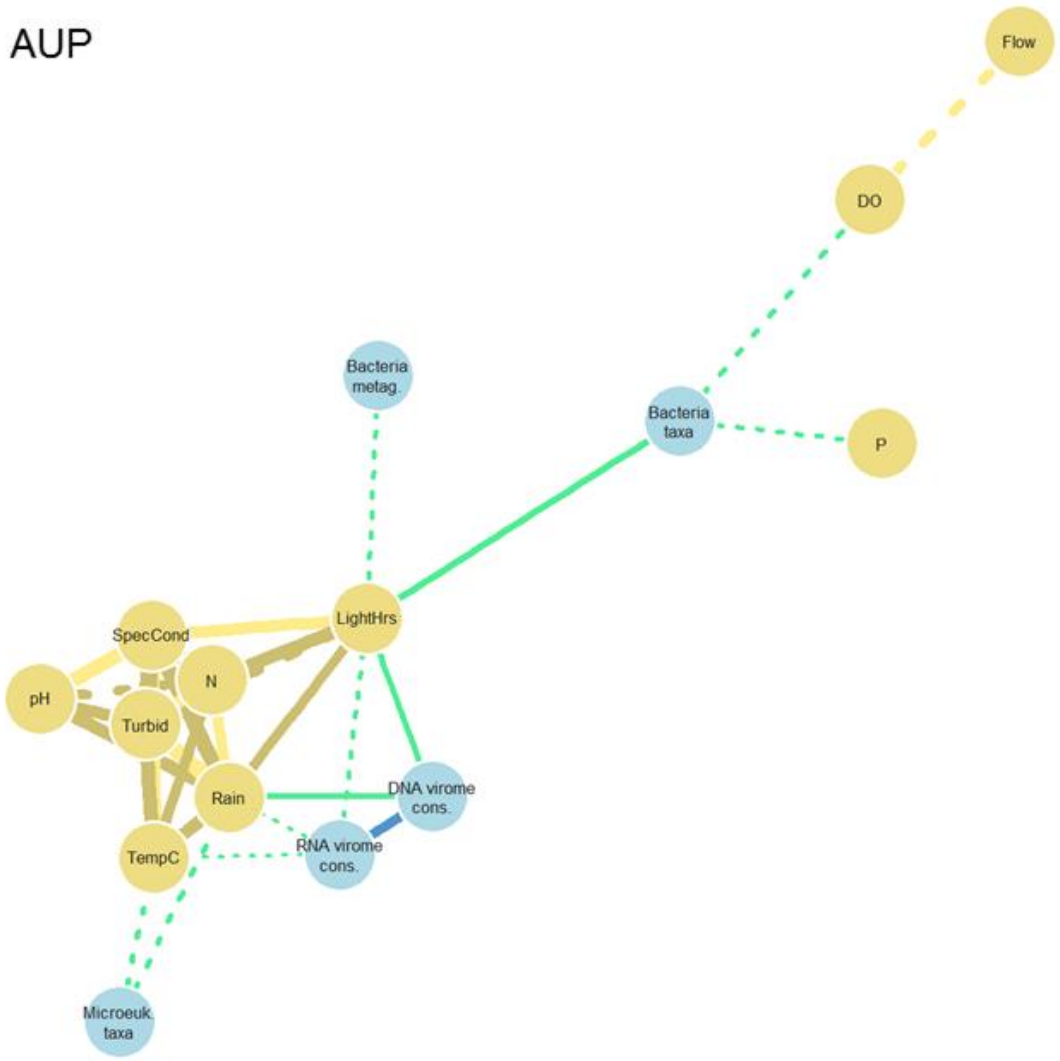

Figure B2 Temporal variation in viruses, bacteria, and microeukaryotes in site AUP (Watershed "A", upstream of agricultural activity, limited land use)

Network of correlations between and within environmental conditions and microbial communities in site AUP. Nodes are environmental conditions (yellow) and microbial communities (blue). Conservative viromes were used (see methods). Edges are coloured by the nodes types they connect. Light yellow edges are positive correlations between environmental measures, dark yellow edges are negative correlations. Solid edges represent statistically significant relationships ($q < 0.1$), dashed edges represent relationships that are strong but have lower statistical confidence ($p < 0.05$). Edge width reflects the strength (R^2) of the represented correlation.

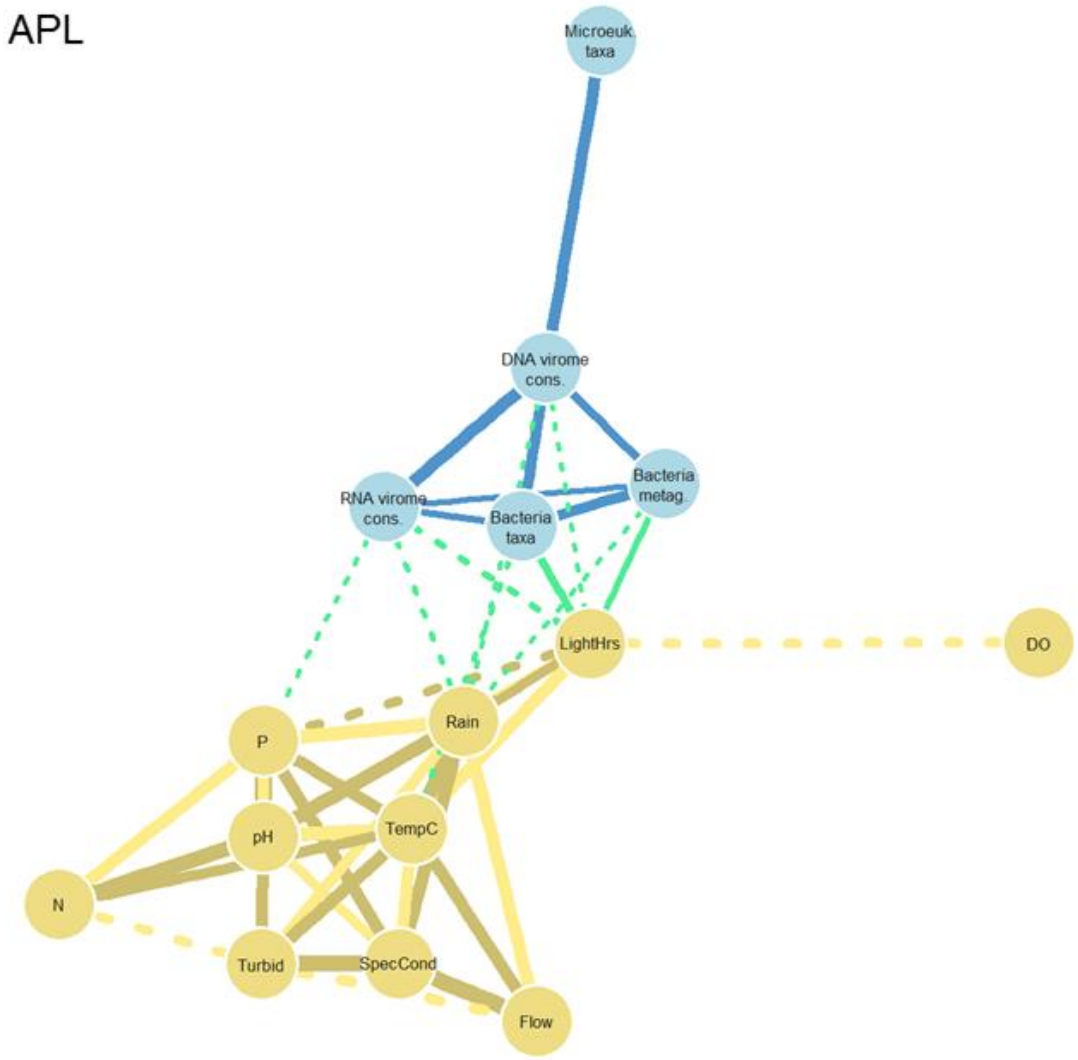

Figure B3 Temporal variation in viruses, bacteria, and microeukaryotes in site APL (Watershed "A", agricultural activity in catchment)

Network of correlations between and within environmental conditions and microbial communities in site APL. Nodes are environmental conditions (yellow) and microbial communities (blue). Conservative viromes were used (see methods). Edges are coloured by the nodes types they connect. Light yellow edges are positive correlations between environmental measures, dark yellow edges are negative correlations. Solid edges represent statistically significant relationships ($q < 0.1$), dashed edges represent relationships that are strong but have lower statistical confidence ($p < 0.05$). Edge width reflects the strength (R^2) of the represented correlation.

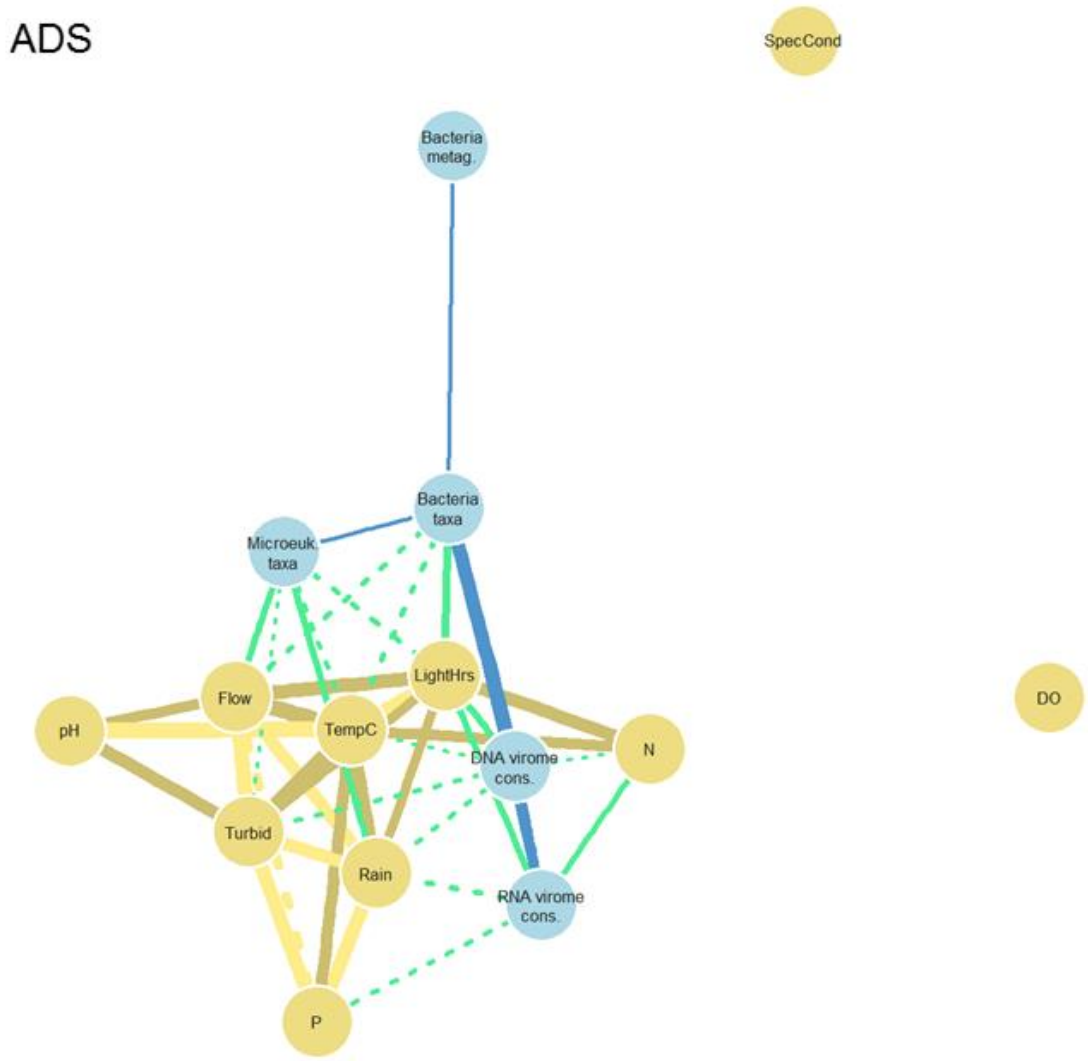

Figure B4 Temporal variation in viruses, bacteria, and microeukaryotes in site ADS (Watershed "A", downstream of agricultural activity)

Network of correlations between and within environmental conditions and microbial communities in site ADS. Nodes are environmental conditions (yellow) and microbial communities (blue). Conservative viromes were used (see methods). Edges are coloured by the nodes types they connect. Light yellow edges are positive correlations between environmental measures, dark yellow edges are negative correlations. Solid edges represent statistically significant relationships ($q < 0.1$), dashed edges represent relationships that are strong but have lower statistical confidence ($p < 0.05$). Edge width reflects the strength (R^2) of the represented correlation.

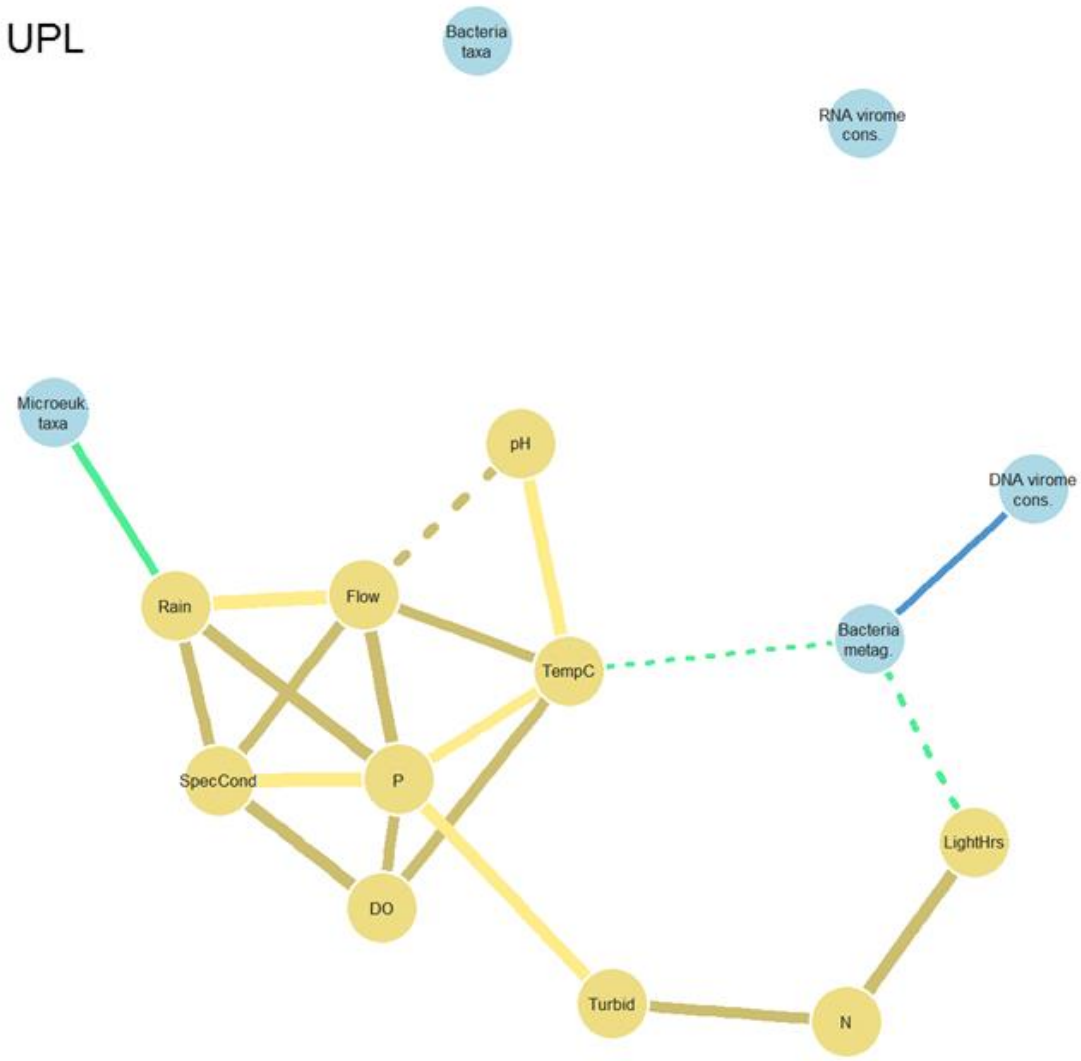

Figure B5 Temporal variation in viruses, bacteria, and microeukaryotes in site UPL (Watershed "U", some urban activity in catchment)

Network of correlations between and within environmental conditions and microbial communities in site UPL. Nodes are environmental conditions (yellow) and microbial communities (blue). Conservative viromes were used (see methods). Edges are coloured by the nodes types they connect. Light yellow edges are positive correlations between environmental measures, dark yellow edges are negative correlations. Solid edges represent statistically significant relationships ($q < 0.1$), dashed edges represent relationships that are strong but have lower statistical confidence ($p < 0.05$). Edge width reflects the strength (R^2) of the represented correlation.

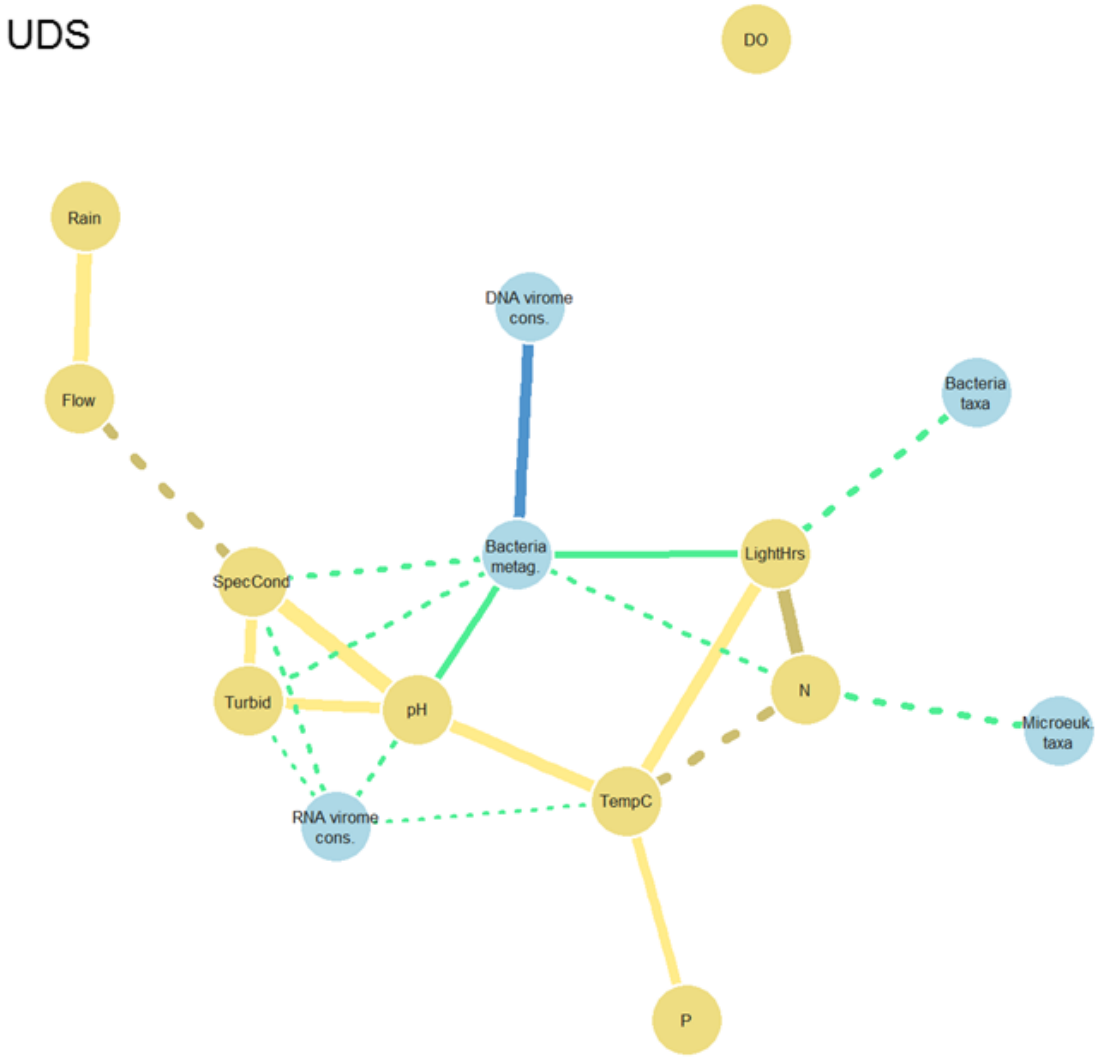

Figure B6 Temporal variation in viruses, bacteria, and microeukaryotes in site UDS (Watershed "U", urban activity in catchment)

Network of correlations between and within environmental conditions and microbial communities in site UDS. Nodes are environmental conditions (yellow) and microbial communities (blue). Conservative viromes were used (see methods). Edges are coloured by the nodes types they connect. Light yellow edges are positive correlations between environmental measures, dark yellow edges are negative correlations. Solid edges represent statistically significant relationships ($q < 0.1$), dashed edges represent relationships that are strong but have lower statistical confidence ($p < 0.05$). Edge width reflects the strength (R^2) of the represented correlation.

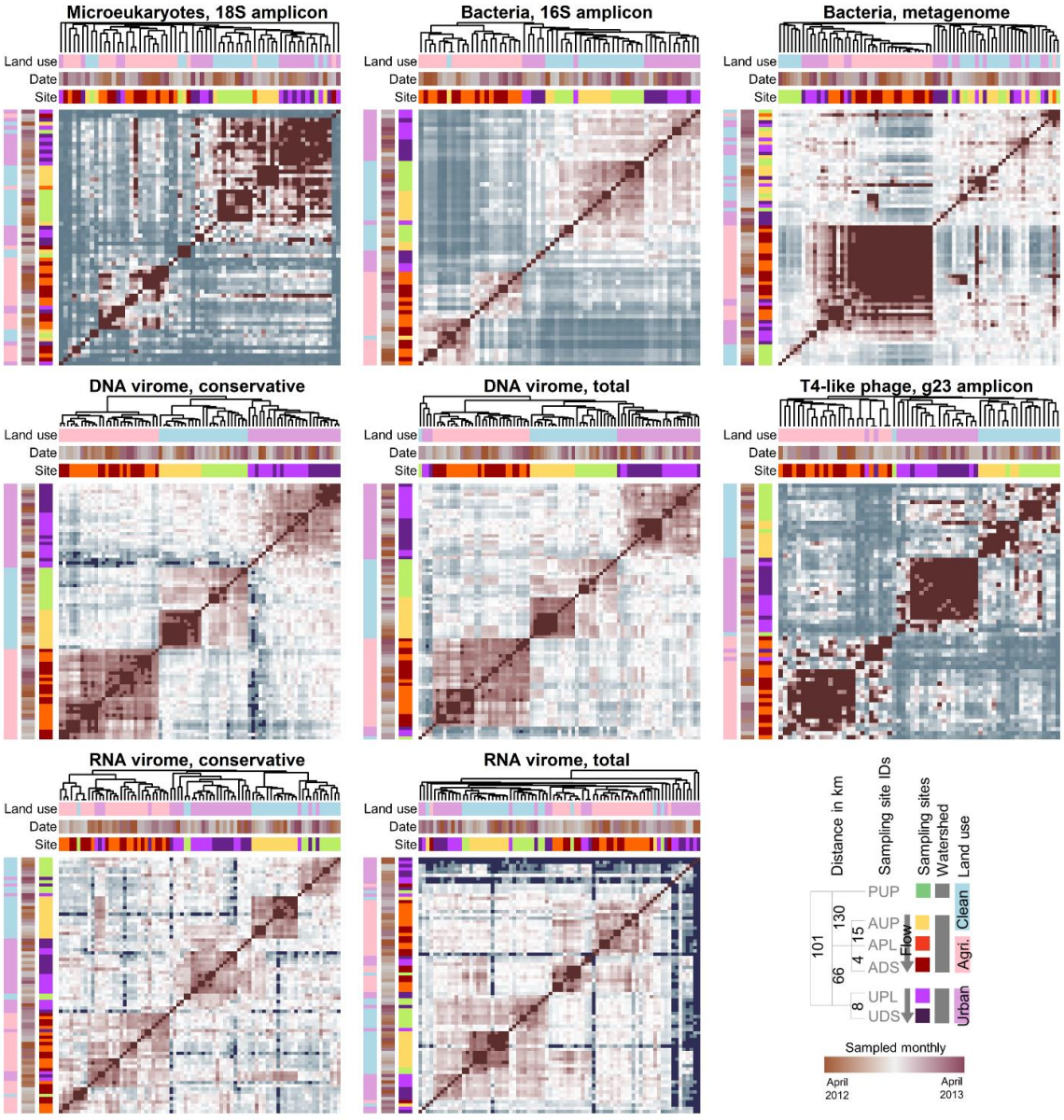

Figure B7 Beta diversity within viral, bacterial, and eukaryotic communities over 1 year of monthly samples.

Heatmaps showing pairwise similarities between samples based on metagenome k-mer profiles or amplicon OTU profiles. Samples were collected from six sites in three watersheds monthly for a year. Columns and rows are symmetrical and ordered using average hierarchical clustering. Heatmap colour represents the strength of similarity, with very high or very low similarity scores collapsed in the gradient (see Methods for criteria). Column and row side bars are coloured by sample site, date, and land use.

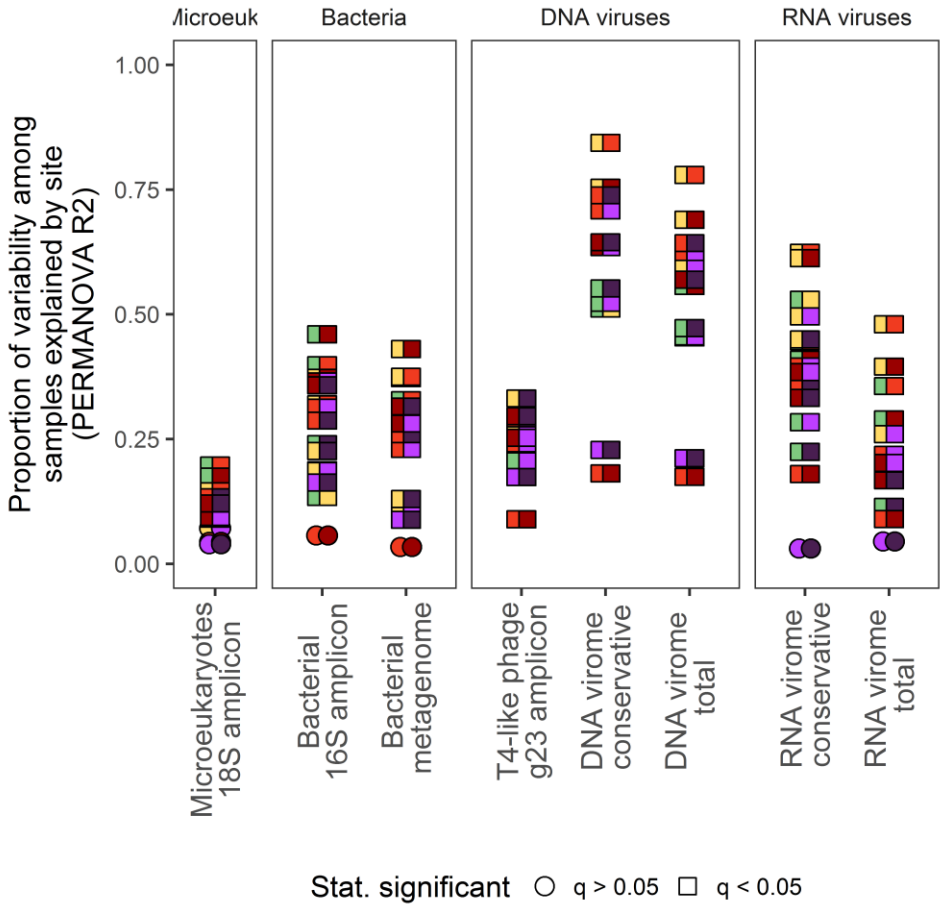

Figure B8 Pairwise distinctiveness of samples based on site.
Horizontally adjacent pairs display the results of a pair-wise site based PERMANOVA. This is site-pair level resolution for the results in Figure 10. Colours correspond to sampling site (see Figure 1a). Shape corresponds to statistical significance, with rectangles having $q < 0.05$ and circles having $q > 0.05$. The higher the value of on the y axis (R^2) the larger the proportion of variability among samples explained by sampling site.

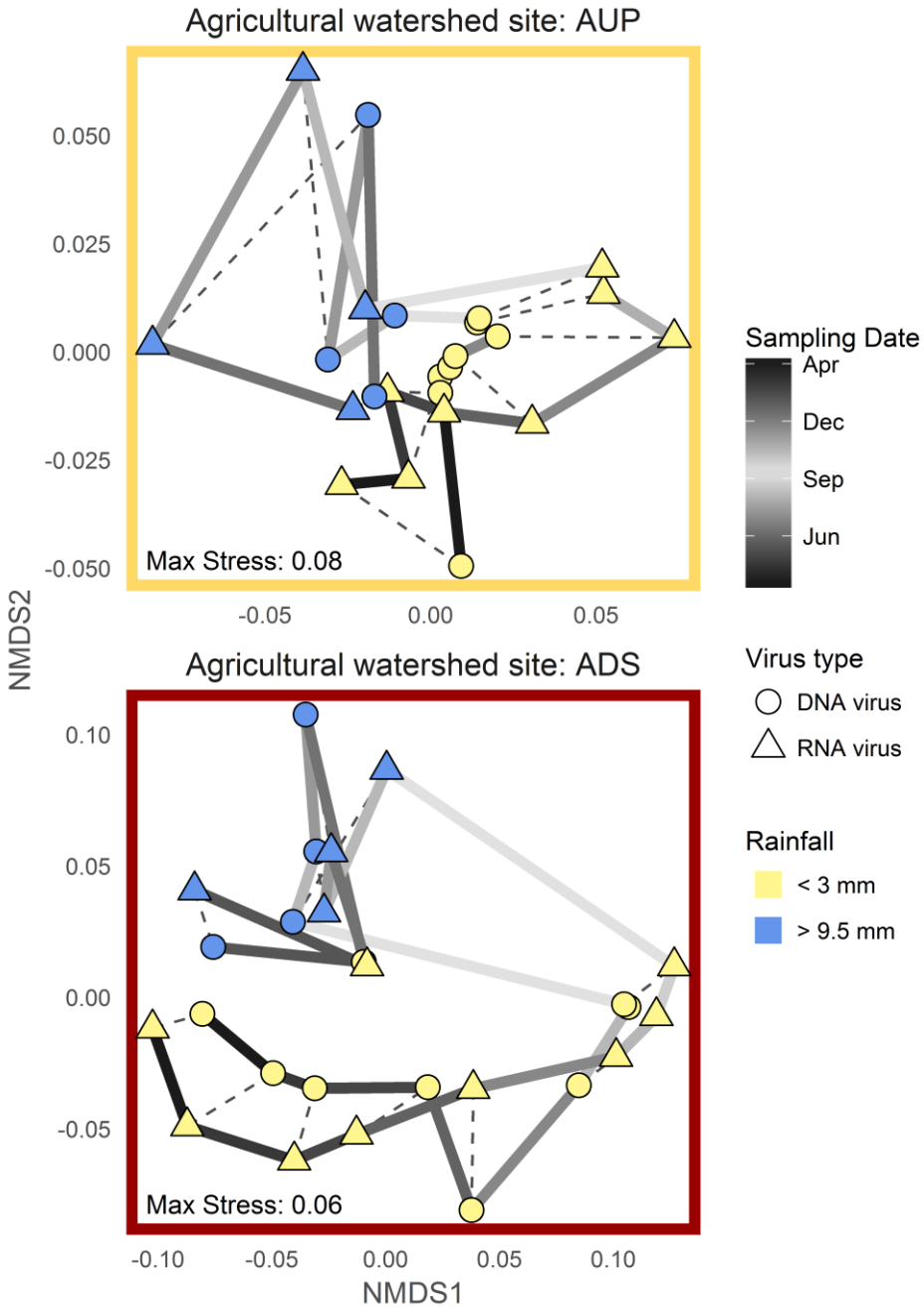

Figure B9 Temporal variation in RNA and DNA viruses in an agriculturally affected site

NMDS plot of DNA & RNA viral communities from a site with minimal land use (AUP) and an agriculturally affected site (ADS). Each point represents a viral community. Dashed lines connect viromes extracted from the same sample. Solid lines trace the temporal path between samples and are coloured by sampling date. Points are coloured by the average rainfall over the three days prior to sampling. The sample with unusually high rainfall in the second half of the sampling year (visible in the ADS plot) was missing from the AUP DNA viruses. The same rainfall measurements were used for both sites.