Applying Metagenomics Analysis Towards a Better Understanding of Freshwater Microbial Communities

by

Michael Alexander Peabody

B.Sc., Simon Fraser University, 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
Summer 2017

Approval

Name:	**Michael Alexander Peabody**
Degree:	**Doctor of Philosophy**
Title:	**Applying Metagenomics Analysis Towards a Better Understanding of Freshwater Microbial Communities**
Examining Committee:	**Chair:** **Dr. Jack Chen** Professor
	Dr. Fiona Brinkman Senior Supervisor Professor
	Dr. Paul Pavlidis Supervisor Professor Department of Psychiatry and Centre for Brain Health University of British Columbia
	Dr. Robert Holt Supervisor Professor
	Dr. Lisa Craig Internal Examiner Associate Professor
	Dr. Iddo Friedberg External Examiner Associate Professor Department of Veterinary Microbiology and Preventive Medicine Iowa State University
Date Defended/Approved:	June 15, 2017

Abstract

Microbial communities may now be studied in more detail using culture-independent methods such as metagenomics (directly analyzing genomes from an environmental sample). One of the many potential applications of metagenomics is in the assessment of water quality. Current methods for detection of fecal pollution in water rely on culture-based microbial testing which is slow and can lack sensitivity and specificity. For the WatershedDiscovery.ca project, it was hypothesized that a molecular-based test, developed based on metagenomics analysis, may be more rapid, and accurate for characterizing fecal pollution.

Many bioinformatics methods for metagenomics sequence classification have been developed, but when initiating this research, no comprehensive evaluation of method accuracy had yet been published. Thus, using both *in silico* and *in vitro* simulated communities, a comprehensive evaluation of metagenomics taxonomic sequence classification methods was performed. Utilizing knowledge gained from this comparative evaluation, a study was undertaken of microbial community dynamics in monthly water samples from sites in urban, protected, and agricultural watersheds collected over a one-year period. Freshwater samples collected from sites affected by agricultural activity showed distinct microbial profiles versus samples collected from unaffected sites, and a notable presence of *Legionella* was discovered in all watershed sampling sites (the largest study of *Legionella* in watersheds to date). Furthermore, biomarkers were developed that could distinguish agriculturally affected samples from pristine samples collected in our watershed study.

Finally, there is a lack of methods for the prediction of subcellular localization (SCL) from metagenomics sequences—of interest for the identification of cell surface/secreted proteins for development of ELISA-based diagnostics and other applications. Thus, PSORTb, a precise bacterial and archaeal SCL program, was modified to enable the classification of metagenomics sequences, and applied to the analysis of the watershed samples. A database of protein SCL associated with PSORTb was expanded to make it suitable for a wider diversity of microbes, particularly those with atypical cell envelopes. Collectively this work expands our understanding of metagenomics software accuracy, and available analysis tools, and provides insight into freshwater microbial community dynamics, with potential application in water quality test development.

Keywords: bioinformatics; metagenomics; subcellular localization; water quality; marker detection; watershed

Acknowledgements

I would like to thank my supervisor, Dr. Fiona Brinkman, for all the support, guidance, and kindness she has shown me over the years. I would also like to thank my committee members, Dr. Paul Pavlidis and Dr. Robert Holt, for their ongoing guidance and feedback. Thanks to all members of the Brinkman lab, both current and past, who have been so supportive and welcoming. A special thanks to collaborators on the PSORTdb and PSORTm projects, especially Matthew Laird, Caitlyn Vlasschaert, Gemma Hoad, and Raymond Lo.

Thank you to members of the Applied Metagenomics of the Watershed Microbiome project, particularly Dr. Patrick Tang, Dr. Natalie Prystajecky, Dr. Matthew Croxen, Dr. Miguel Uyaguari-Diaz, Dr. William Hsiao, Anamaria Crisan, Alvin Tian, Marli Vlok, and Thea Van Rossum.

Thank you also to Dr. Brian Raphael's group at the Centers for Disease Control and Prevention for their collaboration and helpful discussions on the *Legionella* work.

Table of Contents

List of Tables

List of Figures

Glossary

ADS	The sampling site in the agricultural watershed downstream of agricultural “pollution”
APL	The sampling site in the agricultural watershed at the site of agricultural “pollution”
AUP	The sampling site in the agricultural watershed upstream of agricultural “pollution”
BLAST	Basic Local Alignment Search Tool
bp	Base pair
CAMI	Critical Assessment of Metagenome Interpretation
CCME	Canadian Council of Ministers of the Environment
CM	Cytoplasmic membrane
Copiotroph	An organism that tends to be found in environments which are rich in organic matter
ct value	Cycle threshold value. The number of cycles required for the fluorescent signal to cross the threshold (a fluorescent signal significantly above the background fluorescence)
ELISA	Enzyme-linked immunosorbent assay
GO	Gene ontology
HMM	Hidden Markov model
In silico	Performed via computer simulation
In vitro	Studies conducted in a laboratory vessel or other controlled experimental environment rather than within a living organism
K_a/K_s	Ratio of nonsynonymous substitution rate (K_a) to the synonymous substitution rate (K_s)
LD	Legionnaires’ disease
LPS	Lipopolysaccharides
OM	Outer membrane
OQS	Optimal query sequence
OTU	Operational taxonomic unit
PCR	Polymerase chain reaction
PDS	The sampling site in the protected watershed downstream of a reservoir (water first passes through a pipe)
PUP	The sampling site in the protected watershed upstream of the entry point to a reservoir (collected from a river)

qPCR	Quantitative polymerase chain reaction
SCL	Subcellular localization
SD	Standard deviation
SVM	Support vector machine
UDS	The sampling site in the urban watershed downstream of urban “pollution”
UPL	The sampling site in the urban watershed at the site of urban “pollution”
WQI	Water quality index
Alpha diversity (α-diversity)	Within-sample diversity
Beta diversity (β-diversity)	Between-sample diversity

Chapter 1.

Introduction

1.1. Microbial life

Microorganisms, organisms that can not be easily seen with the naked eye, are found ubiquitously. From familiar environments such as soil and the ocean, to environments with extreme properties including frozen environments such as arctic tundra (Neufeld and Mohn, 2005), highly acidic environments such as acid mine drainage (Méndez-García et al., 2014), and high pressure environments deprived of sunlight such as deep-sea hydrothermal vents (Pagé et al., 2004). The total number of prokaryotes on earth is estimated to be on the order of 10^{30} cells, and the total amount of carbon approximately same the as the carbon in all plants (Whitman et al., 1998).

Microorganisms are indispensable for life on Earth. They catalyze essential transformations in biogeochemical cycles, converting the key elements of life into biologically accessible forms. In addition to critical roles in nutrient cycling, microbes are involved in numerous other important ecological activities such as supporting plant health, regulating soil fertility, and modulating and maintaining the atmosphere (Bodelier, 2011; van der Heijden et al., 2008). Microorganisms play a variety of roles in our everyday lives, such as their involvement in the production of bread, yoghurt, cheese, beer, wine, and many other consumable products. Industrial biotechnology can involve the use of microorganisms to generate useful products such as the production of insulin or biofuels, and bioremediation can be used to clean up waste such as in oil spills (Atlas and Hazen, 2011). Many antibiotics that were discovered are produced from microorganisms; for example, streptomycin, the first antibiotic cure for tuberculosis, which was discovered from *Streptomyces griseus*.

Microbes also play an important role in human health. Before the establishment of public health departments and the discovery and use of antibiotics, more than half of deaths were due to infectious diseases, such as diphtheria and tuberculosis (Jones et al., 2012). However, microorganisms are not always harmful to human health. The microbiota is the assemblage of microorganisms in an environment, such as the human

body. The number of microorganisms in the human body is estimated to be roughly the same as the number of human cells (Sender et al., 2016). The microbiota found in the human gut perform many beneficial functions for their human hosts, such as: protecting us against pathogens by competing with them for nutrients or attachment to cell surfaces, fermenting dietary fiber into short-chain fatty acids which can be absorbed by the host, and synthesizing of certain compounds such as vitamin K (Clarke et al., 2014). Although it had been known for awhile that the gut microbiota played some role in health, in recent years there has been increasing attention to their role; there have been a number of studies showing associations between microbiota and host physiology, including an intriguing link between the gut and the brain (Tremlett et al., 2017). A few examples of the links between microbiota and host physiology include associations with cardiovascular disease (Wang et al., 2011), diabetes (Karlsson et al., 2013; Qin et al., 2012), obesity (Le Chatelier et al., 2013; Turnbaugh et al., 2009), rheumatoid arthritis (Scher et al., 2013), and cancer (Castellarin et al., 2012). A better understanding of how microbial community composition affects human health and disease promises to lead to better disease prevention and treatment.

Although microorganisms function within complex communities, traditionally microbiology relied upon culturing techniques to obtain large numbers of clonal microorganisms to study. Once these microorganisms were grown, their genetic, morphological, and physiological characteristics could then be more readily studied. This focus on a single species in pure culture led to understanding how specific species functioned in the laboratory in great detail. However, what was missing was how species, and entire communities, function in their natural environment.

1.2. History of microbial community sequencing

Cultivation independent methods were developed due to the realization that some molecules are essential to life, so will be conserved in all organisms of a certain type (such as bacteria and archaea) and due to the slow rate of evolution in these molecules, they could serve as an evolutionary marker molecule. For example, the 16S ribosomal RNA sequence is a component of the 30S small subunit of bacterial and archaeal ribosomes, which is responsible for translation. The 16S rRNA gene is thus conserved and undergoes slow rates of evolution, properties which Carl Woese and

George Fox capitalized on when they pioneered the use of the 16S rRNA for reconstructing phylogenies (Woese and Fox, 1977).

Norman Pace and his colleagues conducted a variety of studies exploring the diversity in microbial life through sequencing 16S rRNA, including sequencing of bulk DNA from an environmental sample (Lane et al., 1985; Schmidt et al., 1991). These studies of microbial diversity led to the realization that standard techniques for cultivation at the time were only allowing cultivation of as little as 1% of bacterial and archaeal microorganisms observable in nature (Hugenholtz et al., 1998). However, researchers have since managed to culture many microorganisms which were previously considered “unculturable”. These successes were due to a variety of factors, such as the use of low concentration of nutrients (Rappé et al., 2002) or identification of the exotic substrate that a microbe relies on exclusively (Mägli et al., 1996). These advancements led to breakthroughs such as the cultivation of a haloarchaeon that, although abundant in salt lakes around the world, no one had managed to culture in the 25 years since its discovery (Bolhuis et al., 2004; Burns et al., 2004).

It wasn’t until the early 2000s that environmental shotgun sequencing of microbial communities was first undertaken. These studies were done with Sanger sequencing, and due to the cost, initially began with sequencing of viral communities (Breitbart et al., 2002) or relatively simple communities of bacteria and archaea (Tyson et al., 2004). Larger projects were possible only for those with access to large amounts of funds, such as Craig Venter’s project to sequence the oceans across the globe. This started with the Sargasso Sea pilot sampling project in 2003 (Venter et al., 2004), followed by the Global Ocean Sampling Expedition (GOS) starting in 2004. The GOS was a 2-year ocean exploration genome project aiming to assess marine microbial biodiversity in order to allow greater understanding of how ecosystems function, and to discover new genes of ecological and evolutionary importance (Rusch et al., 2007). Generating over 7.7 million sequencing reads, the GOS produced over 90 times the sequencing data of the next largest marine metagenomic dataset at the time (DeLong et al., 2006); the generation of the data from the first leg of the expedition was estimated to cost over $10 million (Knight et al., 2012).

With the advent of high-throughput sequencing, both metagenomics—environmental shotgun sequencing—and microbial profiling through sequencing the 16S

rRNA gene became cheaper resulting in a larger number of large scale projects. The first metagenomics studies utilizing next-generation or high-throughput sequencing were performed in 2005 (Poinar et al., 2006), and due to the drop in sequencing costs associated with the technology, most studies soon started using high-throughput sequencing. This increase in affordability allowed several large-scale projects to be undertaken. Examples are the Metagenomics of the Human Intestinal Tract (MetaHIT) project (Qin et al., 2010), involving the European Union and China, and the Human Microbiome Project (HMP), a United States National Institute of Health initiative (NIH HMP Working Group et al., 2009; Turnbaugh et al., 2007). Other initiatives include the TerraGenome consortium (Vogel et al., 2009), coordinating activities of researchers and projects around the world working on sequencing the soil metagenome, such as the MetaSoil and Canadian MetaMicroBiome Library projects (Delmont et al., 2011; Neufeld et al., 2011). The Tara Oceans Project initiated in 2009, collected over 30,000 samples from 210 sites in every major oceanic region, examining viruses, prokaryotes, and picoeukaryotes (Bork et al., 2015). A final example is the ongoing Earth Microbiome Project, an ambitious, massively collaborative project utilizing crowd-sourced samples, with the aim of “constructing a global catalogue of the uncultured microbial diversity of this planet” (Gilbert et al., 2014).

The ever-increasing throughput of sequencing technologies is allowing microbial sequencing and the study of microbial communities to be applied to many problem areas, from an increased understanding of the process of fermenting foods (Wolfe and Dutton, 2015) or ripening of cheese (Fuka et al., 2013), to improving water quality testing.

1.3. Microbial waterborne pathogens

Many people struggle to obtain access to clean water. Although in many parts of the world, such as Canada, most people have access to a safe supply of water, worldwide, access is still a problem. For example, in 2006 it was estimated that 1.5 million children die each year from diarrheal diseases, and 2.5 billion people had no access to improved sanitation (Fenwick, 2006). Waterborne diseases are caused by infections predominantly transmitted through contact with or consumption of contaminated water. If we consider viruses to be included as microorganisms, freshwater microbial pathogens include a variety of protozoan, bacterial, and viral

pathogens. Protozoan pathogens include *Entamoeba histolytica*, which causes amoebic dysentery (amoebiasis); *Cryptosporidium parvum*, which causes cryptosporidiosis; and *Giardia lamblia*—the most common intestinal parasite—which causes giardiasis, also known as beaver fever. Viral pathogens include such viruses as Hepatitis A and viruses associated with gastroenteritis such as norovirus and enteroviruses. There are a number of bacterial waterborne pathogens, such as *Escherichia coli*, *Legionella* spp., *Vibrio cholera*, and *Shigella dysenteriae*. As section 3.2 of this thesis focuses on analysis of *Legionella* in the samples collected from the Applied Metagenomics of the Watershed Microbiome project, a more in depth description is provided for *Legionella* below. Furthermore, detection of *E. coli* is routinely used in water quality monitoring so *E. coli* will also be provided a more in depth description.

1.3.1. *Legionella*

The genus *Legionella* is a group of bacteria which are opportunistic pathogens that cause legionellosis. Legionellosis is any illness caused by *Legionella*, which includes Legionnaire's disease and Pontiac fever; Legionnaire's disease is a severe multisystem illness involving atypical pneumonia and systemic infection caused by *Legionella* bacteria (Fraser et al., 1977), while Pontiac fever is a self-limited acute respiratory disease causing a mild upper respiratory tract infection resembling acute influenza (Glick et al., 1978). It is not understood why *Legionella* may result in either Legionnaire's disease or Pontiac fever, and it is possible to have simultaneous outbreaks of both illnesses from the same source (Euser et al., 2010). Legionellosis outbreaks are associated primarily with man-made environments such as cooling towers, swimming pools, and whirlpool spas (Fields et al., 2002). Transmission of *Legionella* is through inhalation of aerosolized water droplets from a contaminated source, and person-to-person transmission does not seem to occur.

Legionella are Gram-negative bacteria that are common in many freshwater environments and soil. They are facultative intracellular parasites of freshwater protozoa, particularly amoebae (Rowbotham, 1980). These hosts protect *Legionella* from harsh environmental conditions, within which they are able to multiply intracellularly (Borella et al., 2005). *Legionella* have been described as accidental pathogens of humans as they evolved to parasitize protozoa, and these traits allow them to infect humans; however,

infection of humans is a dead end for replication due to the lack of person-to-person transmission (Chien et al., 2004)

The genus *Legionella* was established after a large outbreak of pneumonia among attendees of a convention of the American Legion, leading to the hospitalization of 147 people and resulting in 29 deaths (Fraser et al., 1977). Following months of investigation, the causative agent was identified as the previously unrecognized *Legionella pneumophila* (Brenner et al., 1979). Although there are now over 60 identified species of *Legionella*, because *L. pneumophila* was the first described and is responsible for most of the reported cases of legionellosis (Marston et al., 1994), *L. pneumophila* has been the most well studied, so the molecular mechanisms of its pathogenesis will be described.

When *L. pneumophila* enters the lung, it is consumed by alveolar macrophages by conventional phagocytosis, although coiling phagocytosis has also been observed (Horwitz, 1984). Once internalized, the *Legionella*-containing vacuole evades the lysosomal network, and instead fuses transiently with mitochondria and vesicles derived from the endoplasmic reticulum. The recruitment of endoplasmic reticulum, starting within minutes of phagocytosis, is accomplished through intercepting endoplasmic reticulum vesicles (Kagan and Roy, 2002). Eventually the membrane surrounding the *Legionella*-containing vacuole comes to closely resemble rough endoplasmic reticulum in appearance, studded with ribosomes, and the bacteria replicate in the vacuole (Isberg et al., 2009). Multiplication of *L. pneumophila* eventually leads to the death of the macrophage, and the bacteria move on to infect other macrophages.

1.3.2. Escherichia coli

E. coli is a gram-negative, facultatively anaerobic, rod-shaped bacterium that is commonly found in the lower intestine of endotherms such as humans and cows. It can be grown and cultured easily, and has been intensively studied for decades; in fact, *E. coli* is the most widely studied bacterial species. Most *E. coli* strains are harmless, and as part of the normal microbiota of the gut, can even benefit their hosts, for example by preventing colonization and infection by pathogenic bacteria such as *Salmonella typhimurium* (Hudault et al., 2001). However, some strains can cause diseases such as gastroenteritis or urinary tract infection, and can cause outbreaks resulting in food recalls

or boil water advisories. One of these strains, *E. coli* O157:H7, produces the shiga-like toxin or verocytotoxin, which acts on the lining of blood vessels, interacting with and inactivating ribosomes (Nguyen and Sperandio, 2012). The resulting halt in protein synthesis leads to the death of these cells, causing hemorrhagic diarrhea and occasionally hemolytic uremic syndrome, a life-threatening complication involving hemolytic anemia, acute kidney failure, and low platelet count.

A well-known outbreak in Canada due to *E. coli* O157:H7 is the one that occurred in 2000 in the small town of Walkerton, Ontario. This outbreak was the result of *E. coli* O157:H7 contaminating the water supply, leading to several deaths and nearly half of the town's approximately 5000 residents becoming ill (Salvadori et al., 2009). Although this event, which became known as the Walkerton tragedy, occurred in large part due to several incidents of human error, it still underscores the importance of water quality monitoring.

1.4. Water quality testing

Fresh water is one of the world's most important natural resources. Not only is it essential to life, but it also supports numerous industries, from commercial fisheries and agricultural, to forestry and recreational uses such as waterparks. The quality of this water is what determines its use to humans, either as drinking water or its use for economic opportunities. Despite the immense importance of high quality freshwater, it is estimated that 80% of the world's population is exposed to threats to their water security (Vörösmarty et al., 2010). Even in North America, the largest waterborne outbreak resulted in over 400,000 Milwaukee, Wisconsin, residents becoming ill with cryptosporidiosis resulting in a total outbreak-associated illness cost of 96.2 million US dollars from both medical and productivity losses (Corso et al., 2003). Here in British Columbia, a single case of gastrointestinal illness costs $1,343 (Henson et al., 2008).

Current methods for detection of fecal pollution in drinking water involve testing samples from the tap, rather than the source watershed. Detection of pathogens is not used for monitoring due to the number of pathogens that would need to be tested for, and laboratory methods are unreliable due to the low pathogen concentrations in the water that can cause illness. Instead, methods rely on culture-based testing of indicators such as fecal coliforms or *E. coli*, which are used as bacterial surrogates for the

presence of feces (Stelma and Wymer, 2012). However, there are a number of drawbacks to these tests. They can be slow and may lack sensitivity and specificity. Results for culture based tests are available only after many hours, so by the time the test results are obtained, contaminated water may already have been consumed. Furthermore, the tests correlate poorly with the presence of pathogens such as protozoan parasites and viruses (Harwood et al., 2005). For example, *E. coli* has been shown to persist and sometimes even grow in the environment (Ishii et al., 2006; Wheeler Alm et al., 2003), so the presence of *E. coli* does not necessarily imply the presence of fecal pollution or pathogens. Similarly, the absence of *E. coli* does not mean the absence of more resistant organisms such as *Giardia* cysts or *Cryptosporidium* oocysts (Wilkes et al., 2009).

1.5. The Applied Metagenomics of the Watershed Microbiome project

In contrast to the traditional culture-based testing, a molecular-based test may be more rapid, accurate, and useful for understanding the source of fecal pollution. The goal of the Applied Metagenomics of the Watershed Microbiome project, which drove much of the research of this thesis, was to use metagenomics to discover novel indicators of fecal pollution in watersheds to be used to build a molecular-based diagnostic test of watershed health. As part of this project, 112 water samples were collected from seven sites in a protected, an urban, and an agricultural watershed over a one-year period (Figure 1.1). For each sample, physical and chemical parameters were recorded, and protist, bacterial, and viral genetic material was extracted and sequenced (Table 1.1).

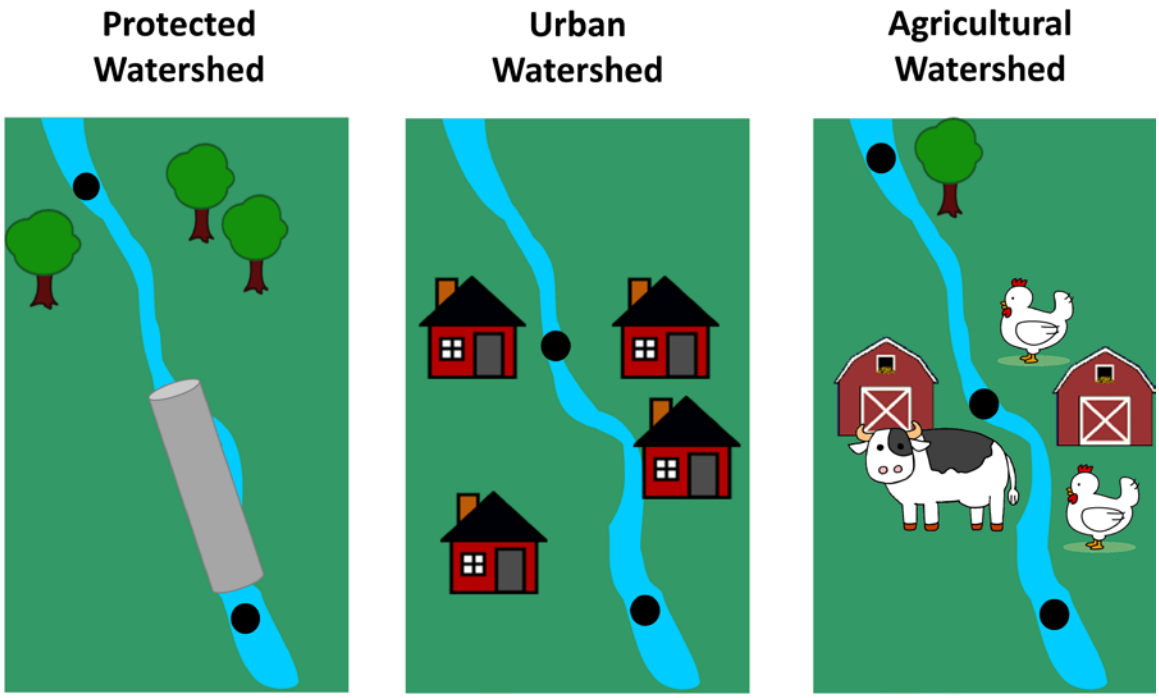

Figure 1.1 Sampling sites in the three different watersheds.
The protected watershed contains two sampling sites: a protected site, and a downstream site collected after reservoir water has passed through a 9-km long pipe. The urban watershed also contains two sampling site: a polluted site collected from a stream that passes through residential development, and a downstream site. The agricultural watershed contains three sampling sites: a site upstream of agricultural pollution, a polluted site surrounding by intense agricultural activity, and a downstream site.

Table 1.1 Environmental variables collected for each of the watershed samples

Type	Parameter	Method/Source
Microbiological	*E. coli,* fecal coliforms, chlorophyll a	Tested at BC PHL (Colilert)
Chemical	Nitrate and nitrite, ammonia, total phosphorous and chloride	Tested at UBC and private laboratory (Maxxam)
Physical & Chemical	pH, total dissolved solids (conductivity), turbidity, water temperature and dissolved oxygen	Hand-held probes on site
Physical	Rainfall data	Canadian Climate Data database

The microbial communities from the collected water samples were analyzed using both amplicon and shotgun metagenomics sequencing. One of the questions to be addressed by the project was which methods or tools should best be used for the taxonomic classification of metagenomics sequences. This was a pertinent question because many methods were being developed, yet when the project began, there was not yet any published comprehensive evaluation of taxonomic classification methods for metagenomics sequences.

1.6. Metagenomics sequence classification

Metagenomics sequence taxonomic classification methods generally fall into four categories, reflecting their different strategies: (1) sequence composition based methods, which are based on characteristics of their nucleotide composition (e.g. tetranucleotide usage or codon usage) (Mande et al., 2012), (2) sequence similarity based methods, which use the results of a sequence similarity search against a database of a reference set of sequences, (3) hybrid methods which incorporate components of the first two, and (4) marker-based methods which identify species based on the occurrence of certain specific marker sequences.

Composition methods generate models from the reference organisms' genomes, and will classify the input sequence reads based on which model(s) fit the read best. They have had trouble with classifying reads of short length (<1000 base pairs), with Phymm being the first method published demonstrating reasonable accuracy at short read lengths (Brady and Salzberg, 2009).

Sequence similarity based methods, on the other hand, perform very well at identifying reads from genomes within the reference database that they search against, even at read lengths as short as 80 base pairs (Ander et al., 2013). However, many reads from metagenomics samples come from genomes that are not in any reference database (Rappé and Giovannoni, 2003). Similarity based methods have traditionally used BLAST (Altschul et al., 1990), and have been generally slower to run compared to composition based methods.

Hybrid methods combine the similarity approach and the composition approach, with the goal of improving classification or speed. For improving classification, scores

may be combined from both the similarity portion and the composition portion of the method for each prediction (Brady and Salzberg, 2009). Another hybrid strategy, aimed at increasing speed, is to use the composition approach to narrow down the set of candidate organisms, and thus have the similarity search occur against a fraction of the original database (Mohammed et al., 2011a).

A related group of methods try to determine community composition from metagenomes by utilizing marker genes. These methods differ from methods that perform taxonomic classification, as they do not to try to classify all of the reads. Instead, they focus on classifying only marker genes to try to determine the microbial community composition of the sample. Most marker based approaches utilize universal genes. However, another approach, utilized by MetaPhlAn, involves use of clade-specific marker genes (Segata et al., 2012).

The first step in a marker based approach is to identify reads that hit to one of the markers. As the size of the reference database of markers used by these methods is relatively small, these methods are comparatively quick to run. In addition to focusing on a limited set of markers, which greatly reduce the computational cost of analysis, these methods are not affected by differences in genome size. If the goal of the analysis is to identify the community composition of the sample, taxonomic classification methods are biased by genome sizes, as organisms with larger genomes will generate more reads. Amplicon sequencing using the 16S rRNA gene also suffers bias due to variability in 16S rRNA copy number (Větrovský and Baldrian, 2013). Thus, marker based approaches using shotgun metagenomics sequencing data may provide the least biased relative abundance information for organisms in the community.

Although there have been many methods developed for the taxonomic classification of metagenomics sequence reads, there is a notable lack of programs for certain other forms of analysis of metagenomics sequences, such as the prediction of subcellular localization of proteins encoded by metagenomics sequences.

1.7. Protein subcellular localization prediction

Identification of the subcellular localization (SCL) of a protein aids in determination of the protein's function, and has many other practical applications. For

example, these cell surface and secreted proteins are involved in interactions with the environment, which in the case of pathogens, are often host-pathogen interactions such as adherence, toxin synthesis, invasion of host cells, and defending against the host's responses. Therefore, these secreted and cell surface proteins are targets for drug and, due to their interaction with the immune system, vaccine development (Maione et al., 2005; Rodríguez-Ortega et al., 2006; Vogel and Claus, 2011). Furthermore, secreted and cell surface proteins may serve as diagnostic biomarkers, for both pathogens or environmental microorganisms (Segata et al., 2011). A final example of the utility of cell surface and secreted proteins is their use in a variety of industries, such as the detergent, pharmaceutical, food, pulp, and biofuel industries (de Champdoré et al., 2007; Graham et al., 2011; Uthandi et al., 2010).

PSORT I was the first comprehensive protein subcellular localization predictor developed, generating predictions for multiple localizations (cytoplasmic, cytoplasmic membrane, periplasm, and outer membrane) in Gram-negative bacteria using "if-then" rules derived from experimental observations, and included features such as hydrophobicity, sequence motifs, and sorting signals (Nakai and Kanehisa, 1991). Since PSORT I was first released in 1991, there have been a variety of protein subcellular localization programs developed. Predictors use an assortment of methods, such as neural networks (Reinhardt and Hubbard, 1998), support vector machines (Bhasin et al., 2005), adaboost (Niu et al., 2008), the k-nearest neighbours algorithm (Nakai and Horton, 1999), and ensemble approaches integrating the results of multiple classifiers (Wang et al., 2015). One of these ensemble methods for protein subcellular localization prediction, PSORTb, was developed by the Brinkman lab.

1.7.1. PSORTb

The first version of PSORTb was introduced in 2003, and was the first method to make predictions for all five localization sites characteristic of Gram-negative bacteria (Gardy et al., 2003). PSORTb is composed of several modules, examining input protein sequences for the presence of signal peptides, similarity to proteins of known localization, amino acid composition, and transmembrane α-helices and motifs corresponding to specific localizations. It integrates the results of these modules using a Bayesian network to produce a final probability value for each localization site. The evaluation of PSORTb used 5-fold cross validation, whereby the training dataset was

divided into 5 subsets, and 1 of the subsets was used to test the model, while the other 4 subsets were combined and used to build the model. This is then repeated 4 more times (for a total of 5 rounds), so that each subset is the test subset exactly once. PSORTb was designed to favour precision, and to make only predictions that it was confident in making. Thus, the results of the 5-fold cross validation were that PSORTb attained an overall precision of 97% while maintaining an overall recall of 75%.

PSORTb 2.0 was then released in 2005, extending to make predictions for Gram-positive as well as Gram-negative bacteria, and improving performance by increasing recall while maintaining a high level of precision (Gardy et al., 2005). This improved performance was achieved through an expanded training dataset, expanded similarity search (SCL-BLAST) and motif-based analyses, and replacement of a single cytoplasmic composition-based support vector machine (SVM) with multiple SVMs, one for each localization site, that used a feature space comprising of frequent subsequences rather than overall amino acid composition.

The most recent version of PSORTb, version 3.0, was released in 2010. It extended PSORTb 2.0 to make predictions for Archaea and bacteria with atypical membrane/cell wall topologies (Yu et al., 2010b). Furthermore, important subcellular localization subcategories were introduced, such as host-associated and flagellar, and one again the training dataset was expanded, improving recall while maintaining a high level of precision. Various other changes were also made, such as improvements to software usability and modifications to certain modules, an example being the replacement of a trans-membrane α-helix predictor with a newer, open source trans-membrane α-helix predictor.

Starting with PSORTb 2.0, the manually curated training sets used for building PSORTb and evaluating it were made available in a database called PSORTdb (Rey et al., 2005a). The PSORTdb database was created in the hopes that it would be useful for those studying microbes as well as those working on subcellular localization prediction, for its use as a training dataset for new predictors.

1.7.2. PSORTdb

PSORTdb contains two independent but linked databases: a manually curated database of proteins of known localization that have been determined through laboratory experimentation (ePSORTdb), and a database of proteins computationally predicted using PSORTb (cPSORTdb). The first release of PSORTdb contained 2171 proteins in ePSORTdb; many of these entries were obtained by automatically parsing bacterial proteins for their subcellular localization annotations from Swiss-Prot (Boeckmann et al., 2003), and verifying annotations through a search of the literature. Additional protein annotations were obtained through curation of other literature sources such as microbiology textbooks and reference articles. The proteins in cPSORTdb came from 140 completely sequenced bacterial genomes that were analyzed by PSORTb 2.0.

Following the release of PSORTb 3.0, PSORTdb was updated and expanded in 2011 (Yu et al., 2011). The database of protein subcellular localization verified by laboratory experimentation (ePSORTdb) was substantially expanded, from 2171 entries to over 13000 entries. These additional entries came from an updated version of Swiss-Prot (version 49), and protein localization data obtained from manual literature search. cPSORTdb was expanded by running the newest version of PSORTb, PSORTb 3.0 (Yu et al., 2010b), on the over 1000 proteomes that were available from complete bacterial and archaeal genomes.

An additional feature added to PSORTdb 2.0 was the incorporation of a computational ‘outer membrane detection’ procedure to automatically determine what ‘Gram-stain’ or cellular structure a given bacterial proteome should be analyzed as. This novel method involves using a BLAST search to check for the presence of the *omp85* gene or its orthologs, which code for the outer membrane protein Omp85, in microbial genomes. The presence or absence of Omp85 determines the cellular structure of a bacterium because it is essential to outer membrane biogenesis and viability of Gram-negative bacteria (Voulhoux et al., 2003). A set of four diverse Omp85 proteins needed to be incorporated (from *Neisseria gonorrhoeae, Thermosipho africanus, Synechococcus sp.* PCC 7002 *and Thermus thermophiles*), to ensure 100% precision and 100% recall on a test dataset of 813 bacterial proteomes that had been curated regarding the absence or presence of an outer membrane.

1.8. Goals of present research

When this research began, there had not yet been any comprehensive evaluation of metagenomics methods for taxonomic classification. Thus, as part of the Applied Metagenomics of the Watershed Microbiome project, I performed an evaluation of methods, in part to help us decide which method or methods to utilize for the project. Due to the notably different approaches taken by different methods, the hypothesis of this work was that methods would vary substantially in various aspects of performance, such as sensitivity and length of time the method takes to run. This evaluation of methods is detailed in Chapter 2.

Various aspects of the work that I did on the Applied Metagenomics of the Watershed Microbiome project are described in Chapter 3. The hypothesis of the project was that it would be possible to identify biomarkers of watershed health. Section 3.1 examines the microbial profiles from the samples collected monthly from the protected, agricultural, and urban watershed sites over a 1-year period. Overall ecological trends are described and a comparison is performed between the profiles derived from shotgun sequencing data and 16S rRNA sequencing. Section 3.2 outlines the development of biomarkers to distinguish low water quality from high water quality freshwater samples. In section 3.3 I examine genera of known waterborne bacterial pathogens present in the watershed samples; this section is focused primarily on *Legionella*, which was found at relatively high abundance in all watershed sampling sites.

One of the top requests received for the development of PSORTb has been a metagenomics version of PSORTb, in part because there is a notable lack of metagenomics subcellular localization predictors. Chapter 4 describes the development of PSORTm, a version of PSORTb for metagenomics sequences, and its application to the watershed metagenomics data described in Chapter 3.

Chapter 5 describes Identification of markers for atypical cell envelopes, and their use in expanding our database of Bacteria and Archaea protein subcellular localization, PSORTdb, to better reflect the diversity in cell envelope structures.

Chapter 2.

Evaluation of shotgun metagenomics sequence classification methods

Chapter 2 presents an evaluation of metagenomics taxonomic classification methods. Portions of this chapter have been previously published in the article "Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities" by M.A. Peabody, T. Van Rossum, R. Lo, and F.S.L. Brinkman in BMC Bioinformatics, 2015, 16:362. © 2015 Peabody et al.

I completed all work presented in this chapter with the following exceptions: Thea Van Rossum assisted in writing scripts to create the clade exclusion scenarios, and Raymond Lo created the in vitro mock communities.

2.1. Abstract

Background

The field of metagenomics (study of genetic material recovered directly from an environment) has grown rapidly, with many bioinformatics analysis methods being developed. To ensure appropriate use of such methods, robust comparative evaluation of their accuracy and features is needed. For taxonomic classification of sequence reads, such evaluation should include use of clade exclusion, which better evaluates a method's accuracy when identical sequences are not present in any reference database, as is common in metagenomic analysis. To date, relatively small evaluations have been performed, with evaluation approaches like clade exclusion limited to assessment of new methods by the authors of the given method. What is needed is a rigorous, independent comparison between multiple major methods, using the same *in silico* and *in vitro* test datasets, with and without approaches like clade exclusion, to better characterize accuracy under different conditions.

Results

An overview of the features of 38 bioinformatics methods is provided, evaluating accuracy with a focus on 11 programs that have reference databases that can be modified and therefore most robustly evaluated with clade exclusion. Taxonomic classification of sequence reads was evaluated using both *in silico* and *in vitro* mock bacterial communities. Clade exclusion was used at taxonomic levels from species to class – identifying how well methods perform in progressively more difficult scenarios. A wide range of variability was found in the sensitivity, precision, overall accuracy, and computational demand for the programs evaluated. In experiments where distilled water was spiked with only 11 bacterial species, frequently dozens to hundreds of species were falsely predicted by the most popular programs. The different features of each method (forces predictions or not, etc.) are summarized, and additional analysis considerations discussed.

Conclusions

The accuracy of shotgun metagenomics classification methods varies widely. No one program clearly outperformed others in all evaluation scenarios; rather, the results illustrate the strengths of different methods for different purposes. Researchers must appreciate method differences, choosing the program best suited for their particular analysis to avoid very misleading results. Use of standardized datasets for method comparisons is encouraged, as is use of mock microbial community controls suitable for a particular metagenomic analysis.

2.2. Background

2.2.1. Introduction

Metagenomics involves collecting samples from an environment (water, saliva, etc.) and then extracting and studying the genetic material from the microorganisms present in these samples (Wooley et al., 2010). This approach is transforming microbiology, ecology, medicine, and other research areas investigating various microbiomes, allowing us to analyze for the first time microbial species, including those not culturable, at a level of detail not previously possible (Handelsman, 2004). Metagenomics sequence reads can be taxonomically classified to identify the microbes, or functionally classified (gene functions, metabolic pathways, etc.) to identify the functional potential of the community. There exist two general approaches for characterizing the taxonomic content of environmental samples: (1) sequencing of PCR amplicons corresponding to phylogenetic marker genes (e.g. 16S rRNA; “amplicon analysis”); (2) shotgun sequencing whereby all genomic DNA in the community is sequenced. A drawback of the shotgun sequencing approach is increased cost, but advantages include the ability to gain insights into metabolism and gene function through functional classification, and the avoidance of potentially biased amplification steps (Acinas et al., 2005). Furthermore, a notable subset of taxa cannot be captured by traditional 16S sequencing owing to divergent 16S rRNA gene sequences (Brown et al., 2015). This, combined with the continuing decrease in cost of sequencing, may result in shotgun metagenomics becoming increasingly used for the taxonomic classification of microbial communities.

Taxonomic classification methods generally fall into four categories, as discussed in section 1.6. The first of these is sequence composition based methods, which use characteristics of nucleotide composition, and tend to have difficulty classifying sequences with shorter lengths (Mande et al., 2012). The second category is sequence similarity based methods, which tend to perform fairly well even at lower read lengths (Ander et al., 2013). The third category is hybrid methods, which incorporate components of the first two, in order to increase either efficiency/run time or accuracy. Finally, the fourth category is marker-based methods which utilize certain specific marker sequences to identify taxa.

2.2.2. Tools vary in several additional characteristics which may influence researcher's choice

In addition to the class of method, there are many other characteristics which may affect the consideration of which method to use. For example, whether a method is available via a GUI (graphical user interface), command line, or web server can be an important consideration, as is whether the method can also perform functional (gene function) classification, or how much memory and compute time the method requires. In addition, some methods are limited to certain groups of microbes. Some methods, such as AMPHORA2 (Wu and Scott, 2012), are limited to analysis of Bacteria and Archaea. Others, such as PhyloSift (Darling et al., 2014), can additionally predict Viruses and Eukaryotes. Furthermore, some methods continue to be supported while others are not, and some eventually become unavailable or difficult to access.

Another distinction that can be made is between methods which are rank-flexible, versus rank-specific. Rank-flexible methods vary the rank at which reads are predicted by classifying each read to the lowest taxonomic level at which the given method is confident. An example of a simple rank-flexible method is the lowest common ancestor (LCA) approach, first used by MEGAN (Huson et al., 2007). This approach takes the set of taxa that the read hit in the similarity search (taking only those hits scoring within a threshold of the top hit), and assigns the read to the LCA of this set. In contrast, rank-specific methods give the same rank predictions for all reads.

2.2.3. Clade exclusion is an important technique to evaluate how well methods will perform on environmental samples

Sequence similarity based methods perform very well when identifying query reads identical to genomes/sequences within the reference database that they search against. However, because the majority of microorganisms have not yet had their genome sequenced, in most environments many of the sequence reads that would be generated in a metagenomics experiment would be quite unrelated to any sequences that are in a reference database, or at minimum not identical (Amann et al., 1995). Thus, one of the approaches used in the evaluation of taxonomic classifiers is clade-level exclusion. This involves removing all sequences from a database at a certain taxonomic level and then evaluating the ability to make predictions at higher taxonomic levels. For

example, if performing species level exclusion for *Pseudomonas aeruginosa*, all *Pseudomonas aeruginosa* genome sequences would be removed from the reference database and/or models of the methods being evaluated. Then, the method's ability to classify reads from *Pseudomonas aeruginosa* at higher taxonomic levels (i.e., *Pseudomonas*, *Pseudomonadaceae*, etc.) would be evaluated. Such clade exclusion methodology is one way to avoid obtaining artificially high accuracy levels caused by the problem of testing and training with identical data.

2.2.4. The present work builds upon a previous evaluation performed without clade exclusion

There has been one previous evaluation of metagenomics bioinformatics methods reported (Bazinet and Cummings, 2012). This study was an important first step in comparing many metagenomics classification tools; however, the microbial genomes used in the analysis were found in the reference databases and training sets of the methods evaluated. This means that the accuracy of the methods shown from the study will be considerably higher than when they are used to classify reads from organisms not in the reference databases or training sets. Samples from most environments, such as soil, ocean, and freshwater samples, are very diverse and the majority of organisms existing in these environments have not been characterized. The human gut is an environment in which intense research interest has resulted in substantial effort to sequence relevant microbes (Human Microbiome Jumpstart Reference Strains Consortium et al., 2010); however, even in the human gut, it appears that many species are not present in reference databases (Sunagawa et al., 2013). In addition, the previous comparison relied solely on *in silico* simulated reads. As sequence simulators cannot capture all of the factors that may affect read sampling in metagenomics, *in vitro* communities (i.e., samples of known bacterial cultures spiked into distilled water and sequenced) are an important complementary set of data on which to evaluate methods.

In the present study, a variety of metagenomic taxonomic classification methods are evaluated on mock communities simulated both *in silico* and *in vitro*. The performance of the methods in terms of their sensitivity, precision, and number of incorrectly predicted species are analyzed. In addition, the performance of the methods is compared as simulated read length is increased, and level of clade exclusion is varied. Methods evaluated more fully were chosen to encompass the range of types of

methods available, as well as based on their popularity, and amenability to clade exclusion. We demonstrate how the accuracy of shotgun metagenomics classification methods varies widely. No one program clearly outperformed others in all evaluation scenarios, rather the results illustrate the strengths and weaknesses of different methods for different purposes – information critical for researchers to be aware of when performing their particular analysis.

2.3. Methods

2.3.1. Simulation of MetaSimHC and freshwater *in silico* and *in vitro* datasets

Two different microbial communities were used for this evaluation, both made up of diverse taxa for which completed genome sequences were available. The first was previously proposed as a “high complexity” dataset in (Richter et al., 2008), and will be referred to as MetaSimHC. This was chosen since it has been proposed to be a reference dataset for analysis of methods, and consists of diverse microbial species covering several phyla of both Bacteria and Archaea. The second was chosen with the aim of having a set of species commonly found in freshwater, suitable as a control for a watershed metagenomics project we participated in (Van Rossum et al., 2015). This was done by identifying species that were common among several publicly available freshwater datasets (Ghai et al., 2011; Oh et al., 2011; Smith et al., 2012), and will be referred to as FW (freshwater). The organisms used in each of these datasets can be found in Table 2.1.

Both of these datasets were simulated using MetaSim (version 0.9.5; (Richter et al., 2008)) at sequence lengths of 100, 250, 500, and 1000 bp, with each organism at 1X coverage. Although the sets of sequences of differing read length were generated independently, they are generated at 1X coverage so the effects of sampling only portions of genomes that are predicted particularly well or poorly should be mitigated. No error model was used, because there was not an error model for Illumina reads at the longer read lengths (500 and 1000), and we wanted to be consistent as read length was varied. Also, the *in vitro* dataset gives us data off of an actual sequencer which allows us to see how methods perform on data with real sequencing errors. Clade exclusion was performed at the level of species, genus, family, order, and class.

The FW dataset was simulated both with MetaSim (FW *in silico*) and an *in vitro* mock community (FW *in vitro*). To construct the FW *in vitro*, the bacteria were grown up in pure culture, and then their DNA were extracted and spiked in equal concentrations into sterile, distilled water for sequencing. All complete bacterial and archaeal genomes were downloaded from NCBI on June 17, 2013, for the creation of databases and supervised models used in the different methods. The numbers of genomes left in the databases and training sets of the methods in the evaluation scenarios are shown in Table 2.2. The datasets used in these evaluation scenarios have been deposited to the MG-RAST database and accession numbers can be found in Table 2.3, and the number of reads simulated from each organism for the *in silico* datasets can be found in Table 2.4. Note that while certainly test datasets could be constructed using a larger number of species, it is non-trivial to construct a similar *in vitro,* mock community dataset using a high number of species. We purposefully constructed our dataset to contain taxa with a variety of levels of divergence from one another, including closely related species (i.e. multiple species from the *Pseudomonas* genera). The latter helps evaluate the ability of methods to handle taxa prediction when closely related taxa are present.

Because there is such a large difference in microbial communities (e.g. soil versus acid mine drainage) in terms of number of organisms, which organisms are present, their taxonomic novelty, and diversity in terms of abundance distribution, it is not possible to simulate communities that will be appropriate for all environmental communities. This is why we suggest researchers test their own mock communities that approximate their expected community.

Table 2.1 Microbes used in the two simulated mock communities

MetaSimHC[a]			Freshwater[b] (FW) *in silico* and *in vitro*		
Genus	Species	Strain	Genus	Species	Strain
Agrobacterium	*tumefaciens*	C58	*Bacillus*	*amyloliquefaciens*	FZB42
Anabaena	*variabilis*	ATCC 29413	*Bacillus*	*cereus*	ATCC 14579
Archaeoglobus	*fulgidus*	DSM 4304	*Burkholderia*	*cenocepacia*	J2315
Bdellovibrio	*bacteriovorus*	HD100	*Escherichia*	*coli*	K-12
Campylobacter	*jejuni*	81-176	*Frankia*	*sp.*	CcI3
Clostridium	*acetobutylicum*	ATCC 824	*Micrococcus*	*luteus*	NCTC 2665
Lactococcus	*lactis*	SK11	*Pseudomonas*	*aeruginosa*	PAO1
Nitrosomonas	*europaea*	ATCC 19718	*Pseudomonas*	*aeruginosa*	UCBPP-PA14
Pseudomonas	*aeruginosa*	PA7	*Pseudomonas*	*fluorescens*	Pf-5
Streptomyces	*coelicolor*	A3(2)	*Pseudomonas*	*putida*	KT2440
Sulfolobus	*tokodaii*	str. 7	*Rhodobacter*	*capsulatus*	SB 1003
			Streptomyces	*coelicolor*	A3(2)

[a]MetaSimHC is a test dataset of 11 diverse microbial genomes covering several phyla of Bacteria and Archaea, proposed in (Richter et al., 2008).
[b]Freshwater (FW) is a set of bacterial genomes found in previous freshwater metagenomics studies (see methods).

Table 2.2 Number of genomes left in the reference databases and training sets of the methods used in the evaluation scenarios

Rank of clade exclusion[a]	Number of genomes	
	MetaSimHC	Freshwater (FW)
None	2499	2499
Species	2460	2388
Genus	2344	2261
Family	2198	2047
Order	1688	1695
Class	555	975

[a]Clade exclusion involves removing all sequences from a database at a certain taxonomic level. For example, if performing species-level exclusion for a particular organism, removing all of the genomes from the database of that species.

Table 2.3 Datasets used in the evaluation scenarios and their accession numbers

Dataset	Read length (bp)	MG-RAST accession number
MetaSimHC	100	4545484.3
MetaSimHC	250	4548386.3
MetaSimHC	500	4548993.3
MetaSimHC	1000	4548992.3
FW *in silico*	100	4545483.3
FW *in silico*	250	4548385.3
FW *in silico*	500	4548991.3
FW *in silico*	1000	4548990.3
FW *in vitro*	Average 223	4545485.3

Table 2.4 Number of reads simulated for each organism in the *in silico* datasets

Organism	100 bp	250 bp	500 bp	1000 bp
	Number of reads			
MetaSimHC				
Agrobacterium tumefaciens str. C58	56636	22512	11318	5720
Anabaena variabilis ATCC 29413	71330	27722	13938	7298
Archaeoglobus fulgidus DSM 4304	21978	8550	4180	2156
Bdellovibrio bacteriovorus HD100	37468	15032	7644	3804
Campylobacter jejuni subsp. jejuni 81-176	17194	6800	3414	1810
Clostridium acetobutylicum ATCC 824	41214	16942	8118	4092
Lactococcus lactis subsp. cremoris SK11	26088	10514	5134	2518
Nitrosomonas europaea ATCC 19718	27860	11316	5524	2704
Pseudomonas aeruginosa PA7	66030	26230	13688	6624
Streptomyces coelicolor A3(2)	90092	36814	18198	8838
Sulfolobus tokodaii str. 7	26830	10656	5388	2708
Total	482720	193088	96544	48272
FW *in silico*				
Bacillus amyloliquefaciens FZB42	39074	15954	7956	3856
Bacillus cereus ATCC 14579	54922	21572	10928	5564
Burkholderia cenocepacia J2315	80712	32054	15946	8244
Escherichia coli str. K-12 substr. MG1655	45768	18476	9216	4540
Frankia sp. CcI3	54242	21816	10636	5298
Micrococcus luteus NCTC 2665	25250	9750	5072	2472
Pseudomonas aeruginosa PAO1	62368	25302	12290	6214
Pseudomonas aeruginosa UCBPP-PA14	64830	26024	13304	6750
Pseudomonas fluorescens Pf-5	71150	28546	14130	7064
Pseudomonas putida KT2440	62072	24596	12224	6280
Rhodobacter capsulatus SB 1003	38642	15698	7892	3838
Streptomyces coelicolor A3(2)	90560	35870	18292	8896
Total	650516	259704	129930	65160

2.3.2. Laboratory preparation and sequencing of the mock freshwater *in vitro* community

Bacillus amyloliquefaciens FZB42 (ATCC# 23842), *Bacillus cereus* (ATCC# 14579), *Escherichia coli* K12 (ATCC# 23716), *Micrococcus luteus* NCTC 2665 (ATCC# 4698), *Pseudomonas fluorescens* Pf-5 (ATCC# BAA-477), and *Pseudomonas putida* KT2440 (ATCC# 47054) were obtained as freeze-dried stocks and used per recommended protocol to start cultures in prescribed media. *Burkholderia cenocepacia* J2315 was cultured in Luria broth at 37°C. *Frankia* sp. CcI3 was grown in liquid *Frankia* defined minimal medium (FDM) in stationary culture at 30°C for one week. *Pseudomonas aeruginosa* UCBPP-PA14 was cultured in Luria-Bertani broth at 37°C. *Rhodobacter capsulatus* SB 1003 was cultured on 0.3% yeast extract, 0.3% bactopeptone, $CaCl_2$ (1 mM) and $MgSO_4$ (1 mM) at 30°C. *Streptomyces coelicolor* A3 was cultured in 0.5% Tryptone, 0.3% yeast extract, pH 7.1 at 28°C for one week. For each of the strains of bacteria, after they were plated on the appropriate media, single colonies were picked. These were cultured overnight in 3 ml of appropriate media at the appropriate temperature (as above). *Frankia* sp. CcI3 and *P. aeruginosa* UCBPP-PA14 were cultured for several days until they reached stationary phase. The other bacteria strains were fast growing, so the starter cultures were diluted 1:100, and grown with vigorous shaking (250 rpm) to saturation overnight. Genomic DNA was extracted from these cultures with the NucleoSpin Tissue kit from Macherey-Nagel according to manufacturer's instructions. For Gram-positive bacteria, cells were pre-incubated with buffer containing 20 mg/ml lysozyme for an hour at 37°C, followed by Proteinase K at 56 °C until complete lysis was obtained. The library was prepared using a Nextera XT DNA sample preparation kit following the manufacturer's instructions. This library was sequenced with a MiSeq platform using a V2 500 cycles kit.

2.3.3. Quality control of sequenced reads

Trimmomatic-0.25 (Bolger et al., 2014) was used to (1) trim reads using a sliding window of 15 and PHRED quality score of Q<=20, followed by (2) checking if any of the last 5 bases had a Q<=5, and if so removing up to that base, and finally (3) filtering out any reads with length <85 bases. After quality control, there were 300969 reads with an average length of 223 nucleotides.

2.3.4. Evaluation of methods and metrics

Performance metrics used to evaluate different software are sensitivity, precision, taxonomic distance, and running time. Sensitivity and precision are calculated based on the numbers of true-positives (TP), false-positives (FP), and false-negatives (FN). True-positives are the number of reads assigned correctly, false-positives are the number of reads assigned incorrectly, and false-negatives are the number of reads unassigned.

Sensitivity was calculated as TP/(TP+FN), and precision as TP/(TP+FP). Taxonomic distance was calculated from correctly assigned reads as the average number of ranks above the best possible rank the assignment could be made at, and running time was the number of minutes taken for the program to complete classification.

For sensitivity, precision, and taxonomic distance, the values were averaged over all the species in the test dataset. This gave equal weighting to all of the species in the datasets; otherwise, the species with larger genomes (which have more reads) would have a larger influence on the scores. For the *in silico* datasets, reads were categorized as correctly assigned (TP) if they classified to a node (taxonomic rank) that was anywhere in the path from the correct species to the superkingdom level (e.g. Bacteria) of the NCBI taxonomic tree, and as incorrect if the read was assigned to a node that was not in this path. In the case where overpredictions were considered correct, the taxonomic level that was used to determine if a read was classified correctly was the best possible correct level that could be predicted. For example, under species clade exclusion, reads would still be classified as correct if they were in the correct genus but classified to an incorrect species.

Although most of the methods evaluated were rank-flexible in their predictions, RITA and PhymmBL are rank-specific, and thus were shown for the evaluation only where overpredictions were considered correct. Although RITA does have a rank-flexible mode, it requires having 16S rDNA profiles of a community. PhymmBL provides a confidence score which in theory could provide a cut-off for which rank to assign the reads; however, we would have had to choose the cut-offs ourselves, and previous researchers have found confidence scores to be high for a false positive dataset (Garcia-Etxebarria et al., 2014). MG-RAST was evaluated due to the popularity of the

method, but because it does not allow the user to create custom clade exclusion reference databases, it is an example of a method where we were only able to evaluate it without clade exclusion.

Table 2.5 lists the version numbers of all of the methods evaluated. All methods were run with default parameters except for filtered Kraken (Wood and Salzberg, 2014) which was run using the kraken-filter script with a threshold of 0.20, which moves assignments up to successfully higher levels of the taxonomic tree until the threshold is reached. This separate analysis was done because we noticed that Kraken was tending to overclassify reads and there was an option that would help assign reads with greater confidence. Note that some methods have variations in the way they can be run. For example, some methods can take a variety of similarity search programs as input, or have the option to utilize paired-end sequence read information. In some cases these variations had relatively small differences in sensitivity, precision, and taxonomic distance of methods, and in these cases only one of the variants was presented in the figures to be concise. Briefly, MEGAN4 (Huson et al., 2011) has the option to allow the use of paired-end information from sequence reads, and the paired-end version is presented; MetaPhyler (Liu et al., 2011) can use BLASTX, BLASTN, or a combination of the results, and the results for the BLASTX/BLASTN combination are presented; MEGAN4 and DiScRIBinATE (Ghosh et al., 2010) have the option of taking results as input from either RAPSearch2 (Zhao et al., 2012) or BLASTX, and the RAPSearch2 versions are presented. RAPSearch2 is an alternative to BLAST, which we found to run over 30 times faster than BLASTX, with comparable accuracy (see results).

Table 2.5 Methods that were the focus of this evaluation and their version numbers

Method	Version
CARMA3	3.0
CLARK	1.1.3
DiScRIBinATE	1.0
Kraken	0.10.2
MEGAN4	4.70.4
MetaBin	1.0
MetaCV	2.3.0
MetaPhyler	1.25
PhymmBL	4.0
RITA	1.0.1
TACOA	1.0
MG-RAST	3.3.7.3

Methods were run with default parameters, except for what we called “filtered Kraken”, which used the kraken-filter script with a threshold score of 0.20

2.4. Results

Table 2.6 provides an overview of methods and their features, grouped by their class. Note that it does not include all methods available, and there are more methods being continually published. Included is the number of citations each method has received, to give an indication of how much of an influence or use each method has. However, it should be noted that several of the methods have capabilities beyond just classification (such as comparisons between samples and visualization), and thus may be cited when used for purposes other than classification. Also, it is worth noting that methods that were published earlier may be highly cited, yet newer methods often improve upon their strategies. As discussed below, even with accuracy assessment aside, the different method properties can have different advantages under certain analysis scenarios and so are summarized here. Notably, many methods cannot undergo full, robust evaluation with clade exclusion, since their reference databases cannot be manipulated, and so methods chosen for full evaluation of the accuracy were limited to ones that allowed it.

Table 2.6 List of metagenomics sequence classification methods and their characteristics sorted by class of method

Method Name	Class of Method	Sequence Alignment Method / Composition Method	Standalone[a]/ Web Server	Most Recent Year Published (First Time Published) [b]	Functional Classification if Applicable	References	Number of Citations[c]
MEGAN4	Similarity	MEGABLAST, BLASTN, BLASTX, RAPSEARCH2 (Zhao et al., 2012) / N/A	Yes / No	2011 (2007)	KEGG, SEED	(Huson et al., 2007, 2009, 2011, Mitra et al., 2009, 2011)	1089
MG-RAST	Similarity	BLASTN, BLAT / N/A	No / Yes	2008	SEED, NOG, COG, KEGG	(Meyer et al., 2008)	691
CAMERA	Similarity	All six BLAST programs / N/A	No / Yes	2007 (2011)	Pfam, TIGRFAM, COG, KOG, PRK	(Seshadri et al., 2007; Sun et al., 2011)	324
CARMA3	Similarity	BLASTX, HMMER3 (Eddy, 2008) / N/A	Yes / Yes	2011 (2008)	GO	(Gerlach and Stoye, 2011; Gerlach et al., 2009; Krause et al., 2008)	201
WebMGA	Similarity	FR-HIT (Niu et al., 2011) / N/A	No / Yes	2013	Pfam, TIGRFAM, COG, KOG, PRK, GO	(Wu et al., 2011)	54
DiScRIBinATE (SOrt-ITEMS)[d]	Similarity	BLASTX, RAPSEARCH2 / N/A	Yes / No	2010 (2009)	N/A	(Ghosh et al., 2010; Monzoorul Haque et al., 2009)	48
Ray Meta	Similarity	Exact match k-mers / N/A	Yes / No	2012	N/A	(Boisvert et al., 2012)	34

Method Name	Class of Method	Sequence Alignment Method / Composition Method	Standalone[a]/ Web Server	Most Recent Year Published (First Time Published) [b]	Functional Classification if Applicable	References	Number of Citations[c]
Kraken	Similarity	Exact match k-mers / N/A	Yes / No	2014	N/A	(Wood and Salzberg, 2014)	15
RTM	Similarity	k-mers / N/A	Yes / Yes	2012	KEGG	(Edwards et al., 2012)	12
Genometa	Similarity	Bowtie (Langmead et al., 2009), BWA (Li and Durbin, 2009) / N/A	Yes / No	2012	N/A	(Davenport et al., 2012)	7
LMAT	Similarity	Exact match k-mers / N/A	Yes / No	2013	N/A	(Ames et al., 2013)	6
Sequedex	Similarity	Exact match k-mers / N/A	Yes / No	2012	N/A	(Berendzen et al., 2012)	5
MetaBin	Similarity	BLASTX, BLAT / N/A	Yes / Yes	2012	COG	(Sharma et al., 2012)	4
TAMER	Similarity	MEGABLAST / N/A	Yes / No	2012	N/A	(Jiang et al., 2012)	4
metaBEETL	Similarity	Direct comparison of compressed text indices / N/A	Yes / No	2013	N/A	(Ander et al., 2013)	2
SPANNER	Similarity	BLASTP / N/A	Yes / No	2013	N/A	(Porter and Beiko, 2013)	2
GOTTCHA	Similarity	BWA / N/A	Yes / No	2015	N/A	(Freitas et al., 2015)	0
CLARK	Similarity	k-mers / N/A	Yes / No	2015	N/A	(Ounit et al., 2015)	0

Method Name	Class of Method	Sequence Alignment Method / Composition Method	Standalone[a]/ Web Server	Most Recent Year Published (First Time Published)[b]	Functional Classification if Applicable	References	Number of Citations[c]
MLTreeMap	Marker	BLASTX / N/A	Yes / Yes	2010 (2007)	4 Enzyme families	(von Mering et al., 2007; Stark et al., 2010)	206
AMPHORA2	Marker	HMMER3 / N/A	Yes / Yes	2012 (2008)	N/A	(Kerepesi et al., 2014; Wu and Eisen, 2008; Wu and Scott, 2012)	190
MetaPhlAn	Marker	MEGABLAST, Bowtie2 (Langmead and Salzberg, 2012) / N/A	Yes / Yes	2012	N/A	(Segata et al., 2012)	114
MetaPhyler	Marker	BLASTN, BLASTX / N/A	Yes / No	2011	N/A	(Liu et al., 2011)	42
mOTU	Marker	HMMER3 / N/A	Yes / Yes	2013	N/A	(Sunagawa et al., 2013)	24
Phylosift	Marker	LAST, HMMER3 / N/A	Yes / No	2014	N/A	(Darling et al., 2014)	18
phymmBL	Hybrid	MEGABLAST / IMM	Yes / No	2011 (2009)	N/A	(Brady and Salzberg, 2011, 2009)	182
RITA	Hybrid	Pipeline of BLAST variations / NB	Yes / Yes	2012 (2011)	N/A	(MacDonald et al., 2012; Parks et al., 2011)	38
SPHINX	Hybrid	BLASTX / k-means	No / Yes	2010	N/A	(Mohammed et al., 2011a)	17

Method Name	Class of Method	Sequence Alignment Method / Composition Method	Standalone[a]/ Web Server	Most Recent Year Published (First Time Published) [b]	Functional Classification if Applicable	References	Number of Citations[c]
TaxyPro	Hybrid	CoMet web server / Mixture model	Yes / No	2013	Pfam	(Klingenberg et al., 2013)	3
TWARIT	Hybrid	BWA short read alignment (Li and Durbin, 2009) / k-means	No / Yes	2012	N/A	(Reddy et al., 2012)	2
PhyloPythiaS	Composition	N/A / SVM	Yes / Yes	2011 (2007)	N/A	(Liu et al., 2011; Patil et al., 2011, 2012)	269
TACOA	Composition	N/A / k-NN	Yes / No	2009	N/A	(Diaz et al., 2009)	65
NBC	Composition	N/A / NB	Yes / Yes	2011 (2008)	N/A	(Rosen et al., 2008, 2011)	35
RAIphy	Composition	N/A / RAI	Yes / No	2011	N/A	(Nalbantoglu et al., 2011)	18
ClaMS	Composition	N/A / DBC signature	Yes / No	2011	N/A	(Pati et al., 2011)	10
INDUS	Composition	N/A / k-means	No / Yes	2011	N/A	(Mohammed et al., 2011b)	8
TAC-ELM	Composition	N/A / Neural Network	Yes / No	2012	N/A	(Rasheed and Rangwala, 2012)	5
MetaCV	Composition	N/A / CV	Yes / No	2013	KEGG	(Liu et al., 2013)	4
GSTaxClassifier	Composition	N/A / Bayesian	No / No	2010	N/A	(Yu et al., 2010a)	2

[a]Standalone refers to whether the program can be run locally.
[b]Some methods have had several publications, with later publications regarding improvements on functionality. In these cases the most recent publication was listed, with the first time the method was published in brackets.
[c]Number of citations is based on Web of Science as of April 21, 2015.
[d]DiScRIBinATE is the successor for SOrt-ITEMS so they were included in the same row.
N/A, not applicable; IMM, interpolated Markov model; NB, naive Bayes; SVM, support vector machine; k-NN, k-Nearest Neighbour; RAI, relative abundance index; DBC signature, de Bruijn chain signature; CV, composition vector

2.4.1. Several methods vastly overestimate the number of species present

To assess performance, first the quality of the assignments made by different methods was examined with no clade exclusion, so that as many representative methods could be comparatively examined as possible. The sensitivity, precision, and taxonomic distance (Figures 2.1 and 2.2) were computed on the MetaSimHC dataset with no clade exclusion. Results were as expected, with all methods generally showing a relatively high sensitivity and precision. The exceptions are TACOA (Diaz et al., 2009), which is known to perform poorly on short reads, and MetaPhyler, which is a marker based method and thus classifies only a small proportion of the reads, resulting in low sensitivity (but high precision).

Next, the numbers of incorrectly predicted species, based on different thresholds of percentage abundance in the predicted community, were tabulated (Table 2.7). It is notable that several methods greatly overpredict the numbers of species present, considering that the sequences the methods are trying to classify exist in the reference databases or training sets.

Under genus clade exclusion conditions (Table 2.8), the number of incorrectly predicted species increases further for any method that makes incorrect predictions at the species level.

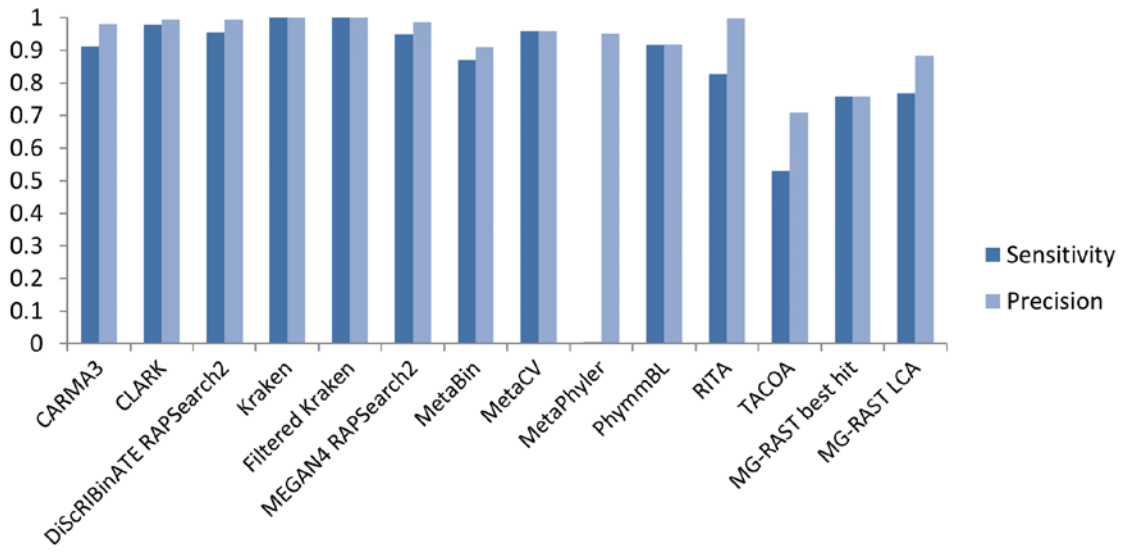

Figure 2.1 Sensitivity and precision of methods on the MetaSimHC dataset of simulated 250 bp reads with no clade exclusion.

Most methods perform very well when the species being classified are in the database or training set that the methods rely on.

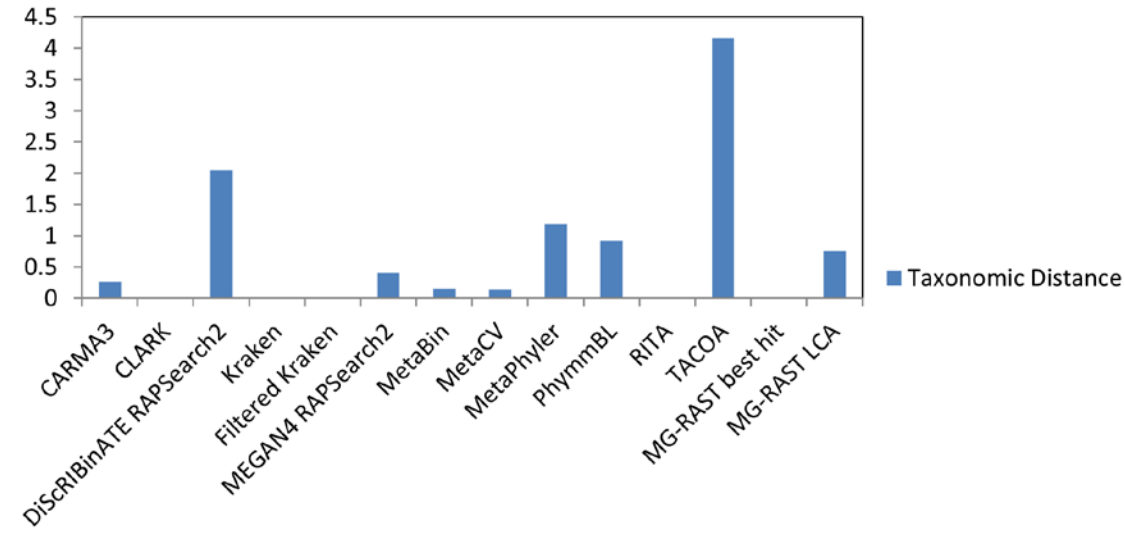

Figure 2.2 Taxonomic distance of methods on the MetaSimHC dataset of simulated 250 bp reads with no clade exclusion.

TACOA is very conservative and assigns reads at non-specific taxonomic levels, whereas DiScRIBinATE and MetaPhyler are relatively conservative.

Table 2.7 Number of correctly and incorrectly predicted species (MetaSimHC)[a] for different thresholds[b] without clade exclusion, illustrating how some methods vastly overpredict the number species, even when the true number of species is low (in this case the true number of species is 11)

	No cutoff[b]		Cutoff > 0.01%[b]		Cutoff > 0.1%[b]		Cutoff > 1%[b]	
Method	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
CARMA3	11	32	11	2	11	0	11	0
CLARK	11	32	11	9	11	2	11	0
DiScRIBinATE RAPSearch2[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Kraken	11	0	11	0	11	0	11	0
Filtered Kraken	11	0	11	0	11	0	11	0
MEGAN4 BLASTN	11	0	11	0	11	0	11	0
MEGAN4 RAPSearch2	11	63	11	19	11	1	11	0
MetaBin	11	262	11	36	11	2	11	0
MetaCV	11	1166	11	38	11	1	11	0
MetaPhyler	11	7	11	7	11	4	9	0
PhymmBL[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
RITA	11	38	11	0	11	0	10	0
TACOA[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MG-RAST best hit	11	622	11	60	11	6	11	2
MG-RAST LCA	11	125	11	7	11	1	11	0

[a]Using the MetaSimHC dataset of simulated 250 bp reads from 11 species

[b]A cutoff of > x%, for example 0.01%, would indicate that only species with a predicted abundance of at least x% of the total set of predictions were considered. Correctly predicted species are any of the 11 species that were used to simulate the reads in the dataset, whereas any other predicted species was incorrect

[c]These methods do not predict to the species level at this read length (they require longer read lengths). See additional analyses at other levels of clade exclusion

Table 2.8 Number of incorrectly predicted species[a] for different abundance thresholds[b] with genus clade exclusion

Method	No cutoff[b]	Cutoff > 0.01%[b]	Cutoff > 0.1%[b]	Cutoff > 1%[b]
CARMA3	71	11	1	1
CLARK	839	467	94	6
DiScRIBinATE RAPSearch2[c]	N/A	N/A	N/A	N/A
Kraken	860	445	95	7
Filtered Kraken	50	39	13	1
MEGAN4 BLASTN	640	493	79	6
MEGAN4 RAPSearch2	648	354	31	6
MetaBin	973	320	31	6
MetaCV	1263	1076	84	7
MetaPhyler	9	9	9	1
PhymmBL[c]	N/A	N/A	N/A	N/A
RITA	934	263	39	14
TACOA[c]	N/A	N/A	N/A	N/A
MG-RAST best hit[d]	N/A	N/A	N/A	N/A
MG-RAST LCA[d]	N/A	N/A	N/A	N/A

[a]Using the MetaSimHC dataset of simulated 250 bp reads.
[b]A cutoff of > x%, for example 0.01%, would indicate that only species with a predicted abundance of at least x% of the total set of predictions were considered. Due to genus clade exclusion, it is impossible to correctly predict any of the species, so only incorrect predictions are shown.
[c]These methods do not predict to the species level at this read length (they require longer read lengths). See additional analyses at other levels of clade exclusion.
[d]Could not perform clade-exclusion on MG-RAST

2.4.2. Sensitivity and precision vary widely between methods, with sensitivity generally decreasing at higher levels of clade exclusion and increasing with read length

The quality of the assignments made by the different methods was further examined under clade exclusion scenarios at different taxonomic levels. Sensitivity and precision were computed on the MetaSimHC dataset (Figure 2.3) and found to vary notably. To examine in greater detail what led to the differences in sensitivity and precision of these methods, the taxonomic distance for each method was evaluated (Figure 2.4). Furthermore, the proportion of reads assigned at each taxonomic rank was determined. An example of the results under the genus clade exclusion scenario is

shown in Figure 2.5, with the data for the rest in Additional File 3 (available online at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0788-5). Additionally, the numbers of reads misassigned and correctly assigned or overpredicted for each rank were compiled (genus clade exclusion Figure 2.6, the rest of the data in Additional File 4, again available online). Many of the methods assign a considerable proportion of reads to the species level, when species level assignment is impossible since the true species are excluded from the database. Also notable is that TACOA assigns the majority of reads to the superkingdom level, so the method will be of limited use for those interested in more specific taxonomic ranks, at least at these shorter read lengths.

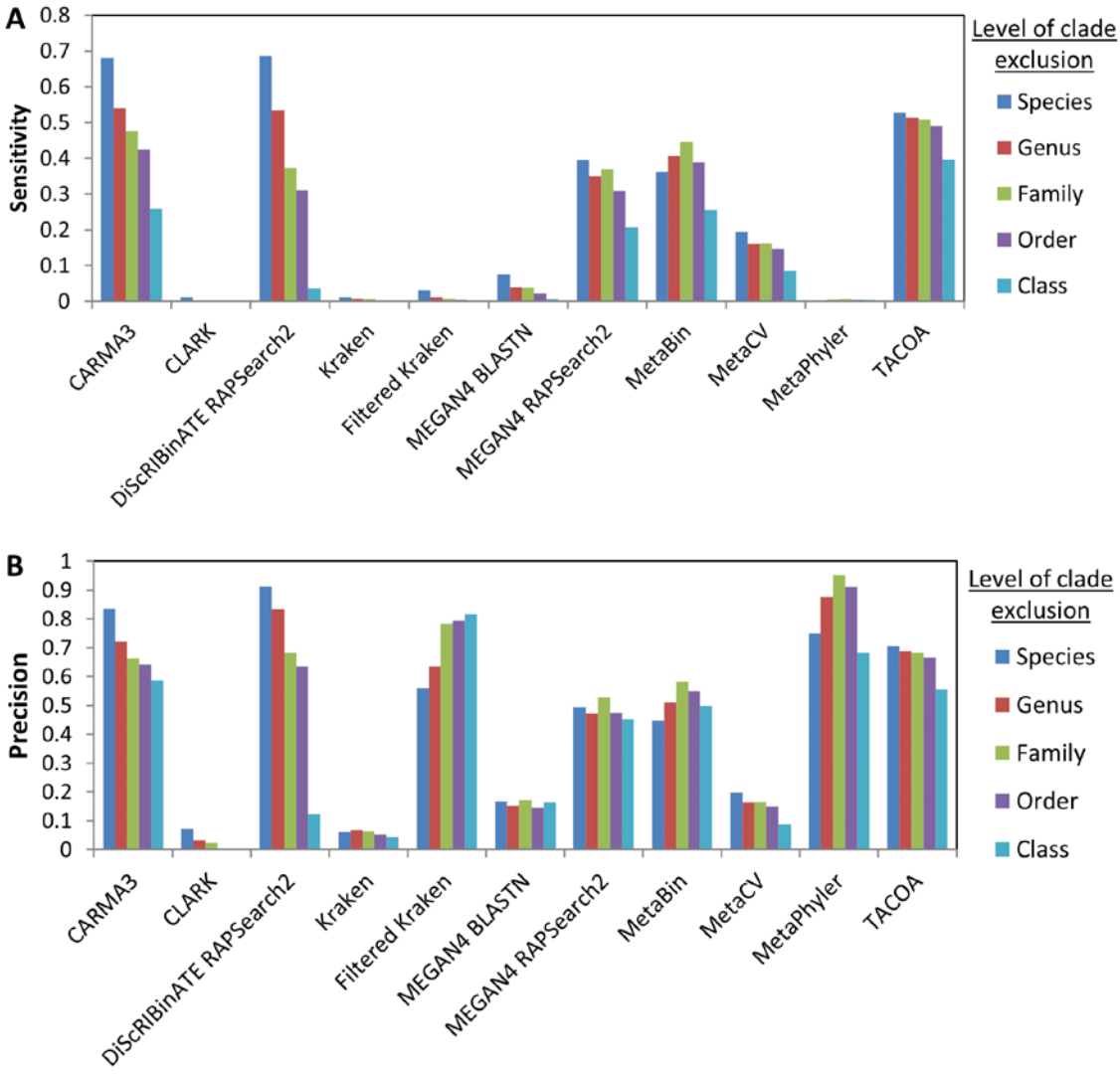

Figure 2.3 Sensitivity (A) and precision (B) on the MetaSimHC dataset of simulated 250 bp reads as clade exclusion level is varied.

There is a wide range of variability in the sensitivity and precision of the methods with sensitivity tending to decrease as the level of clade exclusion moves from species to class. Performance is calculated based on proportion of reads appropriately assigned and averaged per genome (see methods).

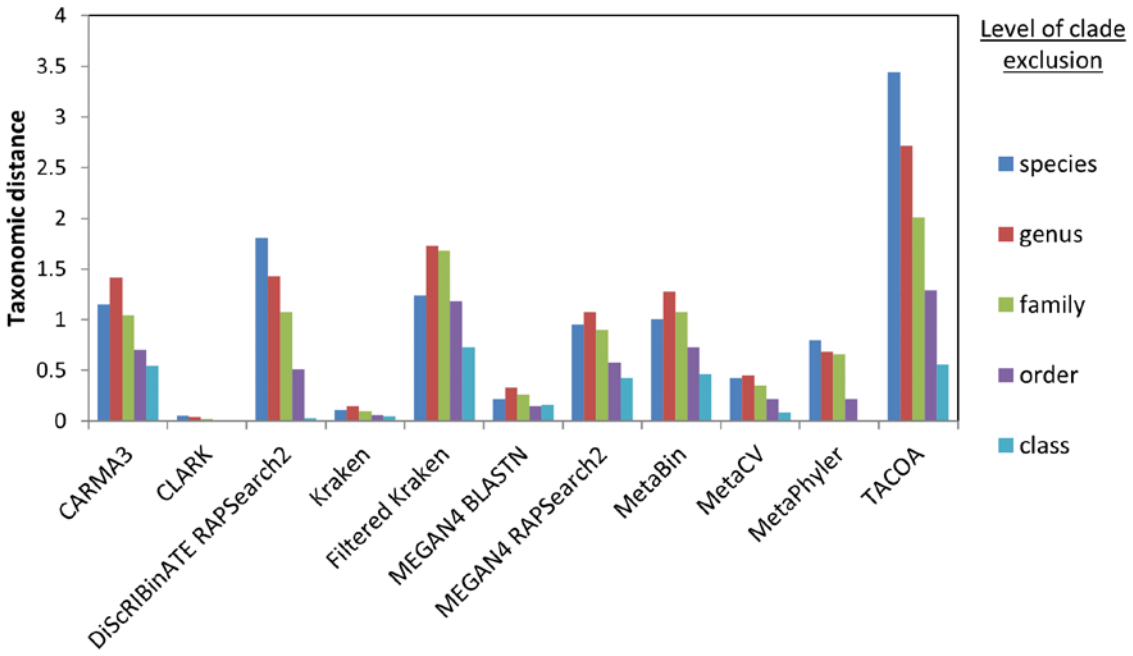

Figure 2.4 Taxonomic distance of methods on the MetaSimHC dataset of simulated 250 bp reads with various levels of clade exclusion.

Methods show a general decrease in taxonomic distance as clade exclusion level is increased, as there are fewer ranks above the best possible rank at which to classify.

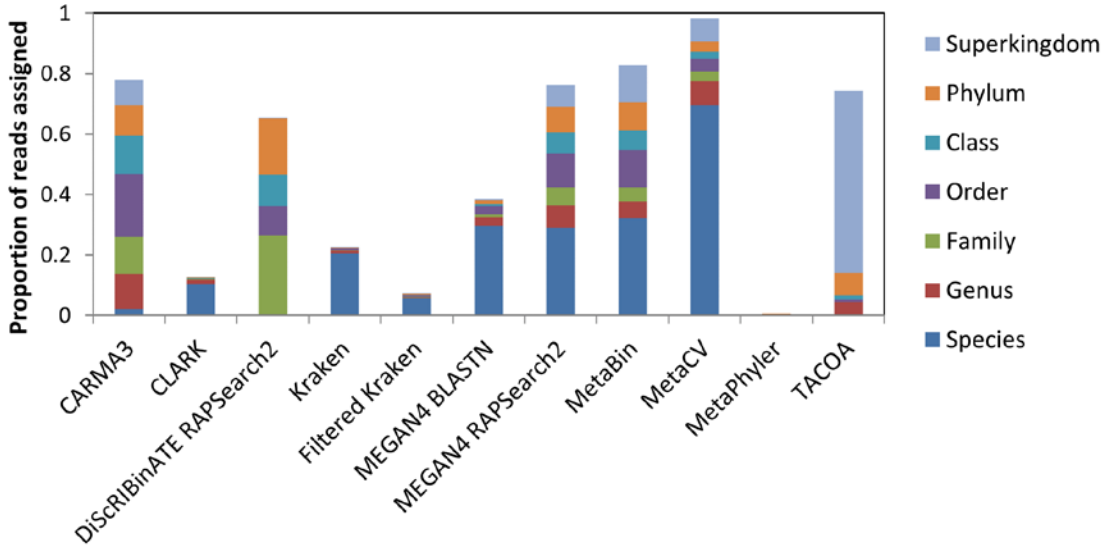

Figure 2.5 Distribution of assignments to taxonomic ranks.

Proportion of reads assigned at each taxonomic rank on the MetaSimHC dataset of simulated 250 bp reads under genus clade exclusion (includes both correct and incorrect assignments). Although the lowest possible correct rank is family, some methods still classify many reads at the species level. CARMA3 and DiScRIBinATE are slightly more conservative, classifying a large number of reads at the family or order levels, whereas TACOA is extremely conservative, classifying the majority of the reads at the superkingdom level.

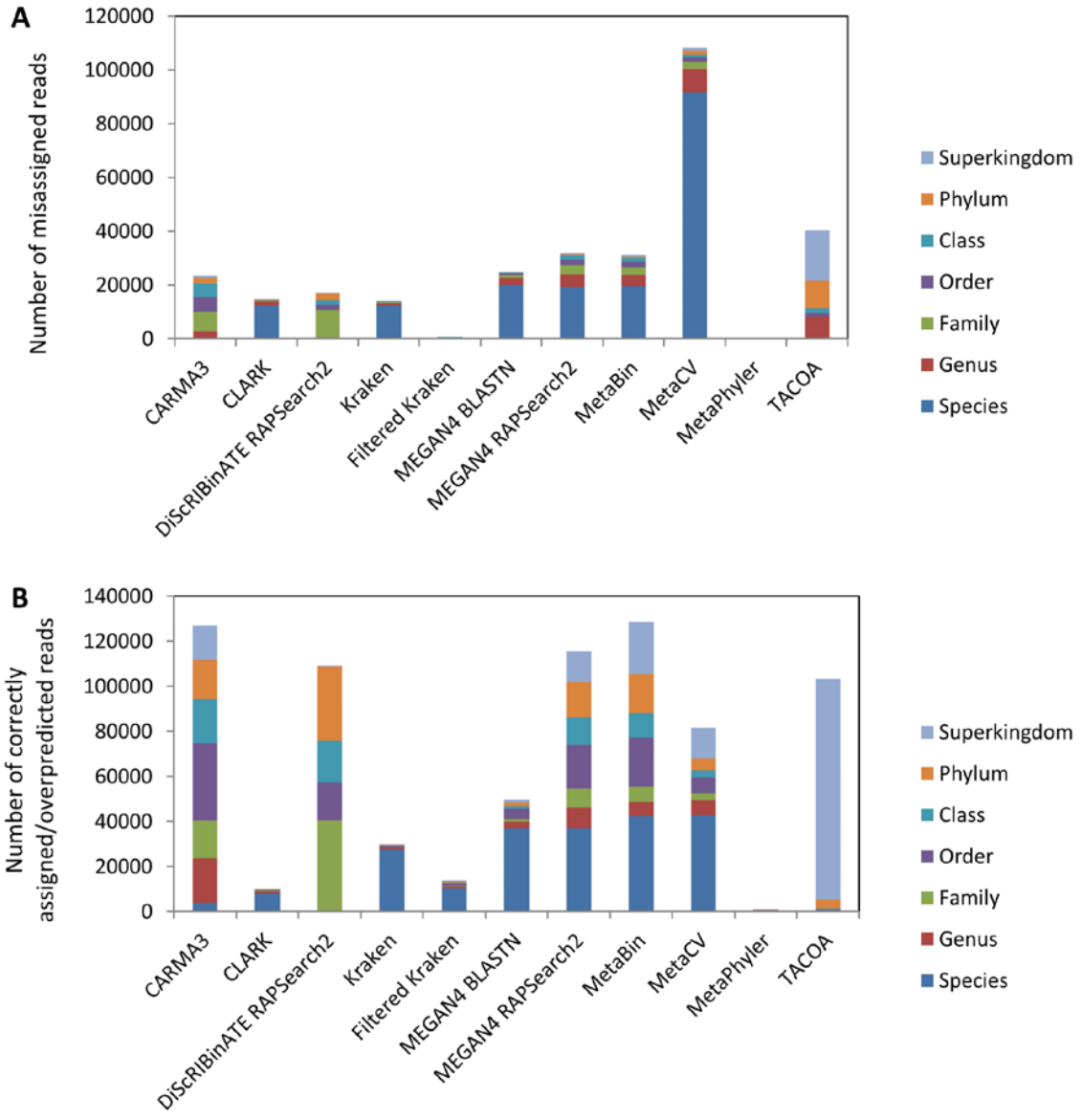

Figure 2.6 Distributions of misassigned (A) and correct or overpredicted assignments (B) to each taxonomic rank on the MetaSimHC dataset of simulated 250 bp reads under genus clade exclusion.

In some cases, overpredictions (e.g. predictions made to an incorrect species in the correct genus) are less problematic than incorrect predictions (e.g. predictions made to an incorrect genus). Thus, sensitivity and precision were recalculated after reclassifying overpredictions as correct classifications (Figure 2.7). There was notable increase in sensitivity and precision for methods such as MEGAN4 and MetaBin which are less conservative in their predictions. For more conservative methods such as CARMA3 and DiScRIBinATE, there was little change.

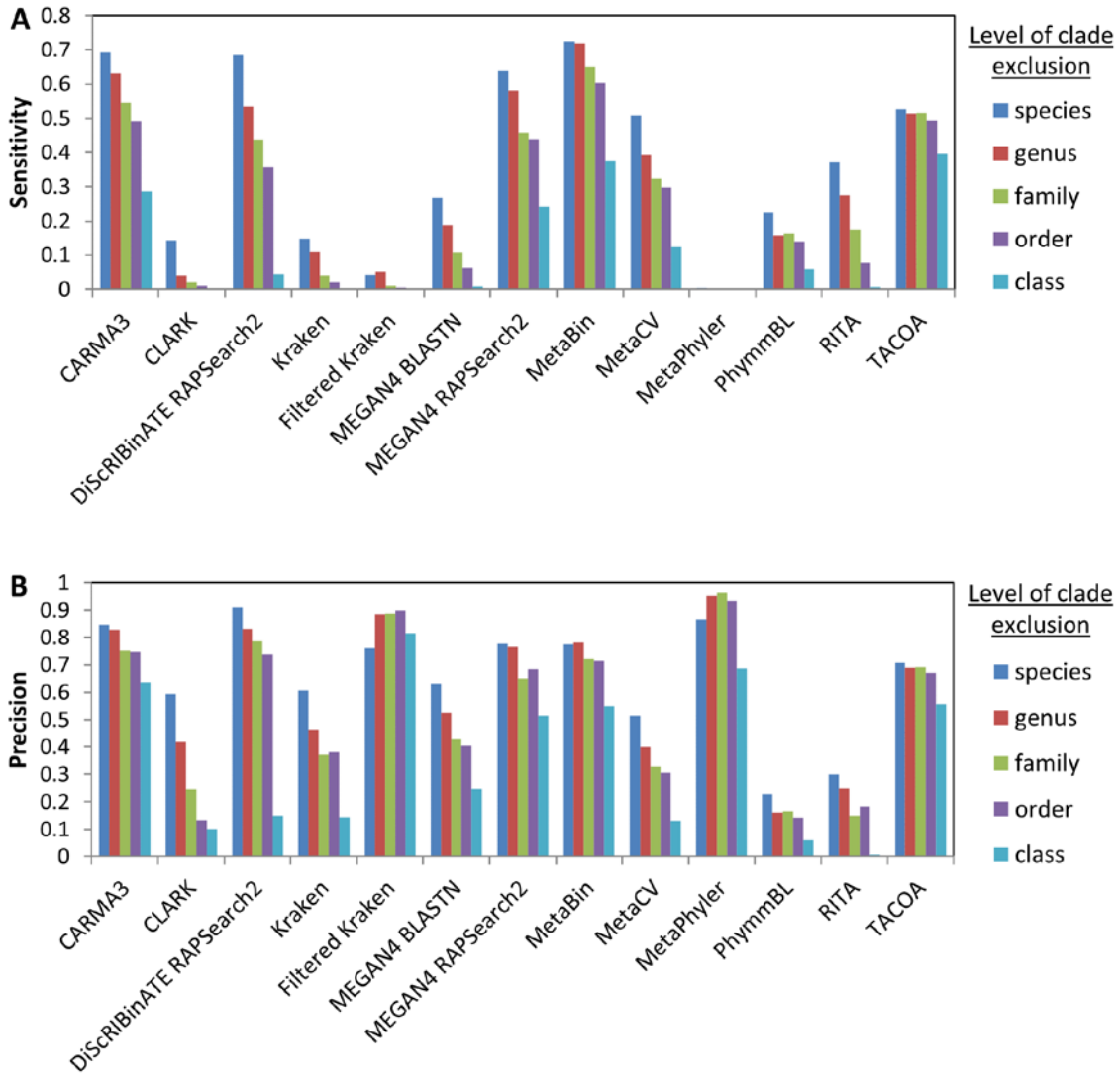

Figure 2.7 Sensitivity (A) and precision (B) on the MetaSimHC dataset of simulated 250 bp reads with overpredictions classified as correct.
See methods for the definition of overpredictions. Methods such as MEGAN4 which classify many reads at lower taxonomic levels see a considerable increase in performance, whereas more conservative methods such as CARMA3 see only a slight improvement.

The changes in sensitivity, precision, and taxonomic distance as read length increased was then examined. This was done on the MetaSimHC dataset (Figure 2.8). Sensitivity followed the expected trend of increasing along with read lengths; however, precision and taxonomic distance showed no clear trend and remained relatively unchanged.

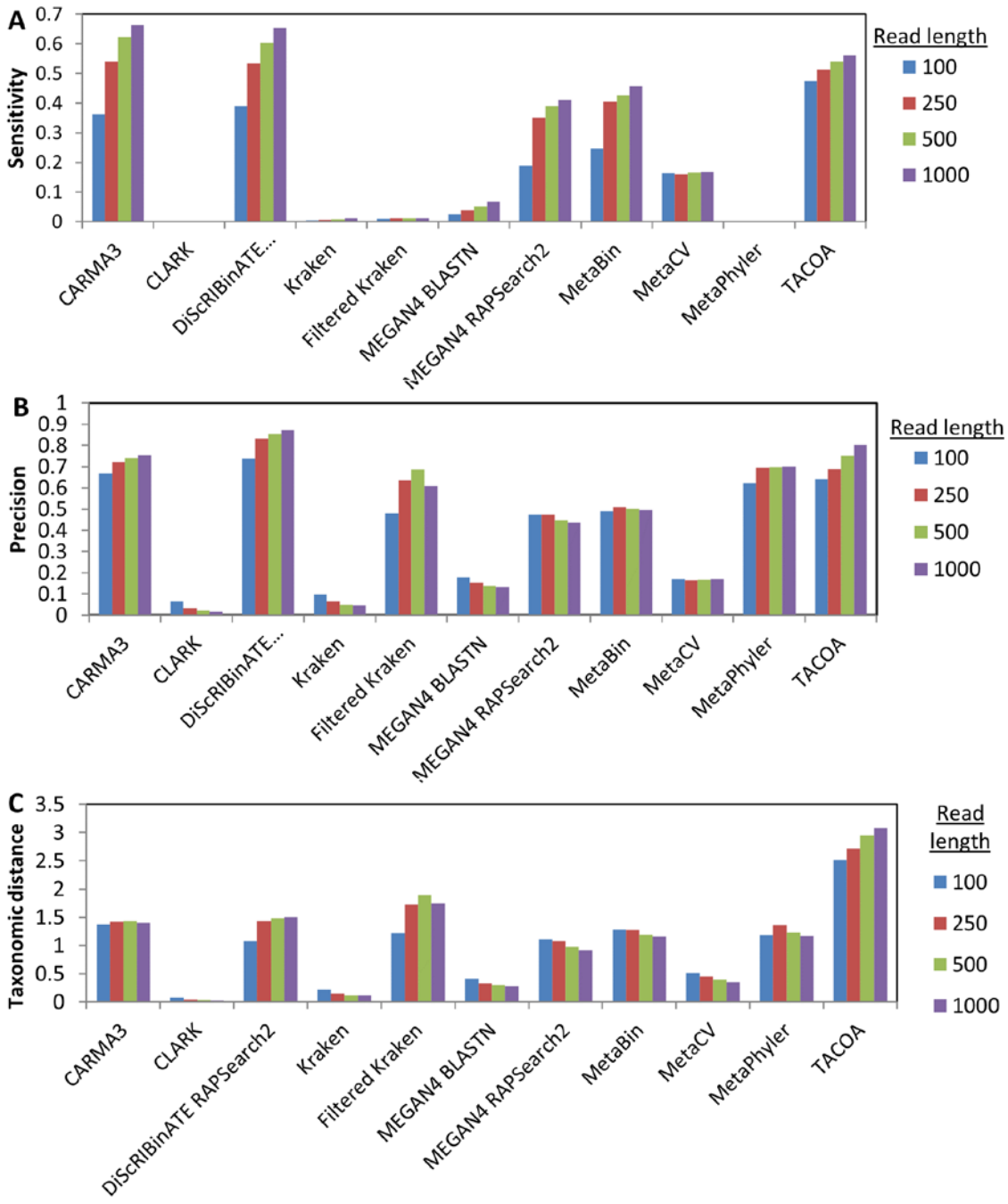

Figure 2.8 Sensitivity (A), precision (B), and taxonomic distance (C) of methods on the MetaSimHC dataset as read length is varied.

Performance of methods on the MetaSimHC dataset simulated at lengths of 100, 250, 500, and 1000 bases with genera clade exclusion. The sensitivity of the methods tends to increase with increasing read length whereas precision and taxonomic distance do not seem to change substantially

2.4.3. Analysis of the FW dataset reveals similar performance between *in vitro* data and *in silico* data, and between the FW and MetaSimHC datasets

A comparison between the FW *in silico* versus *in vitro* datasets is illustrated in Figure 2.9 under species clade exclusion, and in Figure 2.10 without clade exclusion. For the *in vitro* dataset, as it is not possible to determine which read absolutely should be associated with which organism in the mock microbial community, a hit to any of the taxa in the FW dataset was considered correct. In addition, this meant the sensitivity, precision, and taxonomic distance were based on all of the reads classified rather than averaged over all taxa.

The results are similar between the *in vitro* and *in silico* communities, suggesting that for this simple community the methods evaluated are relatively robust to Illumina sequencing errors with the sequencing technology used. A comparison of results between MetaSimHC and FW *in silico* revealed that the relative performance of methods remained similar when analyzing these two different datasets (Figure 2.11). Additionally, the numbers of incorrectly predicted species, based on different thresholds of percentage abundance in the predicted community, were again tabulated for the *in vitro* data (Table 2.9).

Many of the methods incorrectly predict hundreds of species, with MetaCV incorrectly predicting 1226 species, although after filtering out low abundance predictions the numbers of incorrect predictions were drastically reduced. Under genus clade exclusion conditions (Table 2.10), the number of incorrectly predicted species increases further, and even after filtering out low abundance predictions there were sometimes considerable numbers of false species predictions. The number of incorrectly predicted species is higher for the *in vitro* data relative to the *in silico* data (Table 2.11). The greater number of incorrectly predicted species is particularly notable in some methods that perform very well on the *in silico* data, such as MEGAN4 BLASTN, which goes from 0 incorrectly predicted species to 110. The performance for each of the component genomes on all *in silico* datasets is provided in Additional File 5 (available online, see https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0788-5).

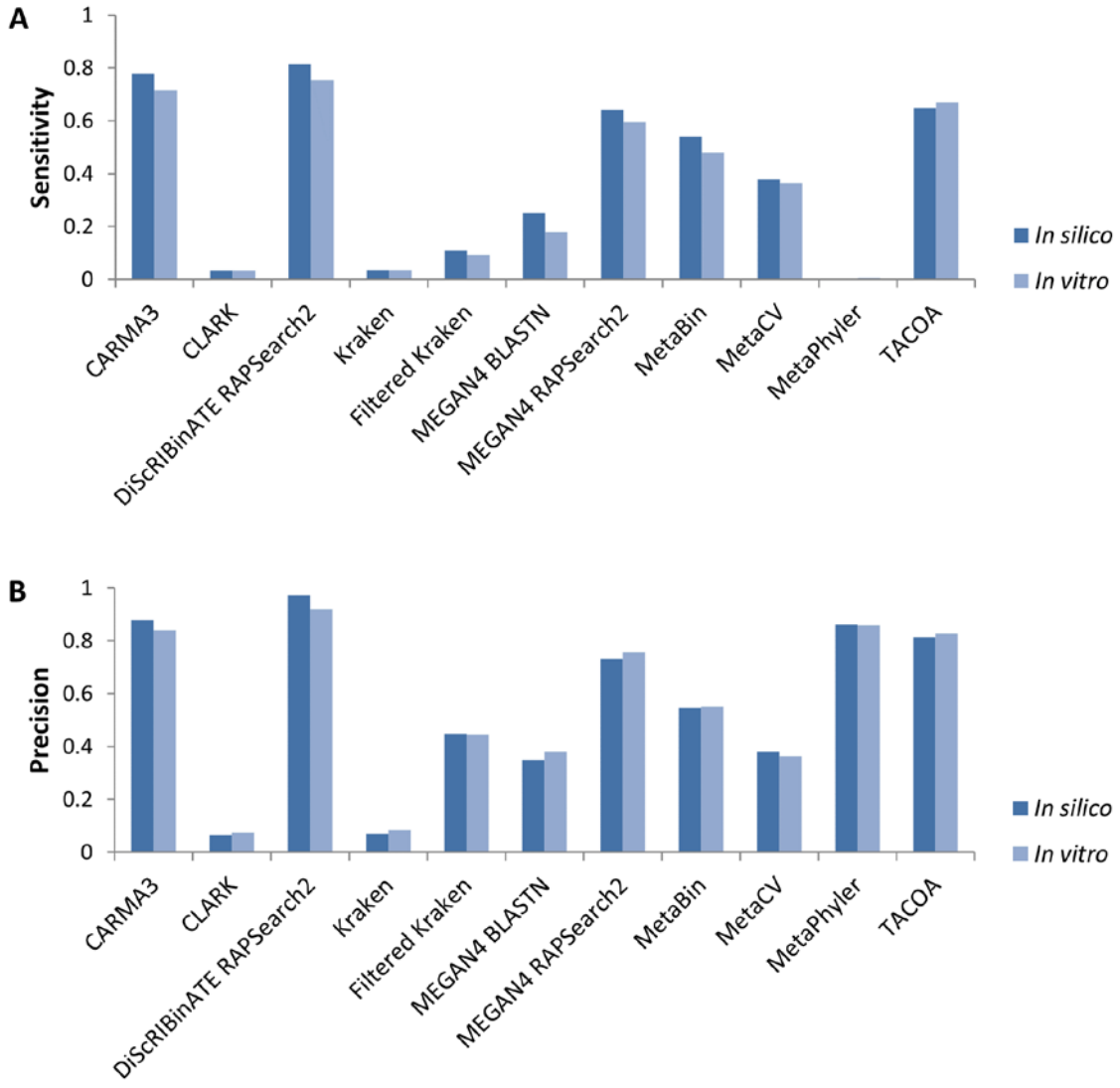

Figure 2.9 Sensitivity (A) and precision (B) of methods on the FW dataset comparing the performance on the *in silico* community versus the *in vitro* community under species clade exclusion.

The results are similar between the *in vitro* and *in silico* communities, demonstrating that methods appear to be relatively robust to real Illumina sequencing errors for this simple community.

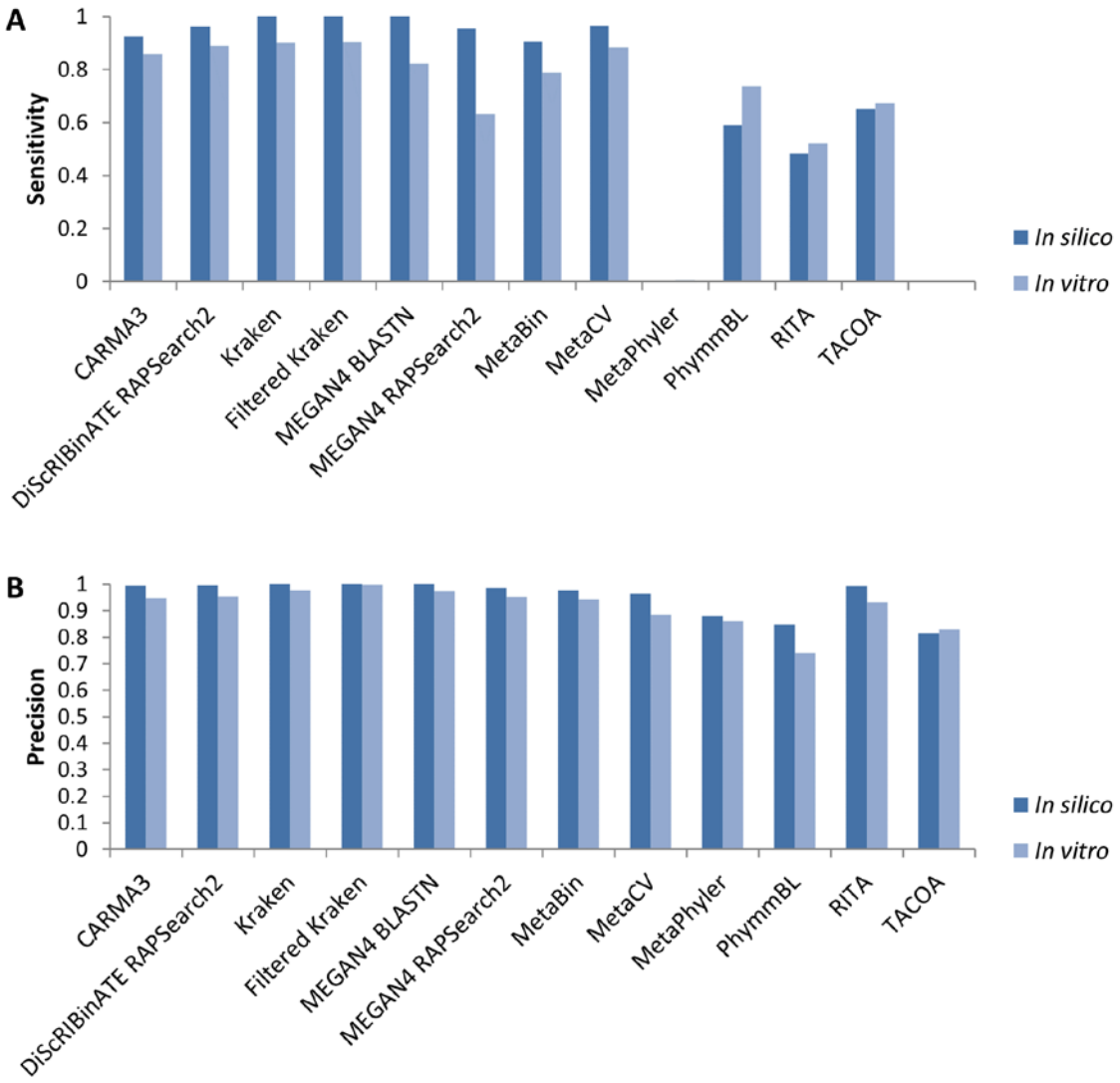

Figure 2.10 Performance of FW *in silico* versus FW *in vitro* without clade exclusion.

Sensitivity **(A)** and precision **(B)** of methods on the FW dataset comparing the performance on the *in silico* community versus the *in vitro* community.

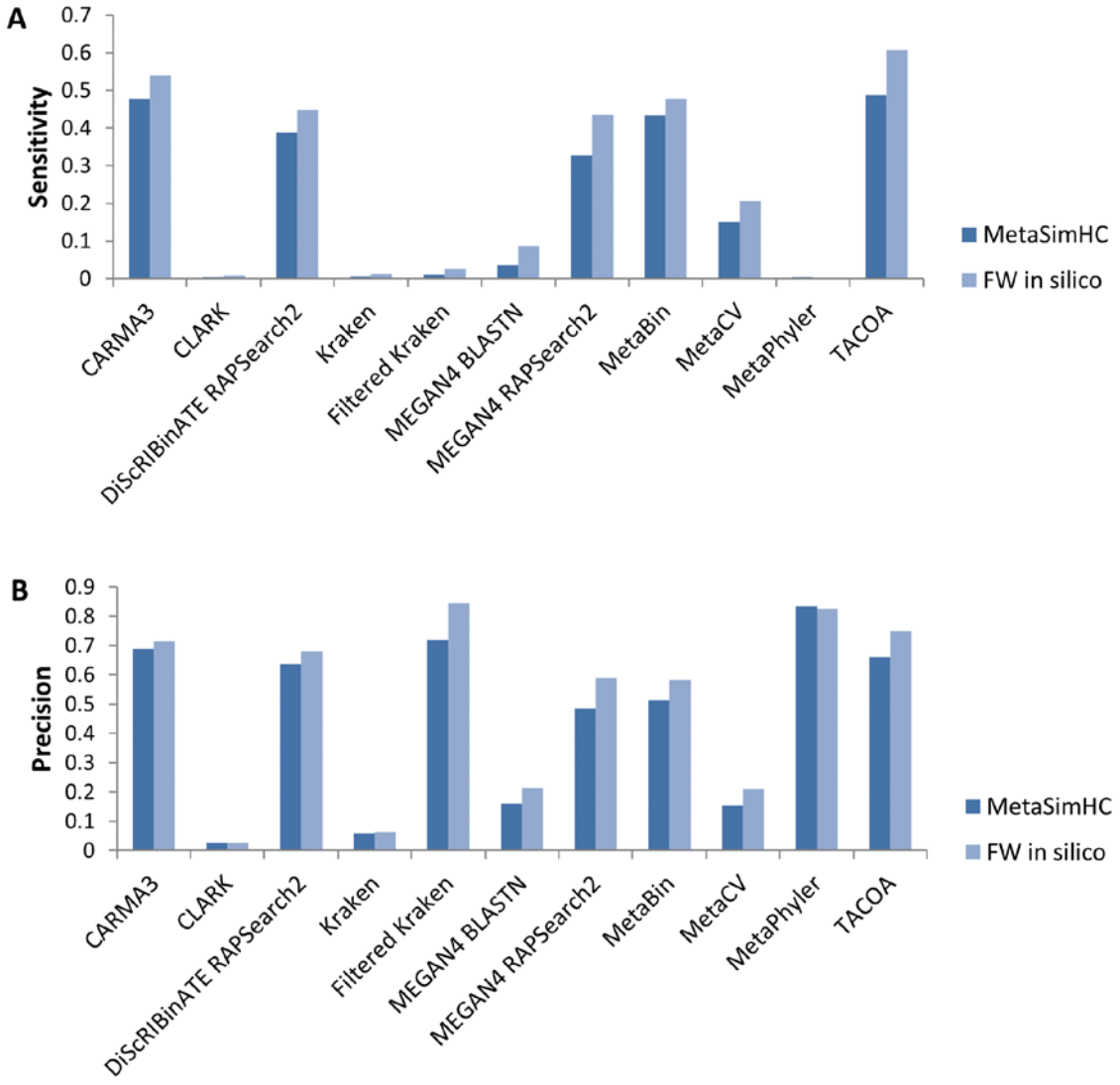

Figure 2.11 Sensitivity (A) and precision (B) of methods on the MetaSimHC dataset compared to the FW *in silico* of simulated 250 bp reads.

Values are averaged over all levels of clade exclusion from species to class. Although the microbes in the dataset changed, the relative performance of the methods remains very similar.

Table 2.9 Number of correctly and incorrectly predicted species (FW *in vitro*)[a] for different thresholds[b] without clade exclusion

	No cutoff[b]		Cutoff > 0.01%[b]		Cutoff > 0.1%[b]		Cutoff > 1%[b]	
Method	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
CARMA3	11	56	11	4	11	0	10	0
CLARK	11	364	11	25	11	5	11	0
DiScRIBinATE RAPSearch2[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Kraken	11	327	11	25	11	5	11	0
Filtered Kraken	11	14	11	1	11	0	11	0
MEGAN4 BLASTN	11	110	11	19	11	3	9	1
MEGAN4 RAPSearch2	11	183	11	41	11	1	9	1
MetaBin	11	561	10	77	10	6	10	1
MetaCV	11	1226	11	232	11	6	10	1
MetaPhyler	11	9	11	9	11	5	7	1
PhymmBL[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
RITA	11	466	10	80	10	10	10	1
TACOA[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MG-RAST best hit	11	927	10	180	10	36	10	8
MG-RAST LCA	11	476	11	69	11	5	11	1

Some methods vastly overpredict the number species, even when the true number of species is low (in this case the true number of species is 11)

[a]Using the FW *in vitro* dataset of sequenced reads from 11 species.

[b]A cutoff of > x%, for example 0.01%, would indicate that only species with a predicted abundance of at least x% of the total set of predictions were considered. Correctly predicted species are any of the 11 species that were used to simulate the reads in the dataset, whereas any other predicted species was incorrect.

[c]These methods do not predict to the species level at this read length (they require longer read lengths). See additional analyses at other levels of clade exclusion.

Table 2.10 Number of incorrectly predicted species[a] for different abundance thresholds[b] with genus clade exclusion. Even more incorrectly predicted species are predicted under these conditions versus without clade exclusion

Method	No cutoff[b]	Cutoff > 0.01%[b]	Cutoff > 0.1%[b]	Cutoff > 1%[b]
CARMA3	102	9	4	0
DiScRIBinATE RAPSearch2[c]	N/A	N/A	N/A	N/A
Kraken	741	422	145	10
Filtered Kraken	87	39	10	5
MEGAN4 BLASTN	447	231	25	2
MEGAN4 RAPSearch2	517	273	32	3
MetaBin	905	316	36	3
MetaCV	1253	901	144	3
MetaPhyler	6	6	4	1
PhymmBL[c]	N/A	N/A	N/A	N/A
RITA	865	502	182	16
TACOA[c]	N/A	N/A	N/A	N/A
MG-RAST best hit[d]	N/A	N/A	N/A	N/A
MG-RAST LCA[d]	N/A	N/A	N/A	N/A

[a]Using the FW *in vitro* dataset of sequenced reads from 11 species.
[b]A cutoff of > x%, for example 0.01%, would indicate that only species with a predicted abundance of at least x% of the total set of predictions were considered. Due to genus clade exclusion, it is impossible to correctly predict any of the species, so only incorrect predictions are shown.
[c]These methods do not predict to the species level at this read length (they require longer read lengths). See additional analyses at other levels of clade exclusion.
[d]Could not perform clade exclusion on MG-RAST

Table 2.11 Number of incorrectly predicted species[a] for different abundance thresholds[b] without clade exclusion. Fewer incorrectly predicted species are predicted with the *in silico* data that does not contain errors versus the *in vitro* data containing sequencing errors (Table 2.9)

	No cutoff[b]		Cutoff > 0.01%[b]		Cutoff > 0.1%[b]		Cutoff > 1%[b]	
Method	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
CARMA3	11	56	11	4	11	0	10	0
CLARK	11	364	11	25	11	5	11	0
DiScRIBinATE RAPSearch2[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Kraken	11	327	11	25	11	5	11	0
Filtered Kraken	11	14	11	1	11	0	11	0
MEGAN4 BLASTN	11	110	11	19	11	3	9	1
MEGAN4 RAPSearch2	11	183	11	41	11	1	9	1
MetaBin	11	561	10	77	10	6	10	1
MetaCV	11	1226	11	232	11	6	10	1
MetaPhyler	11	9	11	9	11	5	7	1
PhymmBL[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
RITA	11	466	10	80	10	10	10	1
TACOA[c]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
MG-RAST best hit	11	927	10	180	10	36	10	8
MG-RAST LCA	11	476	11	69	11	5	11	1

[a]Using the FW *in vitro* dataset of sequenced reads from 11 species.
[b]A cutoff of > x%, for example 0.01%, would indicate that only species with a predicted abundance of at least x% of the total set of predictions were considered. Correctly predicted species are any of the 11 species that were used to simulate the reads in the dataset, whereas any other predicted species was incorrect.
[c]These methods do not predict to the species level at this read length (they require longer read lengths). See additional analyses at other levels of clade exclusion.

2.4.4. There is substantial variation in the computational cost of different methods

To evaluate how long the various methods took to run, 22,000 reads of 100 bp, 250 bp, 500 bp, and 1000 bp, and an additional 44,000 reads of 250 bp were simulated using the MetaSimHC dataset. The time taken by the different methods to complete an analysis of these sequences varied widely, and nearly all methods scaled roughly

linearly with both read length and number of reads on our datasets (Figure 2.12). Sequence similarity based methods that rely on BLASTX take considerably longer than all other methods except TACOA, taking over 24 hours for just 22,000 reads of 250 bp under the CPU conditions in the test (one Intel Xeon E5-2660 2.2 GHz CPU and 282 GB of RAM). At the other extreme, Kraken and CLARK took less than one minute to classify all of the reads.

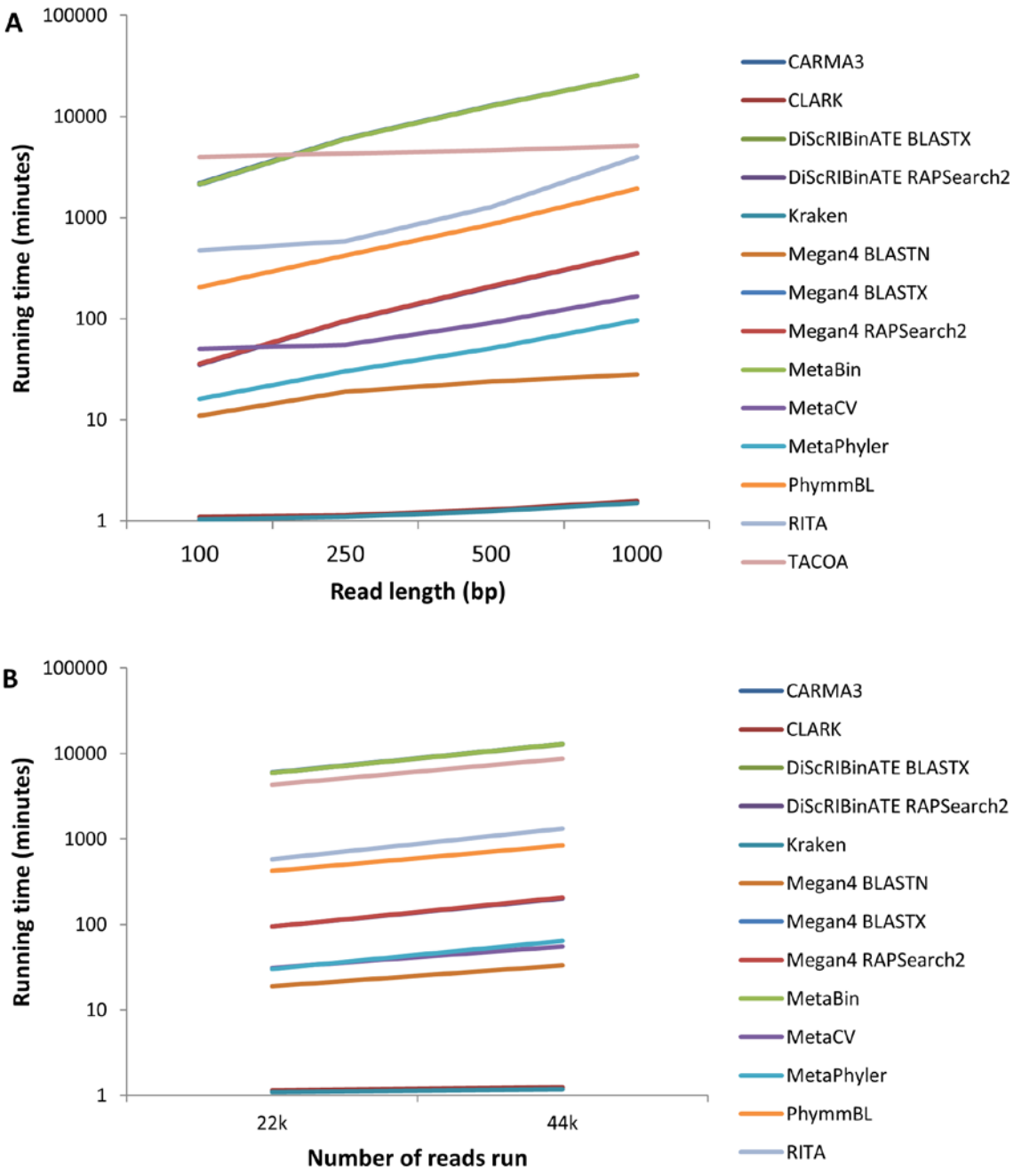

Figure 2.12 Comparison of running time.
Running time for the various methods was calculated on a MetaSimHC dataset of 22,000 simulated reads of various read lengths **(A)**, or 22,000 and 44,000 reads of 250 bp **(B)**. The similarity search programs take up the majority of time to run for the methods that rely on them, so these methods overlap on the graphs (e.g. MEGAN4 and DiScRIBinATE BLASTX, CARMA3, and MetaBin all use BLASTX).

2.5. Discussion

All of the methods analyzed performed very well in terms of sensitivity and precision when the query sequences were in the reference databases (i.e. when there was no clade exclusion). Of course, this type of analysis would be expected to give potentially artificially high accuracy values since one is essentially evaluating using test data identical to the reference/training data. Under this type of analysis scenario, the more informative metrics to examine are taxonomic distance and the number of incorrectly predicted species. Notably, several methods substantially overpredicted the number of species present in the simulated communities. This included popular methods such as MG-RAST and MEGAN4. However, most of these incorrectly predicted species are predicted at a very low abundance. By setting a threshold to filter out low abundance predictions, the number of incorrect predictions can be considerably reduced. The thresholds presented here are not intended as suggestions, but rather to demonstrate the principle of using thresholds to filter out incorrect predictions.

Microbial communities in certain environments are very complex, such as those found in soil (Fierer et al., 2012). These environments, which are very diverse and contain a large number of organisms, would have a large proportion of the microbes found at less than 1% of the total abundance of the community, and thus a 1% filtering threshold would filter out many of the microbes actually in the metagenome. If thresholds are used, they should ideally be chosen based on a mock community control that reflects the anticipated level of diversity and complexity expected in the metagenomics analysis being performed. If the goal is to choose thresholds based on relative abundance, genome size of the organisms would also be useful to take into account. Otherwise, if two organisms are present in the community at low levels but one organism's genome is much bigger, the organism with the smaller genome may get filtered out while the organism with the larger genome does not, due to greater number of reads from the larger genome.

It is important for researchers doing metagenomics projects to know the level of precision of the method that they are using, to have an idea of how well they can trust the taxa predicted at lower abundance. There is a trade-off between finding all of the taxa that exist in the sample, and confidence in the prediction of the taxa. Two ways to adjust this trade-off are to choose a more precise (conservative) method, or to alter the

minimum abundance threshold, with only the taxa over this abundance threshold being reported. Some methods already have a way of choosing this threshold. For example, MEGAN4 by default requires at least 5 reads to hit a taxon before the taxon is reported. The reads that are initially assigned to a taxon with fewer than the chosen threshold number of reads are then pushed up the taxonomy until they reach a taxon with a number of reads assigned to it that is over the threshold. However, when many reads are analyzed, overprediction will still occur and we have found for our analyses that it is necessary to use an additional threshold for removal of low abundance reads that are likely false predictions for such methods. Ideally this threshold may be chosen in part from an analysis of an *in vitro* mock community sample – an important experimental control in any metagenomics analysis. Such evaluation of methods using real sequence data also acts as an additional important control regarding other aspects of a metagenomics sequencing pipelines.

As demonstrated in Figure 2.3, the sensitivity and precision of methods vary dramatically. Methods show a general trend of decreasing sensitivity as the rank of clade exclusion increases. This is expected as the sequences left in the database will become increasingly divergent, and the scores of the matches, if any, will decrease. There is a notable decrease in performance for methods relying on sequence composition or nucleotide-based BLASTN similarity searches, versus the protein/amino acid sequence-based BLASTX and RAPSearch2 similarity based methods. This confirms what has been reported previously, that sequence composition based methods have lower performance than sequence similarity based methods at shorter read lengths (Brady and Salzberg, 2009). BLASTN is outperformed by amino acid-based similarity approaches under clade exclusion probably because nucleotide sequence search is well known to be less sensitive for more divergent sequences due to its lower number of different characters (4 bases versus the 20 amino acids).

The differences in performance between methods can be partially explained by the distribution of taxonomic ranks that they assign reads to. As seen in Figure 2.4, CARMA3 and DiScRIBinATE are assigning reads more conservatively; that is, they are assigning many fewer reads to the lower taxonomic ranks. Many of these lower level predictions of other methods are in fact overpredictions, as demonstrated by their large increases in sensitivity and precision between Figure 2.3 and Figure 2.7. Due to the way we evaluated methods, the most conservative methods will show the highest sensitivity

and precision, but may not be making classifications at specific enough taxonomic ranks to be useful. TACOA, for example, shows high sensitivity and precision, yet makes classifications at very high taxonomic ranks that would not be useful for most researchers.

Not surprisingly, the sensitivity increases for methods as read length increases. The most dramatic increase appears to be between read lengths of 100 and 250 bp. Thus, when choosing a sequencing technology, it may be important to try and obtain a sequence read length of at least around 250 bp. The precision and the taxonomic distance of methods remained relatively unchanged. This was likely due to any increased performance in precision and taxonomic distance offset by additionally classified reads (as seen by the increase in sensitivity) with greater dissimilarity to sequences in the databases of methods, which would have poorer performance in terms of precision and taxonomic distance.

Our comparison of the *in silico* to the *in vitro* freshwater community showed similar results in terms of relative performance of the methods. This gives us some confidence in our results of the other *in silico* simulations, as well as demonstrating the robustness of the evaluated methods to real sequence errors for this simple community. However, this would not necessarily generalize to more diverse communities, or other sequencing technologies. The sensitivity and precision of the methods followed the trends seen in the MetaSimHC *in silico* evaluation, although filtered Kraken showed somewhat lower relative precision. Upon further analysis, this appeared to be due to the nature of the way precision was calculated in this comparison. For the comparison to be done fairly between the *in silico* and *in vitro* community, the metrics were based on all reads rather than the average for all organisms. Filtered Kraken seemed to stand out in that for most organisms it classified few of the reads, and the ones it classified were mostly correct. However, for two organisms (*E. coli* and *B. cereus*), the majority of the reads were classified incorrectly. This means that because more of the reads of *E. coli* and *B. cereus* were classified than the other organisms, their (mostly inaccurate) classifications had a relatively large influence on the precision. The numbers of genomes/taxa in the mock communities was small, relative to the anticipated number of species in most real metagenomic analyses, so abnormal results from individual genomes could have a large impact on the results, as seen here with filtered Kraken. It is also notable that *E. coli* and *B. cereus*, mainly due to historical reasons, come from

regions of the taxonomic tree that are not reflective of the typical case for many environments; genomes with high sequence similarity and composition in this part of the tree are classified as the same species, whereas if they were found in other parts of the tree they would be classified as different species or genera (Fukushima et al., 2002; Økstad and Kolstø, 2011). Thus, species that are not yet discovered will not be classified in a similar manner to the genomes in *Escherichia* or *Bacillus*, and so the performance of methods on these genomes likely does not reflect performance on as yet undiscovered microbes in metagenomics samples. However, it must be emphasized that there is no one mock community dataset that can best evaluate all metagenomics software. It is key for researchers to design mock communities for evaluation that are suitable for their experiment, and use this published analysis to appreciate the types of issues they should watch out for.

The differences we saw in computational costs of the different methods were substantial. Although we ran only a few small test datasets of thousands of reads, we were able to clearly show very large differences in computational cost of the methods. Current metagenomics datasets often include millions of reads; without access to large amounts of computing power, many researchers will not find it practical to utilize BLASTX based methods for Illumina sequence sized data sets as are currently produced. The need for a more rapid alternative is already being addressed by such methods as RAPSearch2 (Zhao et al., 2012), LAST (Frith et al., 2010), PAUDA (Huson and Xie, 2014), and DIAMOND (Buchfink et al., 2015). Notably, RAPSearch2 shows similar, or in some cases even increased, performance relative to the same methods using BLASTX, while requiring much less time to run (over 30x faster in our analyses). Many methods provide the option of running multiple threads, so access to additional processors will allow the methods to run substantially quicker. Furthermore, for most methods reads are classified independently from one another, so files of reads can be broken up into multiple smaller files and each file run on a separate processor, and the results of the classifications combined.

In addition to computational cost, the amount of RAM used by different methods varies considerably. Both Kraken and CLARK require large amounts of RAM, but do provide reduced standard databases for users with low-memory computing environments (known as MiniKraken and Clark-*l*). Certain methods also allow users to adjust settings to allow trade-offs between speed, accuracy and RAM usage, such as

the sampling factor value in CLARK. A final consideration of computational resources when choosing a method is the amount of disk space that a method requires. The databases used by some methods require relatively large amounts of disk space, such as the standard database of Kraken which requires at least 160 GB of disk space. Another aspect that may affect method choice is the relative ease of generating new databases for the methods. Certain methods rely on the results of a similarity search, and expanding the database is a relatively simple process of generating a new database for that similarity search, such as BLAST. However, other methods may require substantial computational resources that researchers may not have access to. For example, the authors of GOTTCHA state that the creation of a database from the 2500 prokaryotic genome projects available in 2012 required 2TB of RAM. Other methods, such as many online only methods, do not even allow the modification/expansion of the database.

Protein sequence similarity-based methods (e.g. BLASTX, RAPSearch2) perform very well in clade exclusion scenarios but do not perform as well as nucleotide based methods when there is no clade exclusion. This is likely because a proportion of microbial genome sequences (commonly around 6-14% (Rogozin et al., 2002)) is non-coding. Protein similarity-based methods still have a relatively high sensitivity, generally >0.94 and, as noted in (Gerlach and Stoye, 2011), this is due to many reads overlapping at least partially with a coding region. This explanation makes sense with our finding that as read length is increased, sensitivity of the aforementioned methods increases (from 0.94 at read lengths of 100 to 0.99 at read lengths of 1000 nucleotides for MEGAN4 BLASTX on the MetaSimHC dataset), as it would be less likely that a longer read would cover only non-coding regions. A quick examination of these incorrectly classified reads confirmed that they were the non-coding regions of the genomes, in many cases rRNA genes.

The results presented should guide researchers to the choice of method that best fits their research question and computational resources. Clearly, certain methods perform well in certain situations. Kraken, Filtered Kraken, and MEGAN4 BLASTN perform exceedingly well when there is no clade exclusion, yet their sensitivity is low when there is clade exclusion. However, filtered Kraken classifies only a small percentage of reads when the species present in the dataset is not in the database. For example, filtered Kraken classifies fewer than 8% of the reads under genera exclusion

(Figure 2). A strategy researchers may therefore use is to take their dataset and first run it on filtered Kraken, followed by running the reads not classified by filtered Kraken on a more conservative method such as DiScRIBinATE RAPSearch2. Filtered Kraken would classify the reads from genomes in the reference database, while leaving the majority of reads from genomes not in the reference database unclassified. Then, DiScRIBinATE RAPSearch2, which will assign a much greater proportion of reads from genomes not in reference databases, could be run on the unclassified reads. If a conservative method such as DiScRIBinATE RAPSearch2 is run alone, it may miss many of the assignments of known genomes to the species rank, due to its tendency to make assignments at higher ranks. However, in some cases, such as when analyzing less well characterized microbiomes (such as in water versus human feces) the use of such conservative methods could be entirely appropriate. The pipeline idea of combining methods is integrated into some methods like RITA, which first identifies a highest-confidence set of predictions, then subjects the sequences not yet classified to a series of downstream classification steps. CARMA3 performs well in both the no-clade exclusion scenario (with a small taxonomic distance, classifying many reads to the species level) as well as the clade exclusion scenario. However, CARMA3 takes a considerable time to run, and may not be computationally feasible for those with large datasets and without access to notable compute power. Another technique involving combining methods would be to use multiple methods and look for consistent assignments among methods (Garcia-Etxebarria et al., 2014). Depending on the type of analysis, this could increase precision and confidence in the assignments, although at the cost of sensitivity in most cases and run time (due to running multiple methods).

The test datasets used in this evaluation are limited in their complexity and diversity, as well as the number of reads simulated. For example, often millions of reads are sequenced for metagenomics samples, while our datasets were smaller, containing tens to hundreds of thousands of reads. Furthermore, many environments sampled are far more complex and diverse, containing a much larger number of microbes with varying relative abundance, such as soil or the human gut. Our analyses were also either on *in silico* simulated communities or communities sequenced with a single sequencing technology. The aim of this research was not to recommend any specific method, but to raise awareness of the advantages and disadvantages of different methods and issues in metagenome analyses.

This evaluation highlights that there are large differences in methods on even the relatively simple communities used for our datasets, such as number of organisms predicted, sensitivity and precision, how specific the classifications tend to be (taxonomic rank), and computational resources required to run. However, other factors such as the diversity and microbes present in a community, and the sequencing technology used, will also affect the performance of the methods. Additionally, certain tools may have advantages and be particularly useful for specific environments. For example, some tools contain genomes in their databases that are not present in RefSeq, while most methods use RefSeq exclusively for their databases. An example of this is MetaPhlAn, which includes many draft genomes from the larger Human Microbiome Project (HMP) (Turnbaugh et al., 2007), and thus may be particularly useful for human microbiome samples.

Metagenomics as a field is expanding rapidly. New tools are needed to classify the sequences obtained from these studies. There is a large need, and lots of interest in this, as evidenced by the large number of methods released over the past few years. However, it is non-trivial to perform an evaluation of methods. This is due to the sheer number of metagenomic methods available, the difficulty in setting up some of these methods, and the challenge in performing robust evaluation techniques such as clade exclusion or leave-one-out evaluation. Furthermore, methods available only on the web are generally unable to be thoroughly evaluated as in many cases they do not allow the use of custom reference databases or training sets, and sometimes limit the number of reads that can be uploaded. To address these difficulties, an initiative called the Critical Assessment of Metagenomic Interpretation (CAMI) has been initiated. This community-led initiative had researchers run their own methods on data sets made up of unpublished microbial genomes. This is a valuable contribution to methodology assessment, but researchers are still encouraged to use mock microbial communities as controls for their own particular analyses, especially mock communities that reflect the types of microbes, diversity, and complexity they expect to see in their study. Researchers should always perform a metagenomics analysis using appropriate controls to best refine methodology and any threshold cutoffs suitable for the specific analysis needs.

Since this evaluation was published, the results of both another assessment and CAMI have been released (Lindgreen et al., 2016; Sczyrba et al., 2017). The first study

includes an evaluation using *in silico* evolved genomes. This approach, with its artificially evolved sequences, complements the clade exclusion approach taken here where we use both computationally simulated and real sequences. One additional notable difference is that their evaluation looked only at genus and phylum level classifications, whereas this study looks at classifications at all taxonomic levels. Furthermore, they constructed their communities to contain only 5% taxonomically novel sequences (artificially evolved sequences). Therefore, the results are not comparable to our evaluations using clade exclusion where all of the sequences are from genomes not in the reference databases of the methods, and where performance is based on classification at all taxonomic levels rather than examined just at the genus and phylum levels. However, the results of their study would be of particular interest for those studying well characterized communities.

The CAMI study created benchmark metagenomes from ~700 newly sequenced microorganisms. These genomes had varying degrees of relatedness to each other and publicly available genomes. CAMI examined not only taxonomic classification, but also assembly and genome binning, where metagenomic sequences are binned but are not given a taxonomic label. The taxonomic classification part of the study examined 4 different methods, and identified some similar issues to be aware of with taxonomic classification, such as the prediction of many small false bins by many methods. Our study complements the work by CAMI, as we evaluate more methods, and provide performance metrics at different levels of clade exclusion, rather than performance on a single dataset that has varying degrees of relatedness to publicly available genomes.

An issue in evaluations of metagenomics methods is that there does not seem to be a consensus on the way to evaluate performance. Some researchers consider classification of a read to a taxonomic level more specific than what is correct (e.g. a novel *Escherichia* species being assigned to *Escherichia coli* rather than *Escherichia*) as assigned correctly (e.g. (Wood and Salzberg, 2014)). Other researchers, however, classify these overprediction assignments as false positives or mispredictions (e.g. (Ghosh et al., 2010)). Depending on the research goal, one may prefer a more liberal or conservative method. For example, if a researcher is interested in comparing the genera in one metagenomics sample to another sample, overpredictions that are incorrect at the species level will not matter if they are correct at the genus level. The more conservative method may assign the same reads to the family level, and will thus completely miss the

relevant taxa. On the other hand, if a researcher is interested in taking all of the predictions at all taxonomic ranks, they may make erroneous conclusions that a specific species is increased in one sample over another if it is just an overprediction. It should also be stressed that many methods allow flexibility in the parameters used, so it may be possible to tune a method to be more or less conservative. However, some parameters cannot be changed, and there are fundamental differences in the ways reads are classified by different methods. For example, MEGAN4 and MG-RAST make assignments based on bit-score as the sole parameter for judging significance. Other methods, such as DiScRIBinATE, CARMA3, and MetaPhyler, employ additional measures such as alignment parameter thresholds and/or a reciprocal BLAST search step, which have been shown to improve the accuracy of taxonomic assignments in certain scenarios (Mande et al., 2012). For example, using these methods a read from a novel *Pseudomonas* species with a single hit over the bit-score threshold to *Pseudomonas aeruginosa* may not align well enough to be assigned to the species level based on the additional alignment parameters, and thus could be assigned correctly to *Pseudomonas*. However, in MEGAN4 or MG-RAST the read would pass the bit-score threshold and because there were no other hits, it would be assigned directly to *Pseudomonas aeruginosa*.

Again, careful examination of controls (like an *in vitro* mock community sequenced alongside metagenomics samples) may provide insight into the best method to use and suitable threshold cutoffs for low abundance reads, especially if that mock community includes a suitable level of diversity and/or includes species expected in the metagenomics analysis. Developers of new methods are encouraged to enable their method to be evaluated using customized reference datasets, including clade exclusion-based analysis, to enable robust analysis of their method.

2.6. Conclusions

There has been a real need for a comprehensive evaluation of metagenomics classification methods, due to the notable number of new methods being released. In this case we have focused on taxonomic classification, for which an expanded comparative analysis was needed, to build on previous assessments and include more clade exclusion-based analysis. For the methods we analyzed, there is no single method that stands out as superior to all others, as there are a wide variety of characteristics in

which the methods differ – characteristics that may make them more suitable for certain research group infrastructure, and research projects, than others. Few researchers will have the time to evaluate methods robustly themselves, so may just use the method which is most popular or easiest to use, which would not necessarily be well suited for their particular computational resources and/or goals. This evaluation explains some of the issues researchers should consider when choosing an analysis approach for their metagenomics project, and reveals that very misleading results can occur, in particular notable overprediction of the number of taxa and/or missed taxa, if an inaccurate or unsuitable analysis approach is used. The results from this evaluation will hopefully help guide researchers' decisions in selecting appropriate analysis methods suitable for their metagenomics studies. As new methods are developed, further evaluations will need to be performed, including with a reference dataset like MetaSimHC, and/or the CAMI approach. This study provides a model for such analyses to compare method accuracies and benefits, and highlights criteria that should be evaluated. It would be very helpful for evaluation purposes if method developers would allow their method's reference databases to be manipulated, to permit analyses like clade exclusion, in order to avoid biases that can occur when no clade exclusion is performed (including with unpublished genomes as planned for CAMI, depending on the relatedness of other taxa to these unpublished genomes). Regardless, researchers are strongly encouraged to include appropriate negative and positive controls for their metagenomic experiments, including appropriate *in vitro* mock communities reflecting their expected type of data (high/low diversity, well characterized previously or not, etc.) to help fine tune their methodology as appropriate for their specific experiment. Robust metagenomic data analysis is absolutely critical at this stage of the development of microbiome research as a key research area. Microbiome research promises to be widely applicable to many, studying human health, the environment, agriculture, mining and other natural resource management, but it will be valuable only if high-quality, careful analysis is performed.

Chapter 3.

Metagenomics and marker analysis of watershed microbial communities

This chapter presents analyses of the watershed samples collected from the Applied Metagenomics of the Watershed Microbiome project. Section 3.1 details analyses of overall ecological trends and changes at the community level at sites from three different watersheds: a pristine watershed, an agricultural watershed, and an urban watershed. It also compares the microbial communities of the same samples analysed using 16S rRNA sequencing versus shotgun metagenomics sequencing. Section 3.2 describes the identification and verification of biomarkers distinguishing agriculturally affected sites from an upstream site. Finally, Section 3.3 explores freshwater bacterial pathogen abundance in the watershed samples, primarily focused on Legionella which was found in all watersheds and sampling sites.

I completed all work presented in this chapter with the following exceptions: Miguel Uyaguari-Diaz led the collection, processing, library construction, and sequencing of the watershed samples, as well as the qPCR validation testing of biomarkers; Thea Van Rossum performed the shotgun metagenomics read quality control and subsampling; Anamaria Crisan performed the 16S read processing and analyses in Section 3.1; and our collaborators at the Centers for Disease Control and Prevention (Brian Raphael, Jason Caravas, Shatavia Morrison, and Jeffrey Mercante) performed the mip gene and 18S rRNA analyses in Section 3.3.

3.1. Bacterial community dynamics in watershed ecosystems affected by different land use as measured by 16S rRNA and metagenomic sequencing

3.1.1. Abstract

Water quality assessment techniques may benefit from high resolution microbial community data that can be ascertained through genomic sequencing. Here we present a longitudinal study that evaluates the spatial-temporal dynamics of bacterial community composition of six sites drawn from three freshwater watersheds that had distinct land-use patterns: agricultural, urban and protected. We derived the monthly site-specific bacterial community composition over a one-year period using both metagenomic and amplicon sequencing. Sites indicated to be affected by agricultural pollution exhibited the lowest alpha diversity, compared to unaffected sites. We found that agricultural sites showed a dynamic change in bacterial community composition between dry and rainy seasons that was absent or less pronounced in other watershed sites. While the results between amplicon and metagenomic sequencing tended to agree, we show that differences between these methods arise from reference database completeness and method sensitivity. Finally, we identify taxa with marked temporal patterns of abundance, including a possible unidentified species of *Pseudomonas* whose temporal dynamics underlie changes in affected agricultural watershed sites. This study interrogates watershed microbial profiles with multiple sequencing techniques across time and space, illustrating the value of temporal data and providing key insights regarding microbiota variability that will aid further water quality test development.

3.1.2. Introduction

Contemporary methods for monitoring water quality rely on centuries old technologies: coliform and *Escherichia coli* counts that are inaccurate and do not broadly examine bacterial community compositions at source watersheds (Cook et al., 2013; Lin and Ganesh, 2013). By limiting the resolution of water quality testing to specific organisms, these testing methods fail to identify other potentially pathogenic bacterial organisms and may also potentially miss indicators of watershed ecosystem perturbation that affects drinking water quality long-term. Developments in genomic sequencing technologies may offer a better assessment of water quality by more broadly profiling

bacterial communities, thereby offering a more complete picture of water quality and overall watershed health.

Studies of bacterial microbial communities have been conducted primarily through amplicon sequencing of the 16S rRNA gene hypervariable region because it is universal, highly conserved, and shown to be relatively robust to sequencing vendor (Sinclair et al., 2015) and platform (Caporaso et al., 2012). There are, however, limitations to amplicon sequencing that hinder the efficacy of this method, namely the choice of primers used to amplify the 16S rRNA gene (Apprill et al., 2015) and the inability to identify new or divergent strains (Brown et al., 2015). An alternative to 16S rRNA gene sequencing is shotgun metagenomic sequencing, which although more expensive, is showing promise towards more complete profiling of microbial communities in synthetic (Shakya et al., 2013) and real aquatic microbial communities (Poretsky et al., 2014).

Recently there have been an increasing number of studies using metagenomic and 16S rRNA gene sequencing to profile bacterial communities in moving freshwater (lotic) systems. These include studies of the factors influencing the microbial composition of lotic systems (Read et al., 2015; Staley et al., 2014), as well as freshwater to seawater studies in river coastal margins (Fortunato et al., 2013) and estuaries (Campbell and Kirchman, 2013; Liu et al., 2015). There have also been a few studies conducted to ascertain the potential impact to human health. These studies investigate the role of land-use on water quality (Staley et al., 2014), identify potential biomarkers of fecal contamination (Newton et al., 2013), and speculate on the role of external factors that introduce pollutants, such as sewage or feces, on water quality (Korajkic et al., 2014; Sassoubre et al., 2015). Taken together, these studies indicate that studying the entire bacterial community profile of freshwater systems may be useful for identifying and evaluating markers of water quality, or as a direct water quality monitoring tool.

While 16S rRNA and metagenomics surveys offer interesting potential, they are as yet untested techniques for watershed water quality monitoring. Specifically, it is not clear whether these technologies can yield novel, useful, and interpretable biomarkers for monitoring activities. Here, we report upon the bacterial community dynamics at six sites within three different watersheds affected by different land use patterns - a

protected watershed, an agricultural watershed, and an urban watershed. Over a one-year period the bacterial community profiles of these watersheds were analyzed by both 16S rRNA and metagenomic sequencing. This is the largest study to interrogate watershed microbial community profiles with multiple sequencing techniques across time and space, and to specifically explore how this information may inform water quality testing methods.

3.1.3. Methods

Sampling sites

Water samples were collected monthly from seven sites in three watersheds across southwestern British Columbia (BC), Canada (Table 3.1). The watersheds had varying land use: an agricultural watershed, an urban watershed, and a watershed that was used for drinking water and protected from human use. For the agricultural watershed, three sites were sampled including a site upstream of agricultural activity (AUP), within a highly farmed and irrigated floodplain (APL), and a site downstream of this activity (ADS). Two sites were sampled in each of the protected and urban watersheds. For the protected watershed, a forested and protected site that serves as the source of a drinking water reservoir (PUP) was sampled, along with a site where water from the reservoir empties out of a 9 km long pipe (PDS). For the urban watershed, sites selected from a stream passing through 300 m (UPL) and 1 km of residential development (UDS) were sampled.

Samples were collected between April 2012 and April 2013 for the protected (PUP and PDS) and agriculturally affected samples (AUP, APL, and ADS). The urban affected samples (UPL and UDS) were collected from May 2012 to April 2013.

Table 3.1 Description of sampling sites across watersheds with varying land use

Watershed	Site name	Catchment land use	Description
Agricultural	AUP (Agri-Upstream)	Forest & minimal housing	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 & residential	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 & residential	Downstream of urban "pollution". Collected downstream of UPL, after passing through 1 km of residential neighborhood (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-Pipe)	Forest & pipe	Downstream of PUP-fed reservoir, which is 1 km 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.

Reprinted from (Van Rossum et al., 2015) with permission

Specimen collection, filtration, and DNA extraction

At each sampling site, we collected approximately 40L of raw water, which was pre-filtered in the field through a 105-µm spectra/mesh polypropylene filter (SpectrumLabs, Rancho Dominguez, CA) to remove debris and larger particles. Bulk water samples were transported on ice to the laboratory and processed within 24 hours. Samples were serially filtered to generate size-specific fractions relating to microbial class (Uyaguari-Diaz et al., 2015). A 1 µm pore-size filter (Envirochek HV, Pall Corporation, Ann Harbor, MI, USA) was used to capture eukaryotic-sized particles, followed by a 0.2 µm 142 mm Supor-200 membrane disk filter (Pall Corporation, Ann Harbor, MI, USA) to capture bacterial and archaeal-sized particles. The DNA from the bacterial and archaeal cells captured on the 0.2 µm filters was extracted using the PowerLyzer Powersoil DNA Isolation Kit (Mo Bio, Carlsbad, CA, USA), which uses a combination of bead-beating and chemical lysis. Particle-associated bacterial cells would have been removed by the 105-µm and 1 µm filters, so the microbial community collected off the 0.2 µm filter was composed primarily of the free-living bacterioplankton community.

Amplicon and shotgun metagenomic sequencing

We generated amplicons targeting the hypervariable V3-V4 regions of the bacterial 16S rRNA gene (Caporaso et al., 2011; Muyzer et al., 1993). Amplicons were purified with a QIAQuick PCR Purification Kit (Qiagen Sciences, Maryland, MD) according to the manufacturer's instructions and sequencing libraries prepared using the NEXTflex ChIP-Seq Kit (BIOO Scientific, Austin, TX) using the gel-size selection option as per the manufacturer's instructions.

Shotgun sequencing libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina, Inc., San Diego, CA). Instead of bead-size selection, an automated gel-size selection step targeting fragments between 500-800 bp was introduced to the manufacturer's protocol (Uyaguari-Diaz et al., 2015). Then, normalization and pooling of libraries was conducted as recommended by the manufacturer.

Amplicon and shotgun libraries were sequenced on the MiSeq (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 Reference Microbiology Laboratory. Raw sequences are deposited on the short read archive and can be accessed via the BioProject ID 287840.

Shotgun metagenomics quality control

Shotgun sequencing reads were first trimmed to remove low quality bases using Trimmomatic (Bolger et al., 2014). Trimming at the 3' end was performed with a sliding window of 5 bases and a minimum Phred score of 20, and at the 5' end, consecutive bases with Phred scores of less than 20 were removed. Sequencing adapters were then removed using CutAdapt (Martin, 2011), overlapping paired-end reads merged with PEAR (Zhang et al., 2014), and reads shorter than 100 bp were discarded. Samples had between 418,538 to 2,165,162 reads following this processing. All samples were subsampled to 418,500 reads, the smallest number of reads in a sample (sample 32); the January 2013 sample from the PUP site was omitted from analyses due to too few reads (sample 82). Sequences were classified by aligning to the nr database (downloaded on April 9, 2014) with RAPSearch2 (Zhao et al., 2012) version 2.18 with a log_10(E-value) cutoff of 0.1, followed by taxonomic classification with DiScRIBinATE (Ghosh et al., 2010) using default parameters, including a minimum bit score cut-off of 35. Assembly of the samples with a relative abundance ≥ 0.05% of Pseudomonadaceae was performed with MEGAHIT (Li et al., 2015), using a minimum k-mer size of 53 and maximum of 123. Assembled contigs over length 1000 bp were classified by CARMA3 using default parameters (Gerlach and Stoye, 2011). DiScRIBinATE and CARMA3 were used for taxonomic classification as we had found them to perform well in the evaluation of metagenomics taxonomic classification methods, with CARMA3 assigning reads to more specific taxonomic levels but taking much longer to run (Peabody et al., 2015)

16S rRNA amplicon quality control

Quality control for the 16S sequences was performed using Cutadapt (Martin, 2011) followed by Trimmomatic (Bolger et al., 2014) with default parameters, removing reads from the dataset if, following quality filtering and trimming, either read in the pair was less than 200 bp. Following this processing, samples were subsampled to 10,000 reads, with samples that had fewer than 10,000 reads removed from further analysis: the May 2012 samples from the ADS, PDS, and UDS sites, the September 2012 sample from the PUP site, and the December 2012 sample from the UDS site. Mothur's MiSeq

protocol was used to generate Operational Taxonomic Units (OTUs) and their taxonomic assignments (Kozich et al., 2013; Schloss et al., 2009).

Data analysis

All analyses were performed in R (v.3.0.2). Shannon's diversity index (Shannon, 1948) was used as the measure of alpha diversity (within-sample diversity), and Bray-Curtis dissimilarity (J. R. Bray, 1957) was used as the measure of beta diversity (between-sample diversity); both measures were calculated using the vegan package (Oksanen et al., 2015). We clustered the samples using the Bray-Curtis dissimilarity and displayed the results using Principal Co-ordinate Analysis (PCoA) and hierarchical clustering (complete linkage).

We also compared the 16S rRNA RDP (version 9) and DiScRIBinATE taxonomy database to contextualize limitations of the metagenomic and 16S rRNA findings. We compared both the individual taxa identified by either 16S rRNA and metagenomic data and also the OTU abundance. As OTUs were classified to varying taxonomic degrees, we summed OTU abundance from lower taxonomic levels (species and genera) that could be classified to the same family.

3.1.4. Results

We obtained monthly samples from a total of seven sites over a one-year period. The PDS site (Table 3.1) was the only non-surface sampling site and was found to be very different from the other sites (Van Rossum et al., 2015) and was thus excluded from our analyses. The agricultural watershed contained sites that were indicated to be affected by pollutants (agricultural waste runoff), and so for analysis we focus primarily on this watershed and we evaluated how findings in the agricultural watershed extend to urban and protected watersheds that were not indicated to be affected by pollutants.

Amplicon and metagenomics data show similar temporal changes in community composition associated with land use, as well as dry and rainy seasons

We surveyed changes in alpha diversity over time using the Shannon (H') index, and compared the results obtained from the 16S rRNA and metagenomic data (Figure 3.1). We found that 16S rRNA and metagenomic data reported roughly the same alpha

diversity measure for each sample, with two notable exceptions: AUP-April 2012 and AUP-July 2012. Furthermore, both the 16S rRNA and metagenomic datasets found that APL and ADS sites had more variable alpha diversity than the AUP site, apart from the AUP-September sample that appears to have a lower alpha diversity.

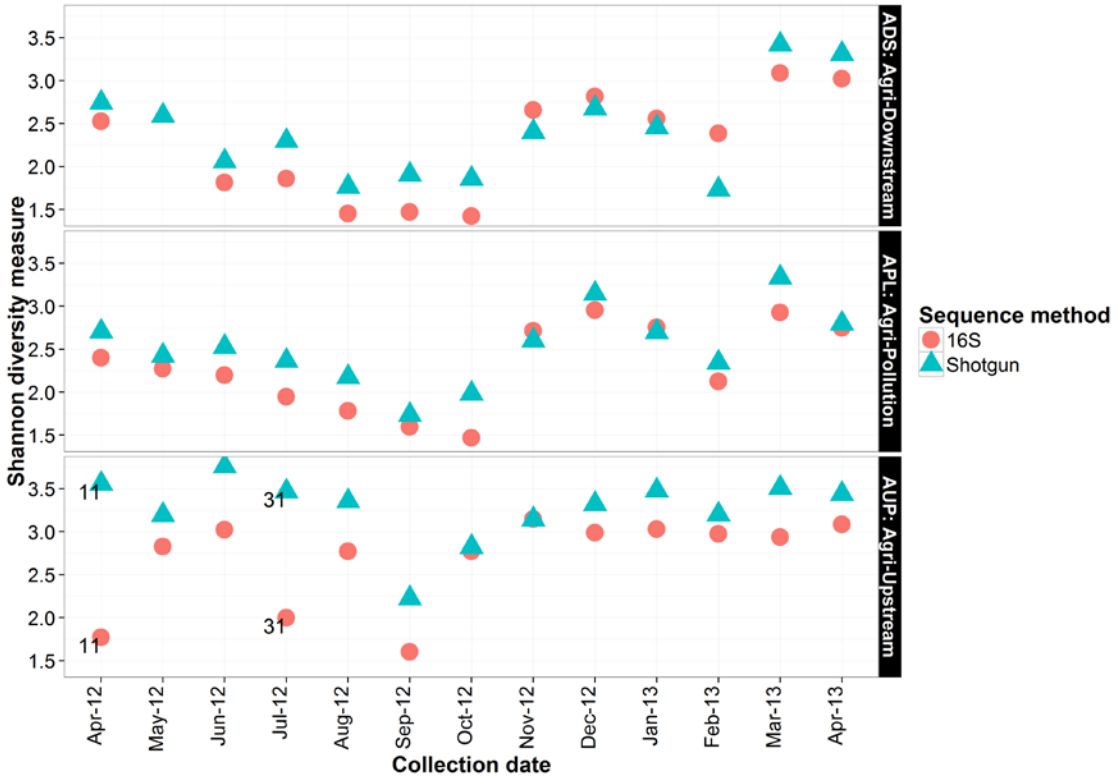

Figure 3.1 Shannon diversity over a one-year period at upstream, polluted, and downstream sites of the agricultural watershed.
The same patterns in alpha diversity are seen by the 16S and metagenomic analysis, however two samples AUP-April 2012 and AUP-July diverge substantially (see discussion).

The alpha diversity variability in APL and ADS sites shows two consistent patterns, a lower alpha diversity from May to October followed by a sustained higher alpha diversity from November to April (Figure 3.2). This trend is found with both 16S rRNA and metagenomic data, but is more pronounced in the 16S rRNA datasets. We found that this temporal change corresponded with the dry and rainy seasons in southwestern BC. This time period corresponds to when manure is permitted to be applied, from April to September each year (Figure 3.2).

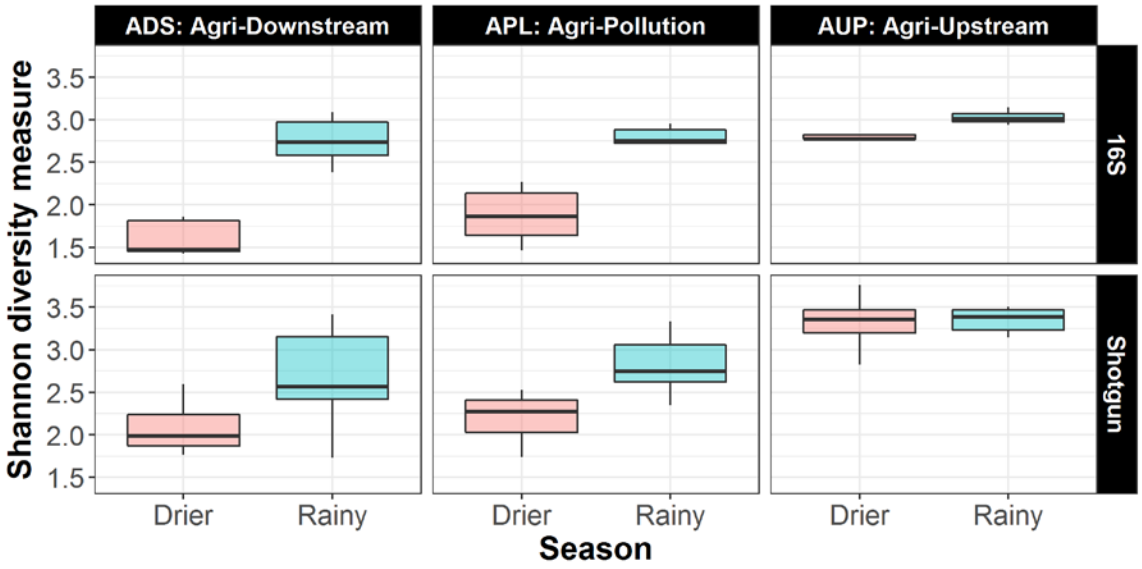

Figure 3.2 Boxplot of temporal diversity trend captured by both 16S rRNA and shotgun metagenomic data.

An increase in alpha diversity is seen in the agricultural polluted (APL) and downstream (ADS) sites in the rainy season (November – April) relative to the drier (May – October) season.

Next, we compared trends in beta diversity captured by the 16S rRNA and metagenomic data. As with the analysis of alpha diversity, we find that APL and ADS are most similar to each other, and the AUP samples appears to stand apart (Figure 3.3). AUP-September appears, once again, to be an outlier and we believe these findings are suggestive of laboratory contamination or sample mislabeling, thus we removed this sample from further analysis. Finally, we wanted to ascertain whether the sequencing-method effects were stronger than underlying biological effects, and to that end applied a hierarchical clustering to the 16S rRNA and metagenomic data (Figure 3.4). We found that samples first clustered into affected (APL, ADS) and non-affected (AUP) sites, and then further sub-clustered into 16S rRNA and metagenomic groups. This suggests that potential method specific effects do not supersede the biological effects for our data.

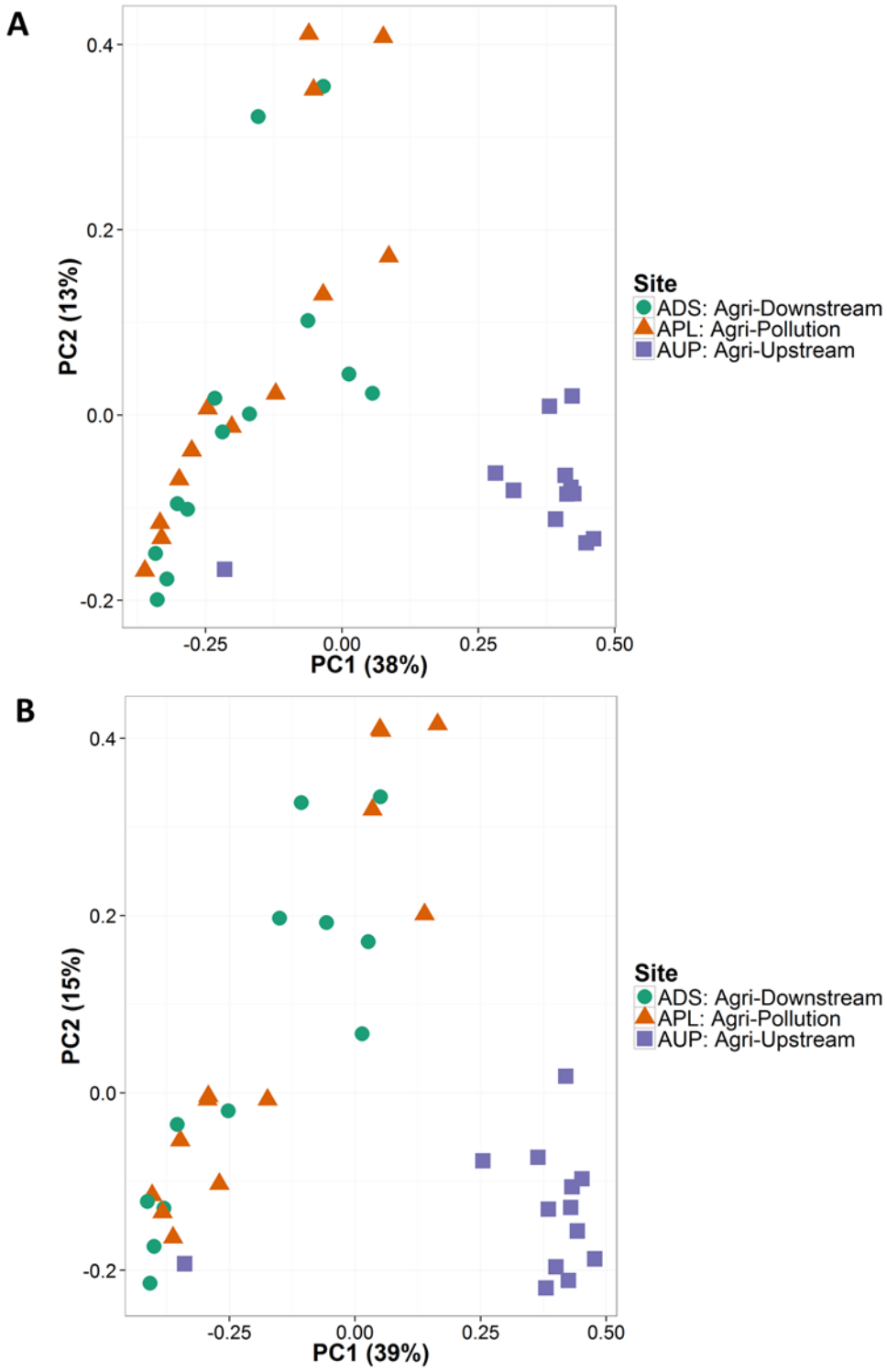

Figure 3.3 Principle coordinates analysis (PCoA) based on Bray-Curtis dissimilarity in the agriculturally affected site for the shotgun metagenomics data (A) and the 16S rRNA data (B).

Both sequencing methods reveal similar clustering of samples, with a separation of the agricultural upstream sites from the affected and downstream sites.

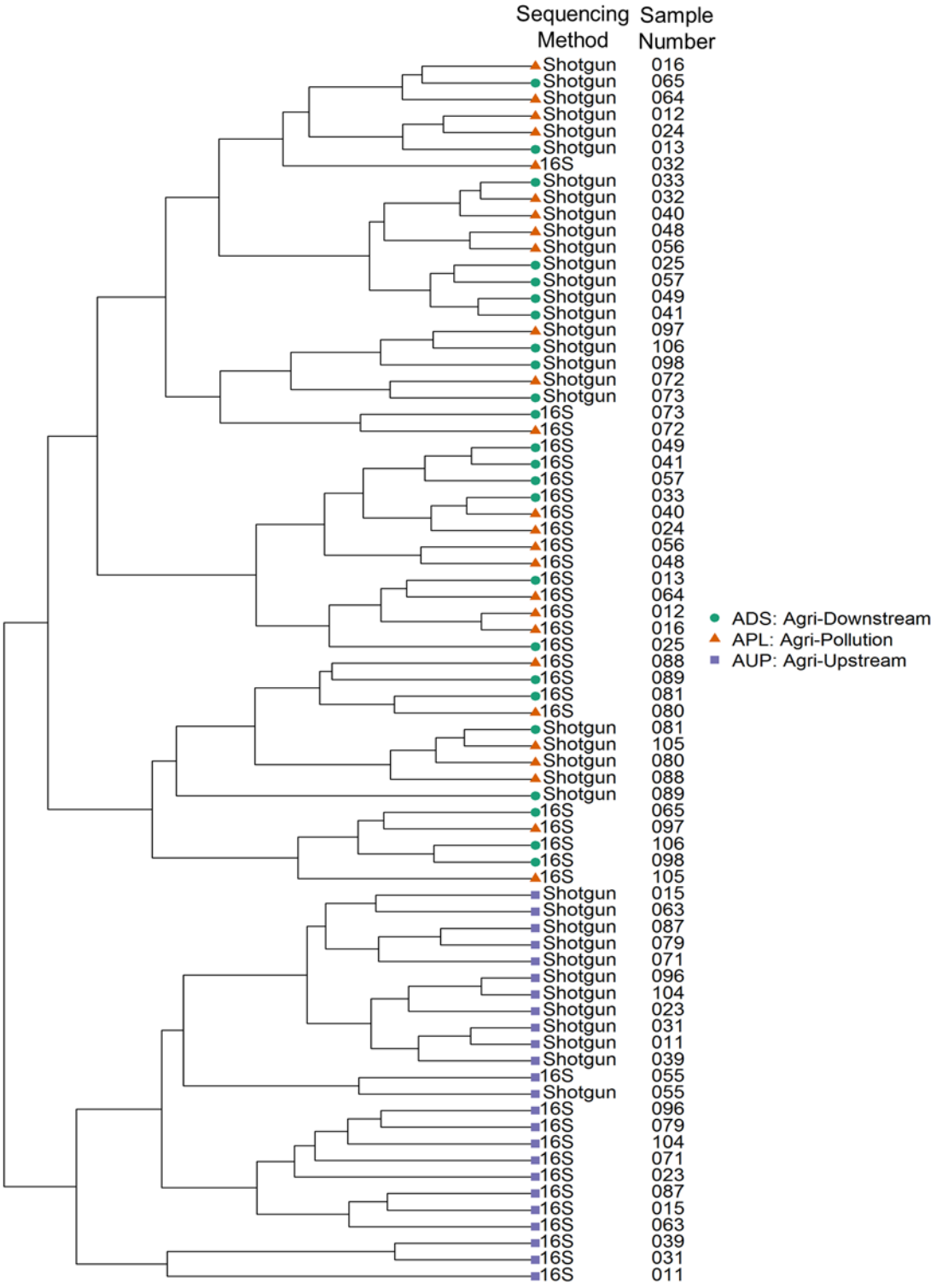

Figure 3.4 Hierarchical clustering of both the 16S rRNA and shotgun metagenomics data based on the Bray-Curtis dissimilarity measure of the agricultural watershed samples.

Samples cluster first by the upstream versus affected and downstream separation, and are mostly further subclustered into 16S rRNA and metagenomic groups.

Finally, we compared the agricultural alpha and beta diversity trends in the agricultural watershed against urban and protected watersheds (Figure 3.5, Figure 3.6). AUP samples had α- and beta diversity that was consistent with protected and urban watersheds. Furthermore, the temporal trend of increasing alpha diversity in the rainy season, observed in APL and ADS sites, is also suggested in the urban watershed, but not in the protected watershed (Figure 3.7).

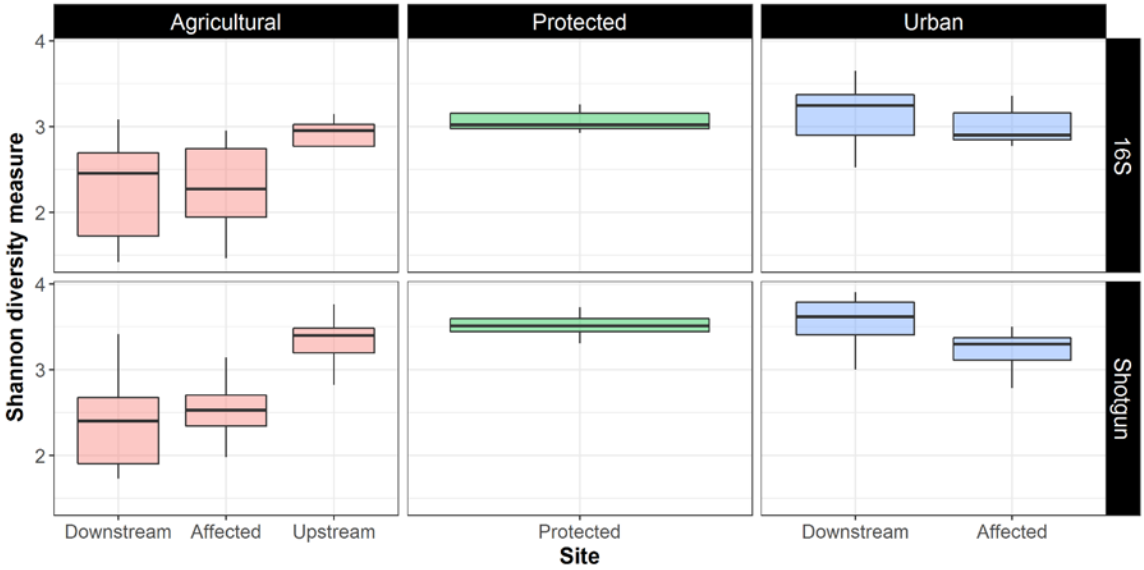

Figure 3.5 Boxplot of the Shannon diversity for the family taxonomic level reveals similar patterns between the 16S rRNA and shotgun metagenomics results for all sites in all watersheds.

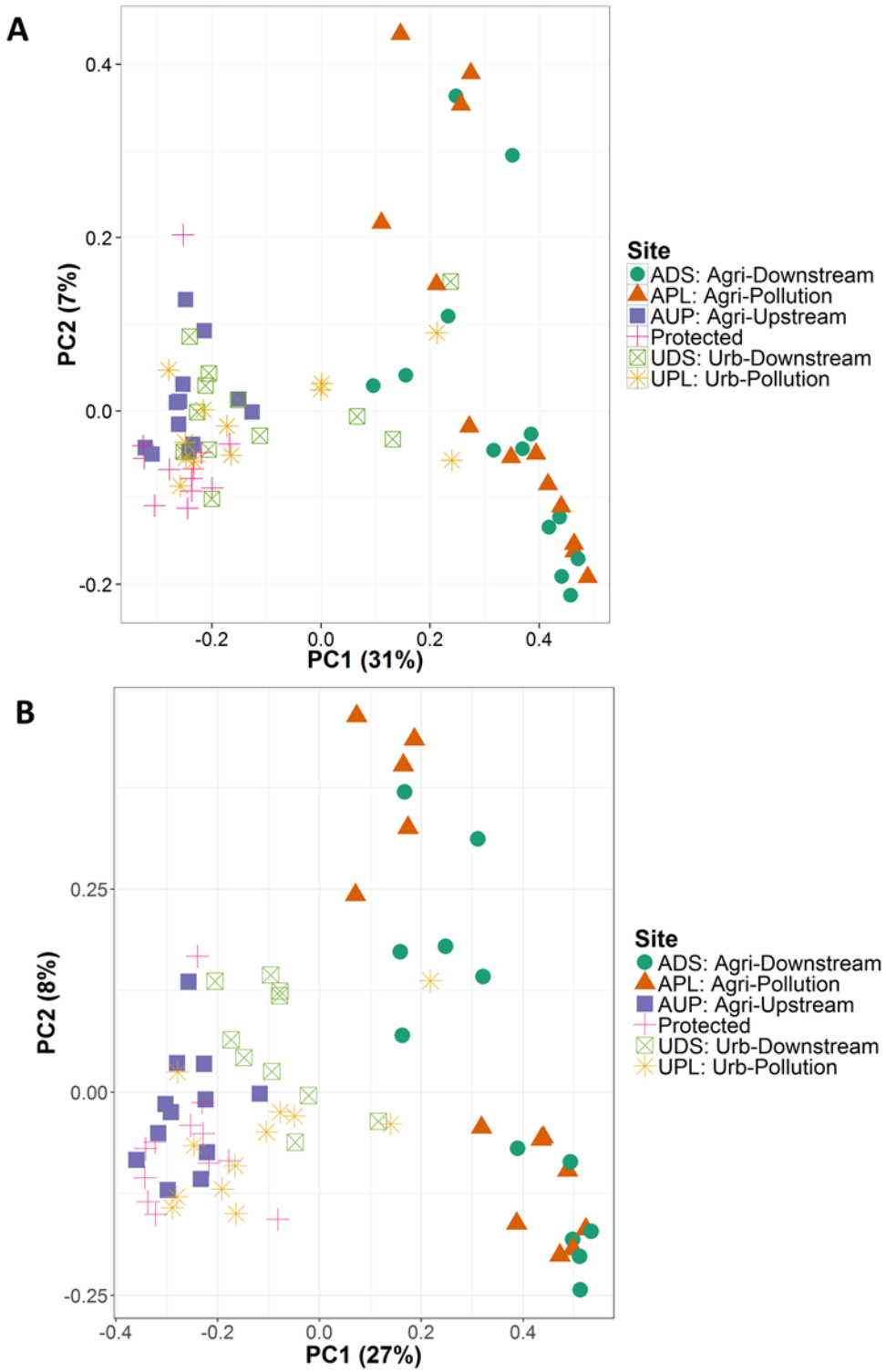

Figure 3.6 Principle coordinates analysis (PCoA) based on Bray-Curtis dissimilarity for the shotgun metagenomics data (A) and the 16S rRNA data (B).

AUP and PUP (Protected) cluster together, APL and ADS cluster together, and UPL and UDS are spread across the two clusters.

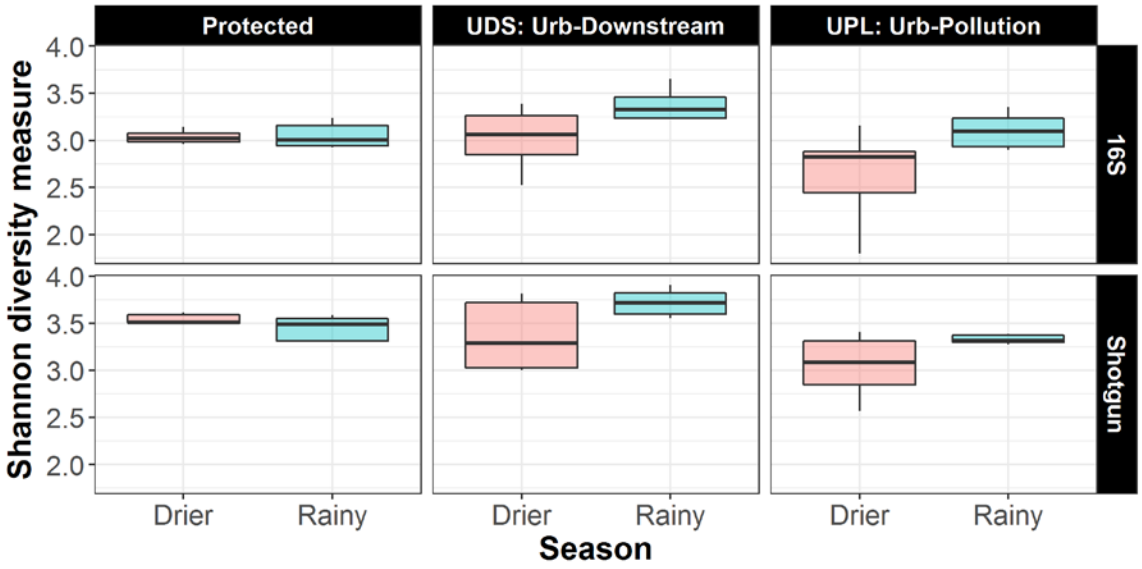

Figure 3.7 Boxplot of alpha diversy of the drier vs rainy seasons in the protected and urban watershed sites.

A slight increase in alpha diversity is seen in the urban sites during the rainy (November – April) versus the drier (May – October) season. Little change in alpha diversity is seen in the protected site.

Changes in alpha and beta diversity are recapitulated by changes in the relative abundance of specific taxa

The changes in α- and β- diversity are recapitulated by a changing bacterial community composition at the level of individual families (Figure 3.8). This shift was most prominently driven by a decreased abundance of two families (Comamonadaceae and Microbacteraceae) seen in both the 16S rRNA and metagenomic data, with a complementary increase in relative abundance of several other families. We did not find large shifts in the community composition of the AUP site, nor the sites of the protected and urban watersheds; however, we did note a decreased abundance of Comamonadaceae in the urban watershed post-September, similar to what was seen in the APL and ADS sites (Figure 3.9).

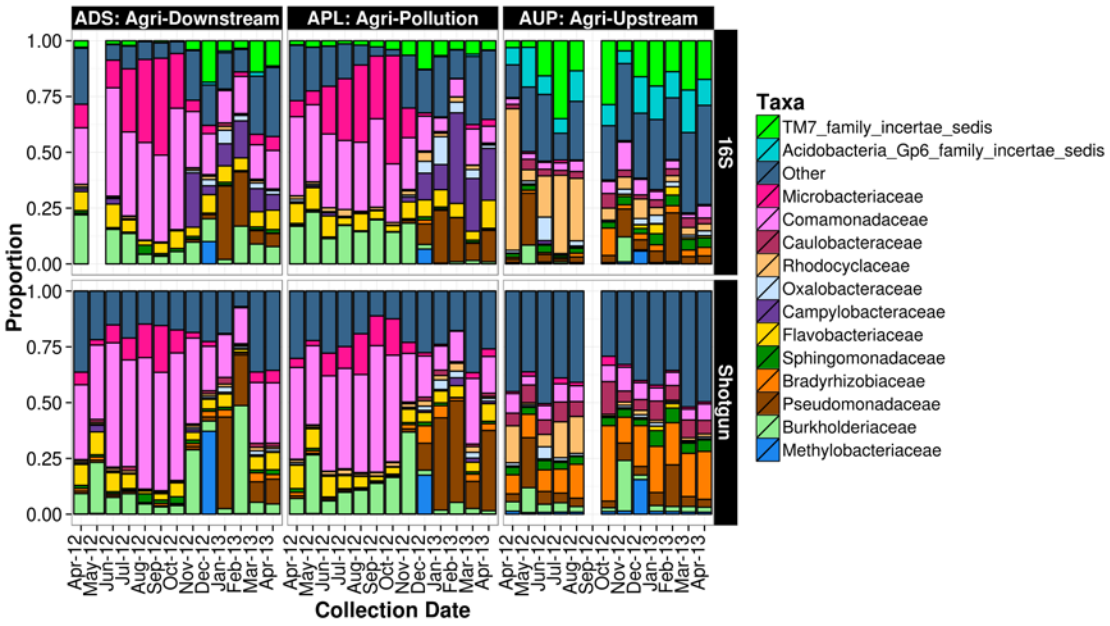

Figure 3.8 Community composition of the agricultural watershed sites over a one-year period.

A distinct shift in composition is seen from the April 2012 – October 2012 period to the November 2012 – April 2013 period in the agricultural polluted (APL) and downstream (ADS) sites. This shift is most prominently driven by a decreased abundance of two families seen in both the 16S rRNA and metagenomic data, Comamonadaceae and Microbacteraceae, with a complementary increase in proportional abundance of several other families. An increase in Pseudomonadaceae and Oxalobacteraceae is seen in both the 16S rRNA and metagenomic data, Bradyrhizobiaceae more notably in the metagenomic data, Campylobacteraceae more notably in the 16S rRNA data, and TM7 family incertae sedis only in 16S. Samples with missing data had been removed due to low read count (16S ADS-May, see methods), or suggestive laboratory contamination (AUP-47, see results).

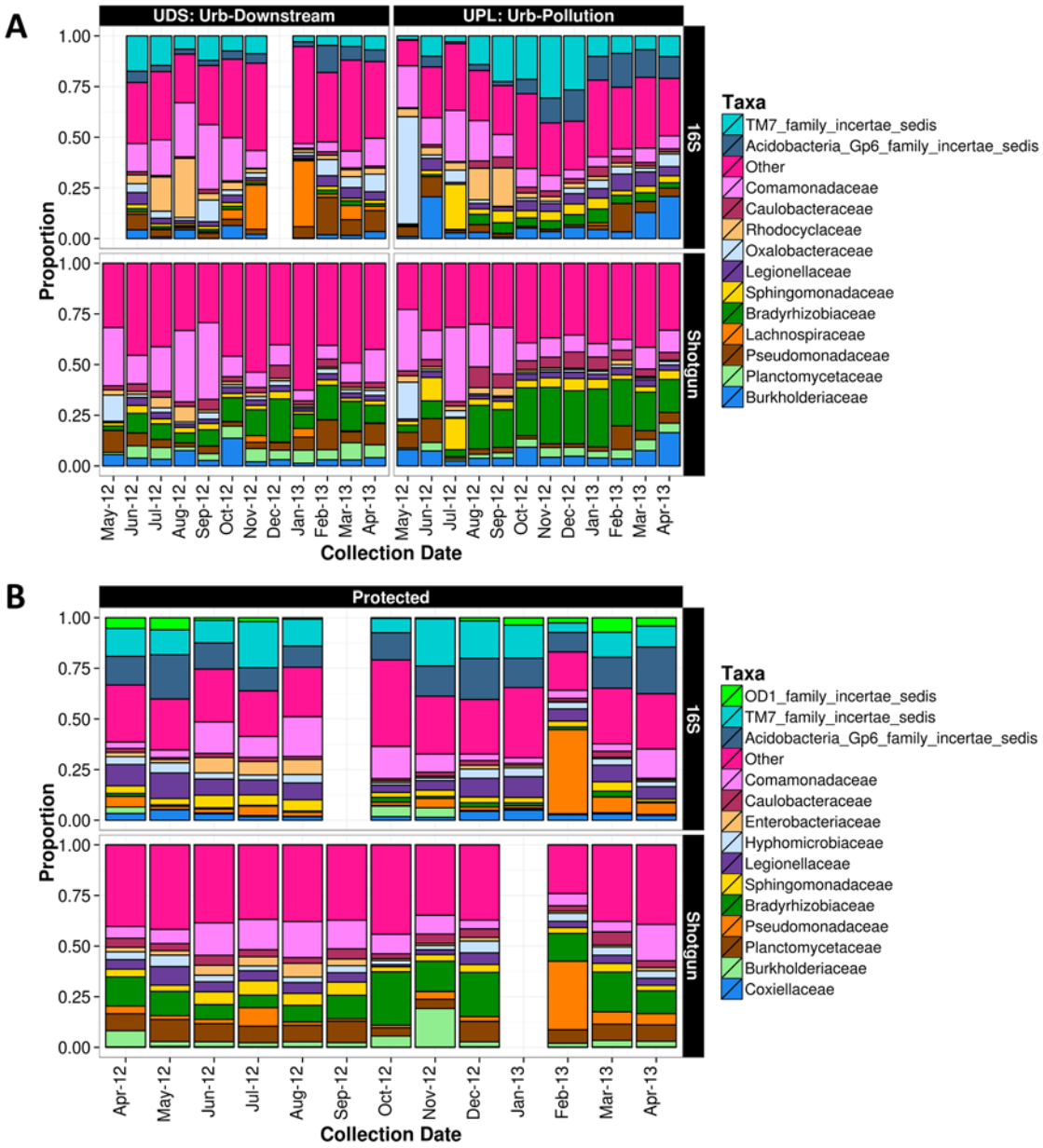

Figure 3.9 Community composition of the urban (A) and protected (B) watershed sites over a one-year period.

There is not a distinct shift in microbial composition seen over seasons, such as the shift from the May – October period to the November – April period seen in the agricultural polluted (APL) and downstream (ADS) sites. Samples with missing data had been removed due to low read count (see methods).

Taxa that are highly abundant and present in both 16S rRNA and metagenomics taxonomic databases are consistently found by both methods

While 16S rRNA and metagenomics data were generally concordant with respect to identifying taxa, we observed differences in the relative abundance of specific taxonomic families that were substantial. For example, in sample AUP-April 2012, which was shown to have the largest difference in alpha diversity between the metagenomic and 16S rRNA data, Rhodocyclaceae made up 60% of the reads assigned to the family level in the 16S rRNA data and 20% in the metagenomics data. AUP-July had the second largest difference in alpha diversity between the metagenomics and 16S rRNA data, which was in large part also due to the difference in Rhodocyclaceae (30% in 16S rRNA versus 15% in metagenomic data), as well as the large proportion of TM7 family incertae sedis in the 16S rRNA data. TM7 family incertae sedis belongs to the candidate phylum TM7, and was not predicted in the metagenomics data because the database of DiScRIBinATE does not include TM7.

We wanted to understand the limitations of the 16S rRNA and metagenomics methods by comparing the family-level taxa captured by either method. In total, we identified 249 and 231 unique families in the metagenomic and 16S rRNA datasets, respectively (Figure 3.10A). Although more unique families could be identified via the metagenomic approach, we found that a much larger proportion of 16S rRNA sequences could be classified to the family level than metagenomic data (Table 3.2).We suspect that the difference in families between metagenomic and 16S rRNA datasets could arise from two sources: taxa with low abundance (captured by one method, but not the other, or incorrectly predicted by one method) and differences in reference databases (16S rRNA RDP (version 9) and DiScRIBinATE taxonomy database).

To address issues of low abundance, we limit analysis to families if their average abundance was greater than 0.01% of the overall dataset abundance in either the 16S rRNA or metagenomic datasets. This increased the concordance between the two methods, while decreasing the number of families identified. Separately, we then limited analysis to only families that were found in both 16S rRNA and metagenomic databases. While this provided greater concordance, there were still differences (Figure 3.10A). We found that taxa predicted by one method but not the other still generally had lower abundance compared to shared taxa (Figure 3.10B). Finally, we limited analysis based

upon both abundance and database completeness and this resulted in the greatest concordance of families detected by both methods; there were only 5 families that were detected in one method but not the other: Holophagaceae and Sinobacteraceae detected only by 16S rRNA and Methanobacteriaceae, Methanocorpusculaceae, and Myxococcaceae detected only by metagenomics.

Taken together, these results suggest that database differences and low abundance taxa account for the greatest differences in taxa that can be captured by only one method. These low abundance taxa may truly be present at low levels, or they may be erroneously detected. Highly abundant taxa that present in both RDP (version 9) and DiScRIBinATE taxonomy databases are well captured by both metagenomic and 16S rRNA sequencing.

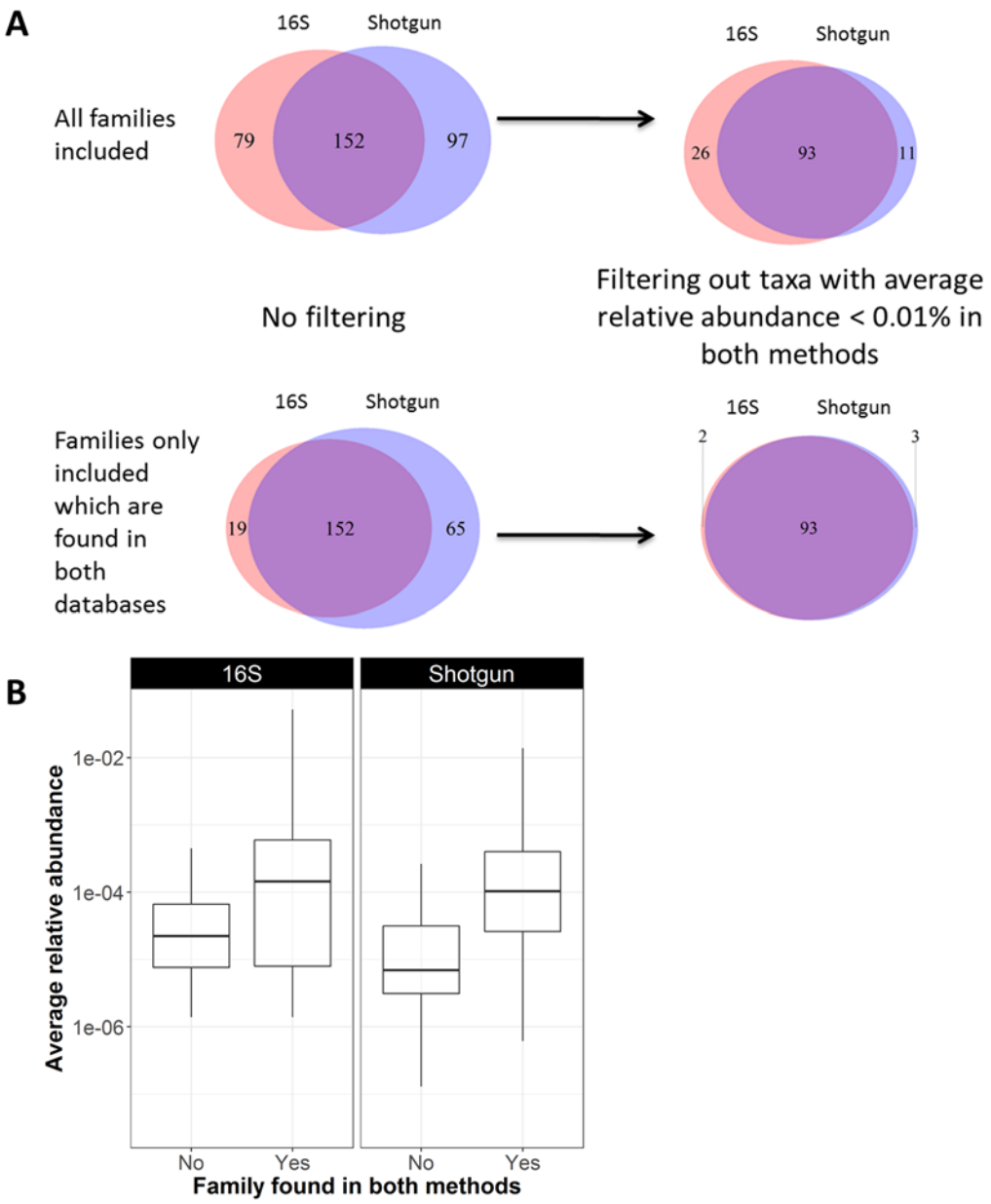

Figure 3.10 The number of shared and unique families found in each of 16S rRNA analysis and the shotgun metagenomics analysis (A), and boxplot of the average relative abundance of families found by only one method versus families found by both metagenomics and 16S (B).

Families found by only one method tend to be at lower abundance, indicating possible false positives (as shown in B).

Table 3.2 Percentage of reads that could be assigned to the family level in the 16S and shotgun metagenomic datasets across watershed sites

Sequence Method	Site Name	Min	Max	Mean	SD
16S	ADS	34.2	56.8	42.2	6.9
16S	APL	16.9	58.5	40.4	10.3
16S	AUP	11.9	38.4	23.6	7.9
16S	PUP	11.3	30.4	19.9	4.6
16S	UDS	14.6	39.2	25.8	7.2
16S	UPL	6	46.4	21.5	10.4
Shotgun	ADS	11.2	37.5	21.2	6.7
Shotgun	APL	14.7	31.5	20.9	5.8
Shotgun	AUP	8.1	19.8	11.1	3.4
Shotgun	PUP	6.6	12.5	8.9	1.9
Shotgun	UDS	8.1	21.4	12.1	3.8
Shotgun	UPL	7.9	20.4	13.1	4.2

A greater percentage of reads could be classified to the family level by 16S relative to metagenomic sequencing.

Of the taxa predicted in both the 16S rRNA and metagenomics analysis, certain taxa are predicted at substantially differing abundance between the two methods

Differences in relative abundance of specific taxa can be due to the sensitivity of the metagenomic and 16S rRNA methods. For example, 16S rRNA data found Acidobacteriaceae in high relative abundance amongst a number of samples, and was one of the top 10 most abundant taxa in the agricultural watershed sites. The metagenomic data did not indicate that Acidobacteriaceae had high relative abundance in the samples. This difference was due to lower sensitivity in the metagenomics data because we could identify the Acidobacteria phylum level but could not further classify those reads at the family level.

To assess how well predicted abundances agreed between 16S rRNA and metagenomics methods, we compared the profiles among families predicted to be present by both methods. We took the 93 families that were predicted by both methods and whose average abundance was greater than 0.01% of the overall dataset abundance in either the 16S rRNA or metagenomic data (see Figure 3.10A) and divided the relative abundance of each family in the 16S rRNA dataset by the abundance in the

metagenomic dataset to get the fold change difference. The median fold change was 1.61, with 19 families having a fold change >= 4, six families having a fold change <= 0.25 (1/4), and 68 families having a fold change between 4 and 0.25 (Table 3.3). Notably, of the families predicted with a fold change in the 16S rRNA versus the metagenomic data of ≥ 4, five of them, Campylobacteraceae, Lachnospiraceae, Ruminococcaceae, Helicobacteraceae, and Prevotellaceae, contain members which are associated with the gut microbiome (Biddle et al., 2013; Eckburg et al., 2005). For example, in the 16S rRNA results, Campylobacteraceae is a notable component of the bacterial communities in ADS and APL samples from November-2012 to April-2013 (Figure 3.8), but is found at much lower abundance in the metagenomics data. This shows that potentially pathogenic families found at high abundance in one method can be almost completely missed in another. Together, these results indicate that if looking for taxonomic markers of fecal pollution within families of known mammalian gut microbes such as Campylobacteraceae, 16S rRNA may be preferable to metagenomic sequencing.

Table 3.3 Families with a fold change in abundance ≥ 4 in one sequencing method relative to the other

Family	Fold change of abundance in 16S/metagenomics
Peptostreptococcaceae	754.3
Halomonadaceae	27.6
Coxiellaceae	27.3
Campylobacteraceae	21.6
Synergistaceae	20.1
Lachnospiraceae	17.3
Fusobacteriaceae	14.3
Veillonellaceae	13.2
Anaerolineaceae	9.9
Kineosporiaceae	9.3
Acidimicrobiaceae	7.0
Ruminococcaceae	5.4
Helicobacteraceae	5.4
Prevotellaceae	4.5
Rhodocyclaceae	4.3
Bacteriovoracaceae	4.3
Cryomorphaceae	4.1
Neisseriaceae	4.0
Rickettsiaceae	4.0
	Fold change of abundance in metagenomics/16S
Planctomycetaceae	151.7
Streptomycetaceae	12.2
Conexibacteraceae	4.2
Cytophagaceae	4.2
Opitutaceae	4.0
Mycobacteriaceae	4.0

Metagenomic data suggested a potentially uncharacterized species of Pseudomonas that relates to temporal alpha diversity changes in agricultural watersheds

Irrespective of the sequence technology used, taxa specific trends were noted that were associated with time, space, and land use. For example, Pseudomonadaceae stood out as being one of the top 10 most abundant families found across all three watersheds, in some samples making up over 25% of the assigned community (ADS-January, Figure 3.8). We also found a temporally interesting abundance pattern, with relatively high abundance seen in all sampling sites in February (Figure 3.11). In samples APL-January, ADS-January, APL-February, ADS-February, and APL-April 2013, Pseudomonadaceae accounted for more than 5% of the relative abundance in the samples. We assembled reads from those samples using MEGAHIT (N50: 1357, maximum contig length: 89619, number of contigs: 10469). Of the 10469 contigs, 4219 had a length of at least 1000, which we required in order to generate reliable species level taxonomic classification. Using CARMA3, we assigned 2060 sequences (49%) to *Pseudomonas*, and 42 (0.01%) sequences were assigned to four *Pseudomonas* species. To compare these values to the proportion of reads classified at the species level when a genome is present in the reference database, we also analyzed the in vitro mock community, which included three different *Pseudomonas* species (Table 2.1), with CARMA3. Out of a total of 96020 reads assigned to *Pseudomonas*, 76254 sequences (79%) were correctly assigned to these three species, 25 sequences (0.02%) were incorrectly assigned to other *Pseudomonas* species not in the mock community, and 19741 sequences (21%) were assigned directly to *Pseudomonas*.

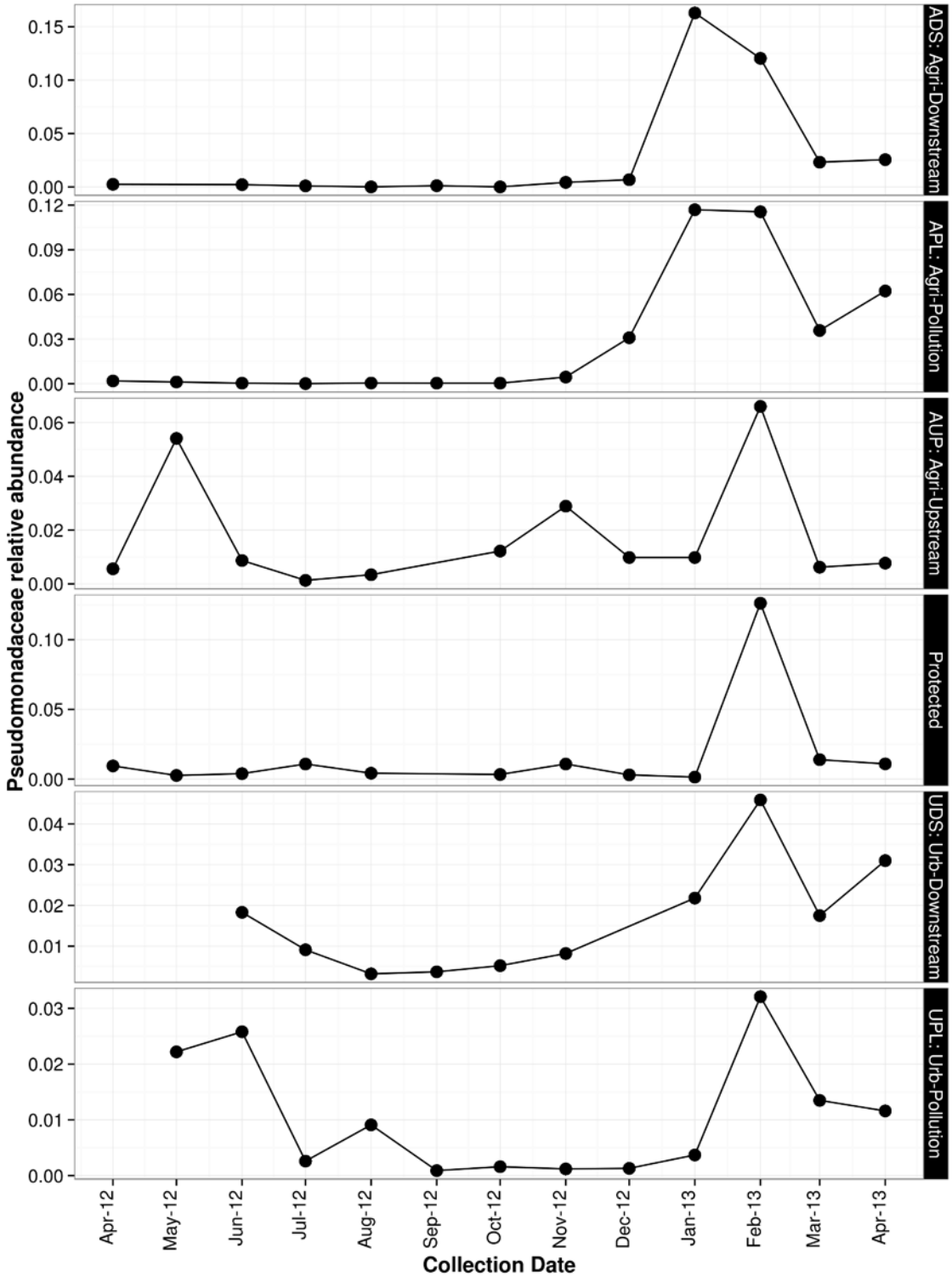

Figure 3.11 Pseudomonadaceae relative abundance across watershed sites from the 16S rRNA data.

There is a peak or near peak in abundance seen in all samples around February. Note the different scales on the y-axis, which was done to emphasize the concordance in the peak in abundance.

3.1.5. Discussion

We presented a comprehensive study of watersheds over space and time, analyzed by both 16S rRNA and metagenomic sequencing. We have shown that 16S rRNA and metagenomics data give relatively consistent bacterial community compositions, although differences existed. These differences were due primarily to taxonomic database differences, and different detection sensitivity between the two methods.

Our study showed a temporal change that could have been missed by less frequent sampling. Thus, the practice to sample a site only once or twice, as is common in prior studies, could have produced spurious associations or completely missed an important trend. For water quality monitoring, high abundance taxa classifiable to a least the family level may yield actionable biomarkers. We show that freshwater ecosystems can have pronounced changes in bacterial community composition, necessitating multiple sampling over time.

Most important, however, we have demonstrated that either metagenomic or 16S rRNA gene sequencing has potential utility in novel freshwater ecosystems that are relatively under-explored in prior research. Our results have implications for the development of future water quality biomarkers.

Temporal changes must be considered when developing water quality biomarkers

Agricultural polluted (APL) and downstream (ADS) sites had fluctuations in α- and beta diversity and community composition that appeared to coincide with changes in seasonal climate and local agricultural practices. The finding of increasing alpha diversity during rainy winter months is consistent with results found in other aquatic ecosystems (García-Armisen et al., 2014; Gilbert et al., 2012). We hypothesize that the lower alpha diversity we observed in our samples between May and October may be the result of high levels of nutrients due to agricultural activity, along with limited disturbance due to minimal rainfall, which could allow certain copiotrophic organisms to thrive and dominate the community, reducing diversity (Fierer et al., 2007). Once the rainy season starts, microbial competitors and predators may enter the watershed through surface water due to increased rainfall, increasing the diversity (Pernthaler, 2005).

Evidence of this temporal change also has implications for analyzing bacterial community structure and developing biomarkers. First, it is important to take multiple measurements over time. Second, if attempting to identify novel biomarkers, ones that may inform water quality, it may be important take seasonality into account. Failure to account for these changes could lead to incomplete community characterization.

Metagenomic data provides a good resolution of community composition and individual taxa

The targeted approach of 16S rRNA gene sequencing in addition to broad prior characterization of bacterial communities via the 16S rRNA gene may make this approach more robust in environmental samples. We found this to be true in our study as the majority of metagenomic data could not be assigned at any taxonomic level and furthermore 16S rRNA data could be more consistently classified to lower taxonomic levels. However, a recent study has shown that a considerable group of bacteria would not be identified by typical 16S rRNA sequencing due to divergent 16S rRNA gene sequences (Brown et al., 2015). As this group, and any possible others with divergent 16S rRNA sequences, have their genomes sequenced and deposited in reference databases, metagenomics will be able to identify these bacteria whereas 16S rRNA will not.

Moreover, the metagenomics data allows the identification of gene functional category biomarkers in addition to taxonomic biomarkers. Functional biomarkers may prove more consistent than taxonomic biomarkers, as different species can fulfill similar functional roles, which is supported by studies finding a lack of similarity in species composition among similar habitats, but similarity in functional composition (Burke et al., 2011). Although there are now methods available to infer functional profiles from 16S rRNA marker gene sequences, the accuracy of these predictions is low in phylogenetically novel and diverse communities (Langille et al., 2013).

Finally, with deep enough sequencing, metagenomics provides the opportunity to assemble microbial genomes, particularly those found at high abundance, which is not possible with 16S rRNA data (Janda and Abbott, 2007). The resolution and sensitivity of metagenomics methods can be improved by greater depth of sequencing; however, even as a first pass with lower levels of resolution, we found that metagenomics data could accurately characterize general trends in community composition and dynamics.

Taxa present in mammalian guts are preferentially found in 16S rRNA analysis

Several of the characterized bacterial families were found to be related to the mammalian gut, namely Campylobacteraceae, Lachnospiraceae, Ruminococcaceae, Helicobacteraceae, and Prevotellaceae. These families would be of interest when developing water quality monitoring biomarkers, because they may identify potentially pathogenic bacteria that are not captured by water quality testing methods currently in use. For example, Lachnospiraceae have been used for developing alternative fecal indicators (Newton et al., 2011), and other families contain pathogenic species. For example, Campylobacteraceae contains *Campylobacter jejuni*, and Helicobacteraceae contains *Helicobacter pylori* (Kusters et al., 2006; Young et al., 2007).

The ability to detect taxa associated with the mammalian gut also points to an important deficiency of bacterial databases, 16S rRNA or otherwise, which is the tendency to be human-centric. Freshwater ecosystems present a particular challenge because many of the organisms that reside in them may be novel, and thus much of their sequence data may not be classifiable to known organisms. For example, an average of 19.9% of 16S reads and 8.9% of metagenomic reads could be classified to the family level in the protected site (Table 3.2). This poses a challenge when trying to develop biomarkers for water quality monitoring that are useful and interpretable, because unclassifiable OTUs require more time and resources to resolve and the lack of supporting contextual information may be unacceptable to end users developing water quality tests.

Select taxa vary notably in abundance and one is found in high abundance across all sites during a single month

Pseudomonadaceae caught our interest as an occasionally abundant, temporally interesting family. Furthermore, a previous group studying the Mississippi river had noticed occasional large abundances of Pseudomonadales over two years (Staley et al., 2014). They found that at 2 out of 11 sites sampled in either year, a majority of sequence reads classified to the order Pseudomonadales. Thus, Pseudomonadaceae was investigated in further detail. Our *in silico* analysis suggest the possibility of a potentially uncharacterized species of freshwater *Pseudomonas*, as reads containing *Pseudomonas* species in a control could be assigned to the species level, whereas none of the reads and only a very small percentage of assembled contigs could be assigned

to a *Pseudomonas* species in the watershed samples. These results suggest the possibility of a previously uncharacterized species of freshwater *Pseudomonas* with relatively rapid changes in abundance, indicative of a copiotrophic lifestyle. Although our results suggest a potentially uncharacterized species of *Pseudomonas*, further work is needed to confirm this finding. Furthermore, future studies could investigate further to see if there is some periodic natural rise and fall in abundance of this species, and/or if it is responding to an environmental trigger.

3.1.6. Conclusions

We have explored the challenges and potential of developing water quality biomarkers by analyzing and contrasting bacterial community spatial-temporal dynamics identified by 16S rRNA gene and metagenomic sequencing. Our work provides important first steps towards more comprehensive guidelines for developing and testing water quality biomarkers in future applications.

3.2. Identification of biomarkers related to watershed health

3.2.1. Abstract

The overarching goal of the Applied Metagenomics of the Watershed Microbiome project was to identify biomarkers that distinguish fecally polluted versus healthy water samples, and to develop molecular assays incorporating these biomarkers. As a first step, biomarkers were identified that could distinguish the agricultural polluted and downstream samples from an upstream sample in an agricultural watershed, and primers and probes for these biomarkers were developed for a qPCR test. Several of these biomarkers were validated back on the watershed samples collected for this project, demonstrating the feasibility of identifying biomarkers and developing a molecular based assay from metagenomics sequences. This work is a first step towards developing a more accurate and rapid molecular based test to replace the current culture based water quality tests.

3.2.2. Introduction

Currently, water quality testing for the presence of feces uses facultative anaerobes, such as *E. coli* or coliforms, as indicators because they are easy and inexpensive to culture. However, facultative anaerobes have been found to make up only a small proportion (around 0.1%) of the fecal microbial communities (Eckburg et al., 2005). Furthermore, the tests correlate poorly with the presence of pathogens such as protozoan parasites and viruses (Harwood et al., 2005). For example, *E. coli* has been shown to persist and sometimes even grow in the environment (Ishii et al., 2006; Wheeler Alm et al., 2003), so the presence of *E. coli* does not necessarily imply the presence of fecal pollution or pathogens. Similarly, the absence of *E. coli* does not mean the absence of more resistant organisms such as *Giardia* cysts or *Cryptosporidium* oocysts (Wilkes et al., 2009). Thus, waterborne outbreaks have occurred even in the absence of positive indicator test results (Goldstein et al., 1996). Finally, culture based tests are relatively slow, and the indicators used do not provide information regarding the source of the fecal pollution, making it difficult to identify and deal with the source of pollution.

Due to these issues with current water quality tests, researchers are interested in developing alternative indicators for water quality assessment. The hope is that these alternative indicators may be also used for microbial source tracking, to identify the host/source of fecal pollution in water. Several alterative indicators have been explored, which are generally fecal anaerobes such as *Bacteroides* (Okabe et al., 2007) or Firmicutes (McLellan et al., 2013; Newton et al., 2011), which unlike *E. coli*, are unlikely to grow outside of the host. The HF183 *Bacteriodes* 16S rRNA genetic marker, developed in 2000 (Bernhard and Field, 2000), has been found in evaluation studies to be the most effective marker of human fecal contamination (Layton et al., 2013). A recent review found that there are inconsistent results of correlations between these alternative indicators and human pathogens reported by different studies (Harwood et al., 2014). This review also found that each marker had strengths and weaknesses; for example, the HF183 *Bacteriodes* 16S rRNA genetic marker is found at relatively high concentrations in sewage, but is not completely specific for human feces. Furthermore, most of the focus has been on human sources, so there are fewer markers for domestic animal sources such as cows and pigs, and the markers that do exist are not well understood with regards to their performance or distribution. Thus, although alternative indicators show promise for use in improved water quality testing including microbial source tracking, there is an opportunity for the further development of new markers.

The ultimate goal of the Applied Metagenomics of the Watershed Microbiome project was to use metagenomics to survey microbial communities, and based on this survey identify microbial biomarkers of watershed health as well as markers indicative of fecal contamination from agricultural species. A molecular based diagnostic test was then to be developed incorporating these biomarkers. As a first step, the project team decided we would first attempt to develop a test that could differentiate between samples collected from the agricultural polluted and downstream sites (APL and ADS) from samples collected from an upstream site (AUP). Therefore, I developed a computational pipeline using freely available tools to identify candidate biomarkers, develop primers for these biomarkers, and test for the specificity of these primers. Using this pipeline, a candidate list of primer and probe sets were ordered, and validated back on our watershed samples.

3.2.3. Methods

Identification of biomarkers and development of primers

Shotgun sequencing reads underwent quality control processing as described in section 3.13. The rest of the pipeline following quality control, the identification of biomarkers and development of primers and probes for qPCR, is shown in Figure 3.12.

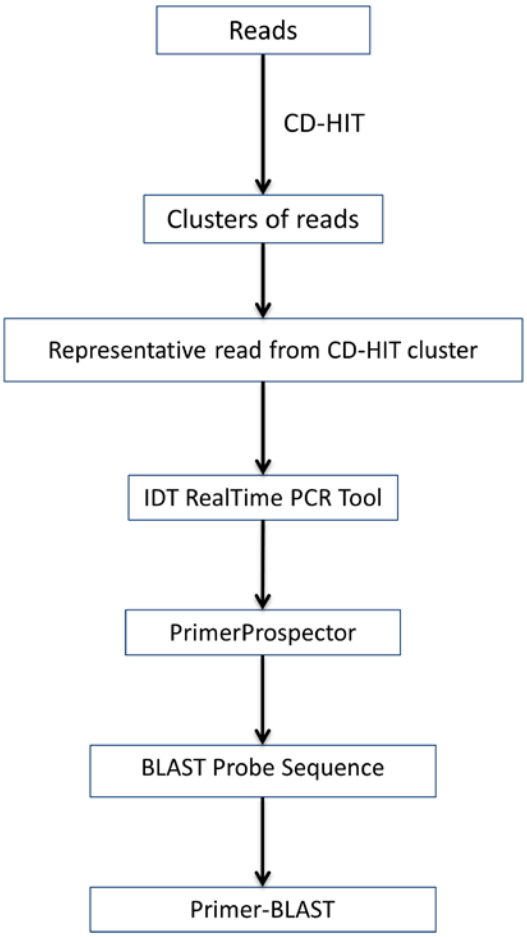

Figure 3.12 Computational pipeline for generating biomarkers and primer/probe sequences for a qPCR test.

Reads are first clustered at 95% sequence identity using CD-HIT. The representative read from the CD-HIT cluster is then run on IDT RealTime PCR Tool to generate a primer pair and probe. The primer pairs are then run on PrimerPropector on the watershed samples to generate *in silico* amplicons, and the probe sequence is checked if it aligns to the *in silico* amplicons with BLASTN. The primers are also checked for specificity with Primer-BLAST against the nr database.

First, the reads were clustered at a 95% sequence identity threshold using CD-HIT version 4.6.1 (Fu et al., 2012); clusters that contained reads exclusively from the agricultural APL and ADS sites, or with a maximum of one read from the AUP site, were considered as candidates for generating primers against. For several candidate clusters, the representative read was taken and used as a query in BLASTN and BLASTX searches against the nr database to identify the gene name and organism name from the best hit. Representative reads were chosen to move forward to the primer design stage so that there was a diversity of taxa and genes represented.

Integrated DNA Technologies RealTime PCR Design Tool (http://www.idtdna.com/scitools/Applications/RealTimePCR/) with default parameters was used to generate a candidate set of primer pairs and associated probes. To ensure the designed primers were target specific, amplification was tested *in silico* against both the watershed samples and the nr database. To test for amplification on the watershed samples, PrimerProspector version 1.0.1 (Walters et al., 2011) was used with the following parameters: 3' mismatch length set to 2, 3' mismatch penalty set to 10, non-3' mismatch penalty set to 1, and both 3' and non-3' gap penalties set to 10. These parameters were set in such a way to be able to extract the amplicons that would be generated if one allowed up to 2 mismatches, as long as they were not in the last two 3' bases.

PrimerProspector performs *in silico* amplification, but does not have the option to test a probe. Therefore, to check if the probe would anneal to the *in silico* amplicons, for each primer pair the output of *in silico* amplified sequences from the watershed samples were used to create a BLAST database, and the probes were aligned to the database using BLASTN version 2.2.31+ (Camacho et al., 2009). The following parameters modified from default: no dust filtering of low complexity regions, mismatch penalty -1, gap open cost 0. The resulting number of amplicons were counted where the probe aligned with the amplicon with 2 or fewer mismatches.

The primers were also tested for specificity against the nr database using Primer-BLAST (Ye et al., 2012). Primer pairs were uploaded and run with default parameters against the nr database, and primers that had hits to multiple species that were not in the same family were filtered out of the candidate set of primer pairs.

High-throughput multiplex quantitative polymerase chain reaction

Primers were first checked for sensitivity and specificity by SYBR green-based PCR (fluorescence seen in a test APL site sample, and not in a negative control). TaqMan probes (Life Technologies, Carlsbad, CA) were then designed and validated for the primers which showed suitable sensitivity and specificity. Assays with cycle threshold values greater than 35 were not included in the high-throughput 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).

DNA extracts from watershed 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 each 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 environmental samples, 16S rRNA primers and probe (positive control), and one no-template control.

3.2.4. Results

Following the clustering of reads at 95% sequence identity threshold using CD-HIT, there were 891 clusters of size 40 reads or greater that had reads exclusively from the agricultural APL and ADS sites, or with a maximum of one read from the AUP site.

From this initial set of clusters, the representative read from each cluster was used as a query in BLASTN and BLASTX searches against the nr database to identify the gene name and organism name from the best hit. The first approach was to take the clusters with the largest numbers of reads assigned to them, but the representative reads from these clusters all ended up hitting rRNA genes. Although a few of these were chosen to move forward to the primer design stage, to get a diversity of gene and taxa hits, other clusters from the initial set of 891 clusters were randomly sampled.

This candidate set of clusters and associated representative reads then moved forward to the primer design stage, and run on IDT RealTime PCR Design Tool. The designed primer pairs were then run on PrimerProspector to generate *in silico* amplicons, the probe was checked using BLASTN against the *in silico* amplicons, and primers were tested for specificity against the nr database using Primer-BLAST (see methods). Only 24 primer pairs with probes could be ordered for testing, so these were chosen such that the ratio of in silico amplicons generated in the APL & ADS samples versus the AUP samples was at least 10/1, the number of *in silico* amplicons generated in the APL & ADS samples was high (at least 10), and there was a diversity of genes and taxa represented. These last two criteria generally contradicted each other, as the markers with large numbers of *in silico* amplicons generated tended to be similar taxa and gene functions. Therefore, some taxa and genes were included multiple times, such as the taxon Beta proteobacterium CB and the 16S and 23S rRNA genes.

The 24 read clusters chosen and the gene and taxa obtained from the top hits from representative sequences are shown in Table 3.4. The associated primer and probe sequences, and the number of *in silico* amplicons generated for the APL & ADS and the AUP sites are shown in Table 3.5. Of the 24 read clusters for which primers were ordered, only 12 showed sensitivity and specificity at the pre-high-throughput qPCR stage (test on a single APL sample and negative control), and thus only these 12 clusters were tested on all the watershed samples. These were clusters 4- 9, 12, 14-15, 18-19, and 24. The qPCR results for clusters 15 and 18, both of which had their top hit to *Limnohabitans*, are shown in Figure 3.12. Cluster 15 worked particularly well at generating signal in the APL and ADS samples but not the AUP samples, while cluster 18 showed a fairly similar signal level for all three sampling sites. The qPCR results for the other clusters and positive control are shown in Appendix B. All clusters showed high signal/low cycle threshold values in most of the APL and ADS sites. However, several

clusters also showed similar values in the AUP site samples, and thus were not able to differentiate the APL and ADS samples from the AUP samples. Out of 24 clusters for which primers and probes were designed, 3 of these showed clear differentiation between the APL and ADS samples and the AUP samples (clusters 8, 15, and 24).

Table 3.4 Read clusters used for the generation of primers and the number of reads in the cluster from the APL & ADS sites versus the AUP site

Cluster	Gene	Taxa name	Family	Number of reads in cluster	
				APL & ADS	AUP
1	Elongation factor Tu	Beta proteobacterium CB	Burkholderiaceae	126	0
2	isoleucyl-tRNA synthetase	Ideonella sp. B508-1	Comamonadaceae	55	0
3	23S	Delftia sp. Cs1-4	Comamonadaceae	158	1
4	50S ribosomal protein L16	Beta proteobacterium CB	Burkholderiaceae	112	0
5	23S	Albidiferax ferrireducens T118	Comamonadaceae	116	0
6	16S	Uncultured actinobacterium clone CB31F01	-	151	0
7	Elongation factor Tu	Polynucleobacter necessarius subsp. necessarius STIR1	Burkholderiaceae	120	0
8	Elongation factor Tu	Beta proteobacterium CB	Burkholderiaceae	104	0
9	30S ribosomal protein S19	Bifidobacterium animalis strain RH	Bifidobacteriaceae	55	0
10	30S ribosomal protein S11	Beta proteobacterium CB	Burkholderiaceae	102	1
11	23S	Beta proteobacterium CB	Burkholderiaceae	114	0
12	23S	Polynucleobacter necessarius subsp. asymbioticus	Burkholderiaceae	166	1
13	DNA-directed RNA polymerase subunit alpha	Beta proteobacterium CB	Burkholderiaceae	80	0
14	23S	Polynucleobacter necessarius subsp. asymbioticus	Burkholderiaceae	137	1
15	aspartate carbamoyltransferase	Limnohabitans sp. Rim47	Comamonadaceae	46	0

Cluster	Gene	Taxa name	Family	Number of reads in cluster	
				APL & ADS	AUP
16	FAD-dependent thymidylate synthase	Candidatus Aquiluna sp. IMCC13023	Microbacteriaceae	42	0
17	16S	Uncultured beta proteobacterium clone 27GS2	-	180	0
18	16S-23S	Limnohabitans planktonicus strain Rim42	Comamonadaceae	177	0
19	hypothetical protein	uncultured Spirochaetales bacterium HF0500_06B09	-	80	0
20	succinate dehydrogenase	Polaromonas sp. EUR3 1.2.1	Comamonadaceae	63	0
21	conserved hypothetical protein, partial	Helicobacter pullorum	Helicobacteraceae	78	0
22	16S-ITS1	Beta proteobacterium CB	Burkholderiaceae	157	0
23	DNA-directed RNA polymerase, beta subunit	Beta proteobacterium CB	Burkholderiaceae	63	0
24	molecular chaperone GroEL	Corynebacterium-like bacterium B27	Corynebacteriaceae	46	0

Gene, taxa name, and family come from the top hits from BLAST searches against the nr database. The process for choosing these clusters is outlined in the methods and results.

Table 3.5 Primer and probe sequences tested and their *in silico* amplification

Cluster	Forward primer	Reverse Primer	Probe	*In silico* Amplicons APL & ADS	*In silico* Amplicons AUP
1	ACCGGAGATAGAGAACACGTC	GACTCATACATCCCAACTCCAG	TCTACTGGCATCAAGAACGCACCG	124	0
2	CACAAACCATCACGCACAG	GCTGTCAAACGCATCCATG	AAACACCTGTCATCTACCGCGCT	46	0
3	GAATGCGTCACTTGGCATAAC	CAGGTGTGGAAGCGTAGTAAT	AGGGTCAAGTCGCACGAGCAATTA	104	2

4	CGCAGAAGTCAGGAGAGTAAG	CAATCCAAAGTCACCAAAGGC	CGCAAGGAACAAAAGGGACGTAACAC	58	0
5	AAGACATCAACCGAGATTCCG	TCTTTTCGCCTTTCCCTCAC	TGTCCCGCCCTACTTGTCGC	52	1
6	CTTAGTCCCAATCACCGATCC	ATCATGCCCCTTATGTCTTGG	TTGCAGACCCCAATCCGAACTGA	24	0
7	ATAGGAGCGATGAGTTTTACGG	TTGGGTAAAGACGAAGGTGG	GGGTTCAATCGAGTTGCCAAAAGAC	66	0
8	GGCTGTAAAGTGAGTATGTGGG	TCAAGCCAACACTCAAGACTAC	ACCAAGGTCAAGCAGGCGATAACG	59	0
9	CGCAAGCATGTTCCAGTATTC	ATCAGTATTAGCCACGGTTAGC	AGTTCGCACCAACACGTACCTTCA	21	0
10	CACTGATCGTCAAGGAAATGC	AGGTGTTGATTTACGTGAGCC	CCGCCAGAAGTTGCCCATGAAAG	56	0
11	AAAACGCCCAGACCTAAGAG	CAGCAACTCAACCCTCTACG	ACCACTCCTTCCCATCCCGAAC	40	0
12	TCCTGACTGTTTTAGCCTTCC	GTGCCTCGTGTATCACTGTAG	TTGTTTCCCTCTTGACACCGGACG	34	0
13	AAACGGCTGTGGAGATAGAAC	GCTTCTGCTTCTGAATCTGC	AGGTTGCACGATGGGCGAATACT	31	0
14	CGACCGATAGTGAACAAGTACC	GTTTAGCCTTTCACCCCTATCC	CGTGTTACCGCACCTTCAACCT	46	2
15	CGACCGTTTTGAATCAGGATG	GTGACGTTTGGTATTGCAGTG	CATTCCCCGCCACGATAGACATCA	20	0
16	GTAGTTGATCAGACCCTTGTCG	AGGTTTTATCCGTGGCTGAG	ATGTAACTGTTGAATTGGTGCGCTCG	15	0
17	CCCGCTTTCATCCTTAGATCG	GGGCGTTTGATGGTGATTG	ATTACTCACCCGTTCGCCACTCG	20	1
18	CTACCTTGTTACGACTTCACCC	CCCATAAAACCAGTCGTAGTCC	TCTGATCCACGATTACTAGCGATTCCGA	94	5
19	TTCCTCTTAACCTTCCAGCAC	TCGACGTGAGGTTTTAGCTG	CATTTTGCCGAGTTCCTTAACCGTGG	15	0
20	AACACCATGACAAAACGCAC	TGAGTCATGTCCACGATCAAG	CCGACGCCAAGCCCTACAT	14	0
21	CAGTAGCCAGAGGAAGAGAAATC	ACTTTTCACCTTTCCCTCACG	TGGGTTTGTCCTAATCCGCTTTCGC	13	1
22	ACTTGATCCTATAACGAGCACC	ACAAGACTTAGTGGGACGTTG	TTAGATGATGGTGGAGGATGACGGGA	12	0
23	TGTGTCTTGCCCTTATCGTG	ACATATCACTCTGCCTTACGC	GTTGCACCGTTCGCC	10	0
24	GTCGTAGTCAGAGTCAGTGTTG	GTGGCATCTTCAAGTCGGTAG	TGGGTAAGGCTCGCAAGGTTGT	10	0

This table lists the primer and probe sequences for each cluster, and the number of *in silico* amplicons generated for the APL&ADS sites and the AUP site. In all cases, substantially more *in silico* amplicons were generated for the APL & ADS sites versus the AUP site.

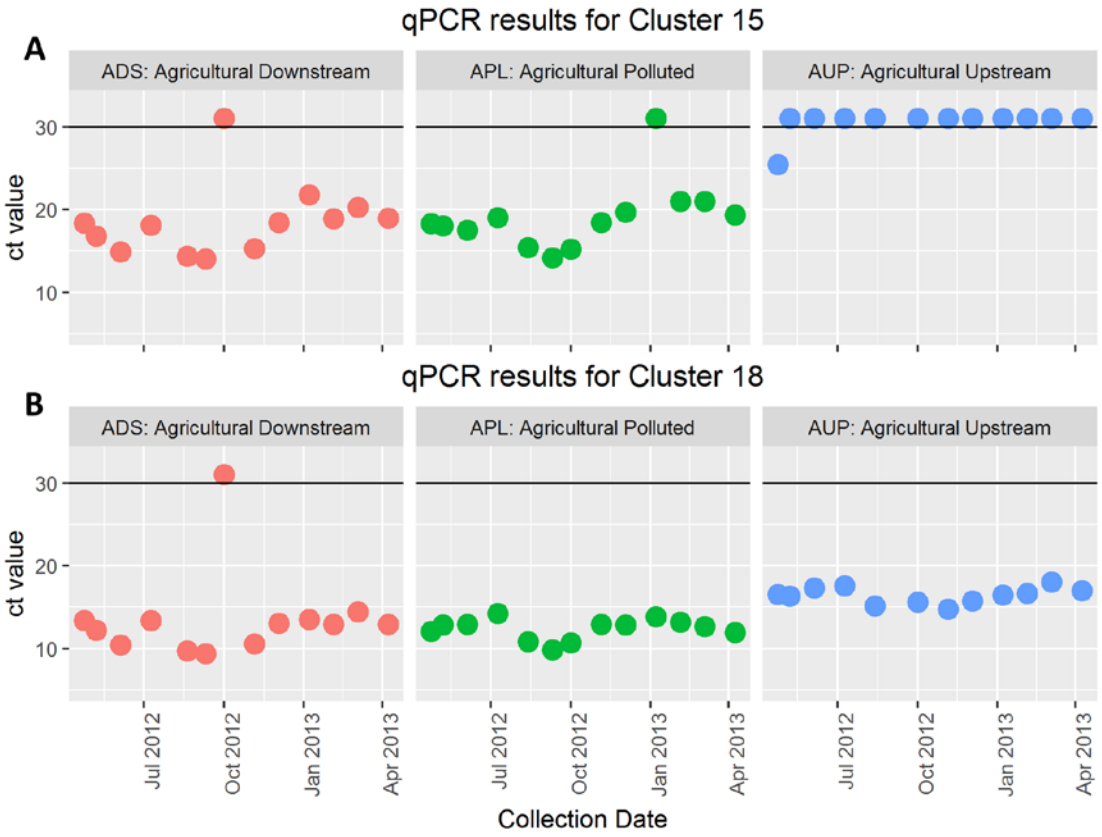

Figure 3.13 qPCR results for two different biomarkers: Cluster 15 (A) and Cluster 18 (B), which both likely target *Limnohabitans*.

These biomarkers were designed from metagenomics data and the primers and probes validated back on the watershed samples, demonstrating the feasibility of this approach to biomarker identification and development of a qPCR test from metagenomics data. However, not all markers were able to differentiate APL and ADS samples from the AUP samples, such as cluster 18 in (B).

Samples above the black line had no detectable signal.

The APL and ADS samples in the undetected region are suspected technical errors. The ADS sample had undetectable signal for a 16S rRNA primers and probe positive control, and the APL sample had undetectable signal for clusters 8 and 24 (see Appendix D).

3.2.5. Discussion

The feasibility has been shown of using freely available computational tools to identify candidate biomarkers from metagenomics samples, and to develop primers for these biomarkers that can differentiate the samples intended to be differentiated. However, of the 24 clusters for which primers and probes were designed, only 3 of these showed clear differentiation between the APL and ADS samples and the AUP samples. Notably, none of these clusters (8, 15, and 24) showed any *in silico* amplification from the AUP samples, whereas some clusters that did not show clear differentiation between

the APL & ADS samples and AUP samples, such as cluster 18 in Figure 3.13, did show *in silico* amplification in AUP samples. Nevertheless, there were still samples that did not show *in silico* amplification in the AUP samples that had cycle threshold values of the AUP samples at a similar level to the APL & ADS samples, such as cluster 19. This amplification *in vitro* that was not seen *in silico* could be due to several reasons, such as not having enough sequencing depth in the samples. Future research could take a more in-depth examination at the factors that distinguish biomarkers/primers that differentiate sites from those that do not.

Primer design and *in silico* amplification were performed on the samples rarefied down to 418,500 reads, the minimum number of reads in any sample. Although this allowed all samples to be considered equally in the primer design and *in silico* amplification stages, it also resulted in a loss of usable information, as any additional reads over 418,500 were not used. Further work could also examine if using all of the reads available results in a greater proportion of biomarkers/primers that are effective at differentiating the sites of interest.

As shown in Figure 3.13, there were two biomarkers that likely targeted *Limnohabitans*, yet one was effective at distinguishing the APL & ADS samples from the AUP samples (cluster 15), and the other was not (cluster 18). *Limnohabitans* is globally distributed in freshwater environments, but is not well characterized as it was first described in 2010 (Hahn et al., 2010). However, studies have demonstrated that *Limnohabitans* species have high rates of substrate uptake and growth, and isolated strains show a diversity of morphologies and different patterns in substrate utilization (Kasalický et al., 2013). Therefore, it is hypothesized that *Limnohabitans* strains each occupy their own specific niche. If the two clusters targeted different species or strains of *Limnohabitans*, this may explain why one of the clusters worked whereas the other did not, and may indicate the importance of generating strain specific markers.

Demonstrating the feasibility of identifying biomarkers and developing primers from metagenomics sequences, and showing this approach could differentiate between agriculturally affected sites versus an upstream site, was a first step towards developing new water quality tests. One of the next steps in testing these biomarkers/primers would be to try these biomarkers on additional samples collected at a later date from these same watershed sites, to see if they work over multiple years (temporally). Another step

would be to gather freshwater samples from further watersheds on which to test these biomarkers. Finally, towards the broader goal of developing better tests for water quality monitoring, other biomarkers could be investigated, perhaps looking at markers specific to certain pathogens.

3.3. Analysis of *Legionella* and other freshwater bacterial pathogen genera in natural environments

3.3.1. Abstract

Here, we investigate the presence of freshwater bacterial pathogens through 16S rRNA amplicon and metagenomic sequencing of DNA collected monthly from the seven sites in the three watersheds with varying land use over a period of one year. *Clostridium*, *Escherichia*/*Shigella*, and *Mycobacterium* were found in some samples, but most notably, *Legionella* was found in all watersheds and sampling sites, comprising up to 2.1% of the bacterial community composition. *Legionella* spp. present in some man-made water systems can cause Legionnaires' disease in susceptible individuals. Although legionellae have been isolated from the natural environment, variations in the organism's abundance over time, and its relationship to aquatic microbiota are poorly understood. This is the first temporal study of this scale examining *Legionella* in the natural environment. The relative abundance of *Legionella* tended to be higher in pristine sites compared to those affected by agricultural activity. The relative abundance of Amoebozoa, some of which are natural hosts of legionellae, was similarity higher in pristine sites. Compared to other bacterial genera detected, *Legionella* had both the highest richness and the highest alpha diversity. Our findings indicate that a highly diverse population of legionellae may be found in a variety of natural aquatic sources. Further characterization of these diverse natural populations of *Legionella* could help inform prevention and control efforts aimed at reducing the risk of *Legionella* colonization of the built environment which could ultimately decrease the risk of human disease.

3.3.2. Introduction

Legionnaires' disease (LD) is a potentially fatal form of bacterial pneumonia caused by various species of *Legionella* (Mercante and Winchell, 2015). Individuals with chronic lung diseases or immune system deficiencies, as well as smokers and those of advanced age are at an increased risk for LD. A milder form of legionellosis, characterized by fever and "flu-like" symptoms, is termed Pontiac fever. In the built environment, *Legionella* can multiply in water that is stagnant, maintained at permissive temperatures (~25-37°C), and lacking appropriate disinfectant levels. Various devices

such as cooling towers, showerheads, fountains, and spas can aerosolize contaminated water. Inhalation of these aerosols by susceptible individuals can result in legionellosis.

Over 60 species of *Legionella* have been identified (http://www.bacterio.net/-allnamesdl.html) and at least a third of these have been linked to human disease (Muder and Yu, 2002). *L. pneumophila* is by far the most frequent cause of LD. Other less common, clinically relevent species include *L. longbeachae*, *L. bozemanii*, *L. micdadei*, and *L. dumoffii* (Amodeo et al., 2010; Fang et al., 1989; Reingold et al., 1984; Yu et al., 2002) *L. pneumophila* is highly diverse containing 17 known serogroups (Mercante and Winchell, 2015). Genome sequence analysis has revealed that much of the genetic diversity among isolates of *L. pneumophila* is driven by recombination (Sánchez-Busó et al., 2014).

During legionellosis outbreak investigations, *Legionella* isolates from potential environmental sources are compared with clinical isolates in an effort to support epidemiological associations. Various subtyping schemes have been used for this purpose such as pulsed-field gel electrophoresis, sequence based typing, and more recently, whole genome sequencing (David et al., 2016; Mercante and Winchell, 2015). Confirmation of the environmental source of *Legionella* may help to shorten the duration of an outbreak by focusing remediation efforts on a specific source and informing ongoing prevention strategies.

The mechanisms by which sources of *Legionella* from the natural environment colonize the built environment are poorly understood. However, it is likely that at least some of the *Legionella* present in source water for the built environment is derived from natural aquatic ecosystems such as rivers, streams, and lakes where *Legionella* have been shown to be widely distributed (Fliermans et al., 1981; Morris et al., 1979). Various studies have demonstrated a link between *Legionella* and various protozoa including amoebae such *Acanthamoeba* spp., *Naegleria* spp., and *Hartmannella* spp. (Barbaree et al., 1993; Rowbotham, 1980). Many of the molecular mechanisms that legionellae use for growth in amoebae apper to overlap with those for growth in human macrophages (Swanson and Hammer, 2000). Moreover, growth within amoebae not only amplifies the number of *Legionella,* but may also enhance bacterial virulence (Cirillo et al., 1994; Swanson and Hammer, 2000). Finally, *Legionella* spp. have been detected in biofilms,

which are considered to be a major reservoir of the organism in colonized man-made water systems (Declerck, 2010; Taylor et al., 2009).

Few studies have attempted a comprehensive analysis of the microbiome of natural aquatic environments. One of these, the Applied Metagenomics of the Watershed Microbiome, has been described in the introduction and section 3.1 of this thesis, as well as previously (Van Rossum et al., 2015). As mentioned earlier, it was a year-long study to understand the microbial community composition in watersheds of varying land-use. In this study, sites within agricultural, urban and protected watersheds were sampled monthly. Size fractionation methods were employed to generate templates for sequencing; 16S and 18S amplicon sequencing were conducted to quantify changes in the microbiome of these environments while shotgun metagenomic sequencing was conducted to understand community structure and function.

In order to better understand the ecology of *Legionella* and other freshwater bacterial pathogen genera in the natural aquatic environment, we evaluated this extensive dataset with the goals of quantifying their abundance among different watersheds. Due to the relatively high abundance and presence of *Legionella* in all watershed sampling sites, we examined *Legionella* in greater depth, determining the diversity of *Legionella* spp. present, and evaluating the role of amoeba on the presence of this bacterium in such environments. Understanding the presence and diversity of *Legionella* in these watersheds may help to improve our ability to control colonization of man-made water systems from natural water sources by these organisms.

3.3.3. Methods

Sampling sites and processing

Water sample collection and processing were as described in section 3.1.3.

Shotgun and amplicon sequencing

Both shotgun sequencing and 16S rRNA amplicon sequencing were performed as described in section 3.1.3. Amplicons targeting the 18S rRNA gene for the eukaryotic-sized fraction were generated as previously described (Uyaguari-Diaz et al., 2016).

Bioinformatic analysis

Quality control and generation of OTUs for the 16S sequences were performed as described in section 3.1.3. After Mothur assigned reads to OTUs, the data was exported in BIOM format (McDonald et al., 2012), and the OTUs were extracted that were assigned to *Legionella*.

Shotgun sequencing quality control was performed as described in section 3.1.3. Following quality control, sequence reads were run on Diamond version 0.8.36 (Buchfink et al., 2015) with default parameters against the NCBI nr database (downloaded February 21, 2017). These results were used as input for taxonomic classification by MEGAN version 6.6.7 (Huson et al., 2016) using default parameters, except for Min Support Percent which was set to 0. MEGAN6 was used as, this analysis was performed most recently, and MEGAN6 was recently updated and compatible with the most recent version of the NCBI nr database (which had additional *Legionella* genomes). The taxonomic classification files were then parsed to extract the assignments directly to *Legionella* at the genus level, or any specific *Legionella* species. Additionally, a custom BLAST database containing > 700 *Legionella* spp. *mip* nucleotide sequences was used to identify raw reads aligning to at least 100 nucleotides and an E value of $1x10^{-10}$ with this gene. BLAST using the NCBI nt database was also performed with the identified reads and those with better alignments to non-*Legionella* targets were excluded.

Analyses were performed in R (v3.2.3), and Shannon's diversity index was calculated using the vegan package (Oksanen et al., 2015). The Kruskal-Wallis with Dunn's test was used to compare the relative abundance of legionellae among the sites, with p-values adjusted using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The September 2012 sample from the AUP site was removed from analyses, as it had been previously noted as unusual and perhaps mislabeled during sample processing (Van Rossum et al., 2015).

Paired-end 18S read files were quality filtered and trimmed with Trim Galore v0.3.7 (available at: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) using a quality score cut-off of 25. Remaining read pairs with more than 10 overlapping nucleotides were joined using FLASH v1.2.11 (Magoč and Salzberg, 2011). The merged read pairs were used as input to QIIME v1.9.1 (Caporaso et al., 2010) with the

SILVA (release 128) 18S database (Yilmaz et al., 2014) used for chimera filtering and open reference OTU assignment.

Accession number

Sequence data are in the NCBI Sequence Read Archive under BioProject ID: 287840.

3.3.4. Results

Detection of non-Legionella freshwater bacterial pathogen genera

Examination of the 16S dataset for *Clostridium*, *Campylobacter*, *Vibrio*, *Escherichia*/*Shigella*, *Salmonella*, *Mycobacterium*, and *Leptospira* showed that there were no OTUs assigned to any of *Campylobacter*, *Vibrio*, *Salmonella,* or *Leptospira*, and *Clostridium, Escherichia*/*Shigella,* and *Mycobacterium* were found sporadically sampling sites (Appendix B).

Detection of Legionella over time at various sampling sites

Analysis of the 16S dataset revealed that *Legionella* spp. were found in all watersheds and sampling sites (Fig. 3.14). Overall, the relative abundances of *Legionella* spp. were ≤ 2.1% of the bacterial taxa present and were significantly (q-value < 0.05, Table 3.6) higher in samples from more pristine sites (PUP and AUP) compared to sites affected by agricultural activity, while the sites affected by urban activity had intermediate levels of *Legionell*a abundance. The abundance of *Legionella* spp. varied substantially over the sampling period (Fig. 3.15).

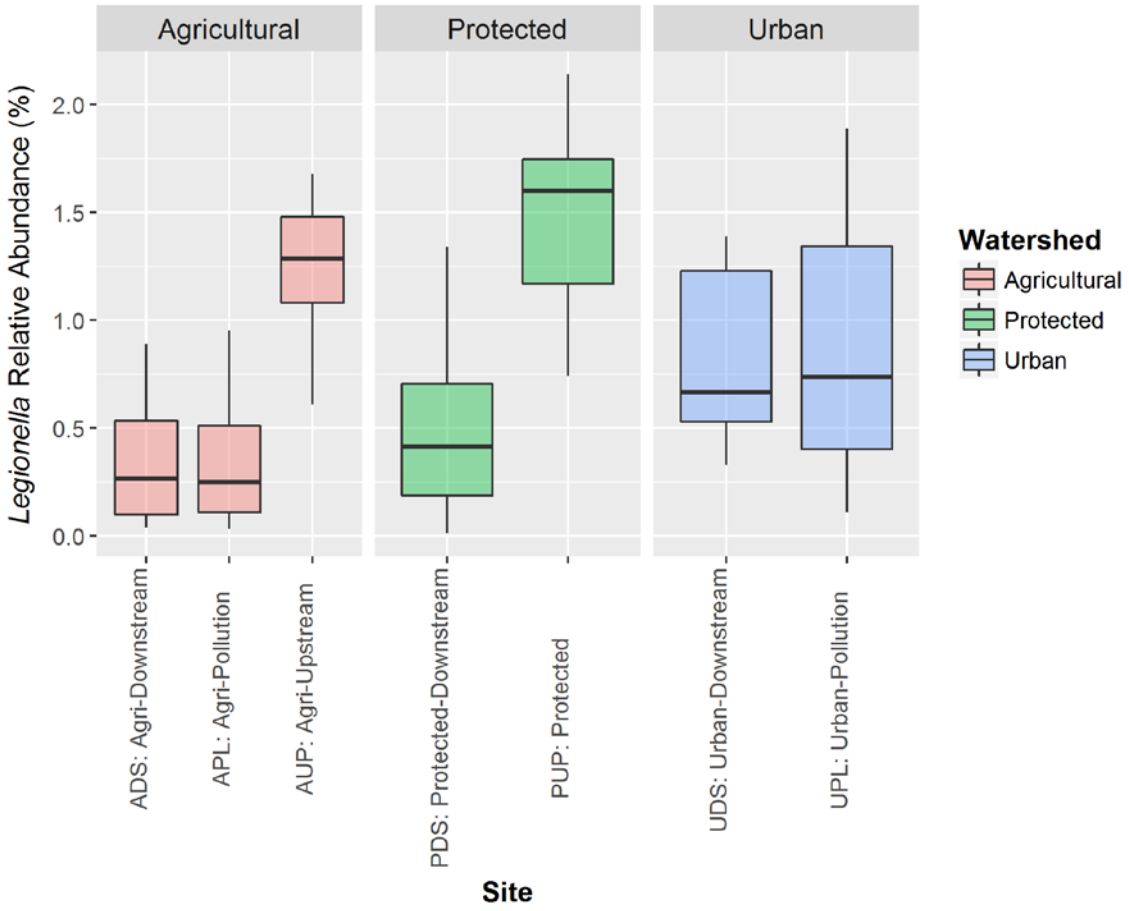

Figure 3.14 Abundance of *Legionella* at various sites.
Boxplots represent the relative abundances of *Legionella* present for all samples obtained from the indicated site determined by 16S rRNA sequencing. Boxes are bounded by the 25th and 75th percentile and the middle line represents the median relative abundance.

Table 3.6 Dunn's test results comparing relative abundance of *Legionella* based on 16S rRNA gene-sequencing data between watershed sites

Site 1	Site 2	q-value
APL: Agri-Pollution	ADS: Agri-Downstream	0.4976
APL: Agri-Pollution	AUP: Agri-Upstream	0.0004*
APL: Agri-Pollution	PDS: Protected-Downstream	0.2694
APL: Agri-Pollution	PUP: Protected	0.0000*
APL: Agri-Pollution	UDS: Urban-Downstream	0.0226*
APL: Agri-Pollution	UPL: Urban-Pollution	0.0237*
ADS: Agri-Downstream	AUP: Agri-Upstream	0.0004*
ADS: Agri-Downstream	PDS: Protected-Downstream	0.2621
ADS: Agri-Downstream	PUP: Protected	0.0000*
ADS: Agri-Downstream	UDS: Urban-Downstream	0.0209*
ADS: Agri-Downstream	UPL: Urban-Pollution	0.0218*
AUP: Agri-Upstream	PDS: Protected-Downstream	0.0045*
AUP: Agri-Upstream	PUP: Protected	0.1826
AUP: Agri-Upstream	UDS: Urban-Downstream	0.1350
AUP: Agri-Upstream	UPL: Urban-Pollution	0.1092
PDS: Protected-Downstream	PUP: Protected	0.0002*
PDS: Protected-Downstream	UDS: Urban-Downstream	0.0869
PDS: Protected-Downstream	UPL: Urban-Pollution	0.0828
PUP: Protected	UDS: Urban-Downstream	0.0207*
PUP: Protected	UPL: Urban-Pollution	0.0206*
UPL: Urban-Pollution	UDS: Urban-Downstream	0.4885

* Indicates a q-value < 0.05

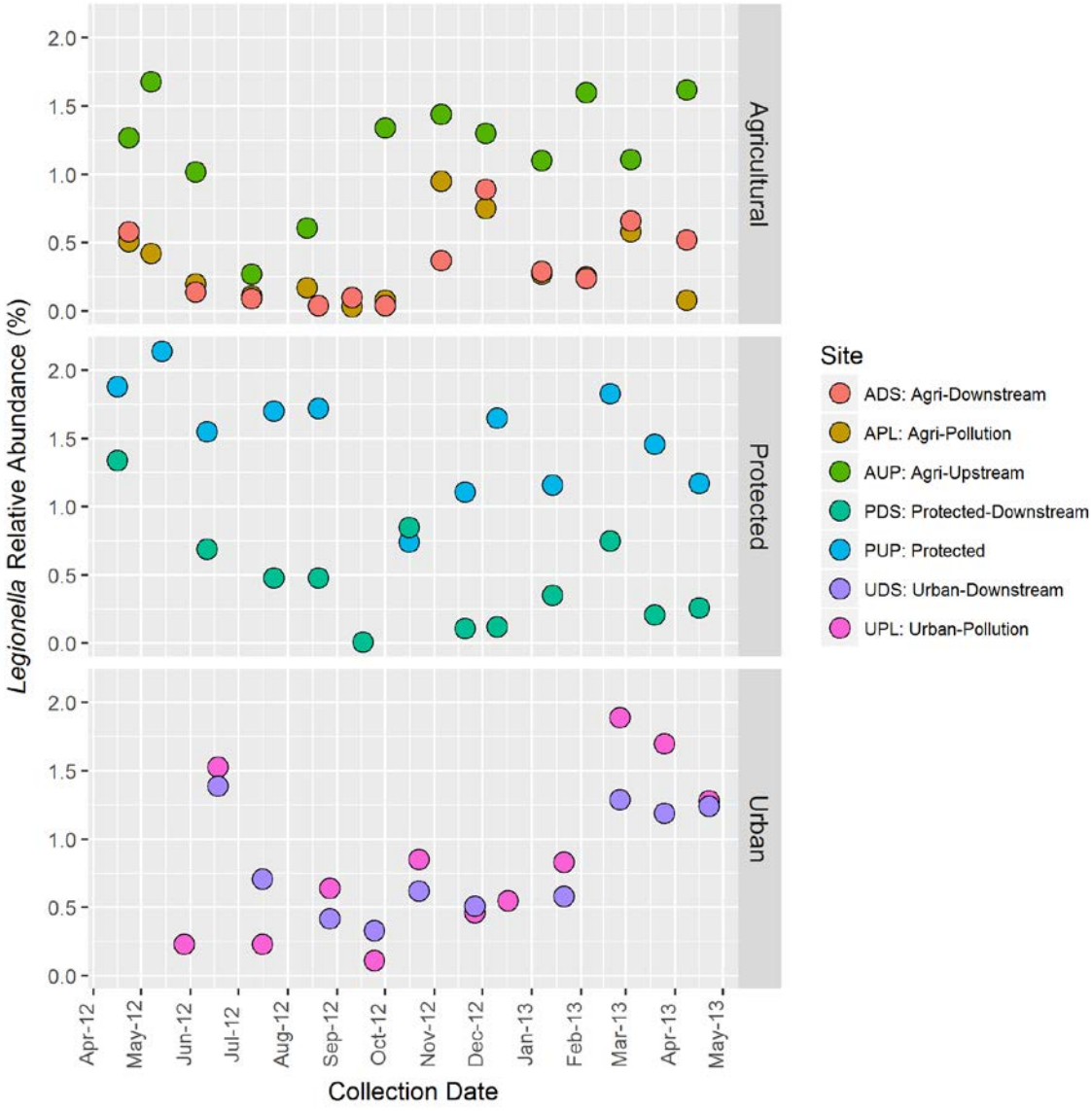

Figure 3.15 Relative abundance of *Legionella* derived from 16S amplicon analysis across sampling sites and date.

The relative abundance of *Legionella* varies substantially over time, and from site to site.

Distribution of Legionella species among sampling sites

To examine the distribution of *Legionella* spp. at a higher resolution than the genus level, the shotgun metagenomics dataset was used to obtain species-level classification. The shotgun metagenomics reads were classified to over 40 *Legionella* species with MEGAN6 albeit with very few reads associated with some species (Fig 3.16). While most reads were assigned only at the *Legionella* genus-level, the most abundant species was *L. pneumophila* which was found at all sampling sites.

In addition, we used a BLAST approach to identify sequence reads from the shotgun metagenomics dataset matching the *mip* gene which is used extensively for *Legionella* species determination (Ratcliff et al., 1998). This analysis also found a low number of reads matching over 35 different *Legionella* spp. among the samples examined (Appendix B). Notably, the range of nucleotide identity among the reads aligned with known *mip* genes varied widely (~68-94%).

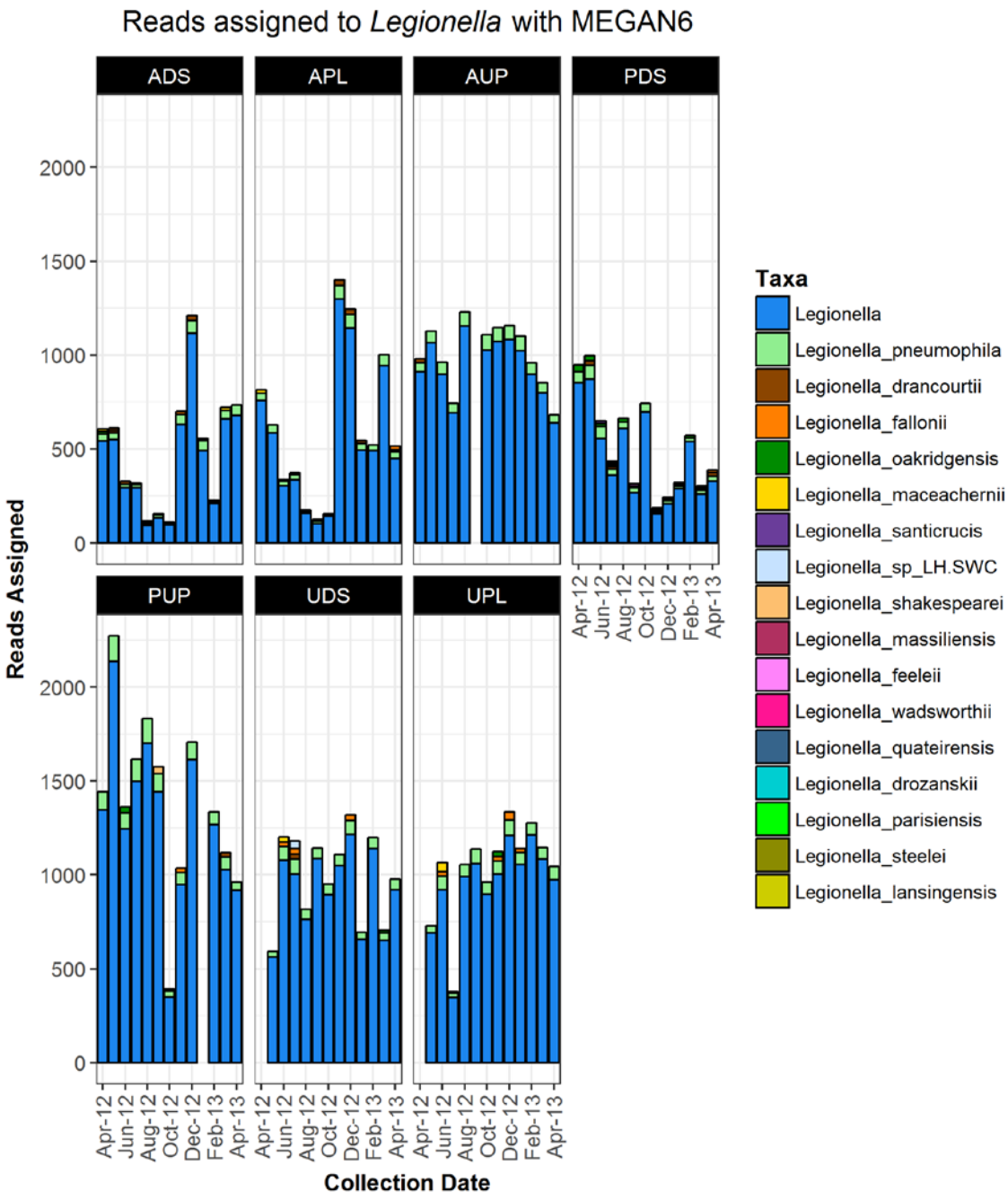

Figure 3.16 Taxonomic classification *Legionella* metagenome shotgun sequencing reads.

The number of metagenomic sequencing reads classified by MEGAN6 analysis to *Legionella* are shown for each sampling site and date of collection. Reads assigned to *Legionella* spp. for which the species-level assignment comprised fewer than 2% of the total reads assigned to *Legionella* are denoted as “Legionella” in the legend (i.e. “Legionella” denotes the sum of all reads assigned to the genus *Legionella* and all *Legionella* species not depicted separately). Species in the legend are ordered from most abundant to least abundant in the overall dataset (all samples).

Diversity of Legionella relative to other bacterial taxa

To better understand the diversity of *Legionella* spp. present in the samples, the 16S rRNA datasets were examined at the OTU level. OTUs approximate microbial taxa, but are not limited to previously sequenced taxa (as in the MEGAN6 analysis), providing a useful complement to the species-level classification of the metagenomics data. A total of 71 OTUs were assigned to *Legionella* supporting the view of a highly diverse population of *Legionella* spp. present in these samples (Fig. 3.17). *Legionella* is the 9th most abundant genus in the dataset, so to rule out the possibility of the richness being driven by relative abundance, the richness of *Legionella* was compared to other abundant genera. The top 50 most abundant genera in the dataset over all sampling sites, based on the sum of all reads assigned to the genera from all samples, were plotted versus the richness or number of OTUs assigned to that genus (Fig. 3.18). Richness was unevenly distributed among the taxa and does not seem to be driven by relative abundance. Most genera have a richness of fewer than 10 OTUs, and only 5 genera have a richness greater than 20 OTUs.

Legionella had markedly higher richness than the mean (8.1 OTUs) and median (2 OTUs) richness observed among the top 50 most abundant genera. This high richness of *Legionella* was even more pronounced when compared to the entire dataset (versus just the top 50), which had a mean richness of 3.2 and a median of 1. The richness of OTUs assigned to *Legionella* was distributed across sample sites and time (Fig. 3.19). Finally, average alpha diversity over all samples of the 50 most abundant genera was also plotted, showing that *Legionella* is not only the richest but also the most diverse bacterial genus identified (Fig. 3.20).

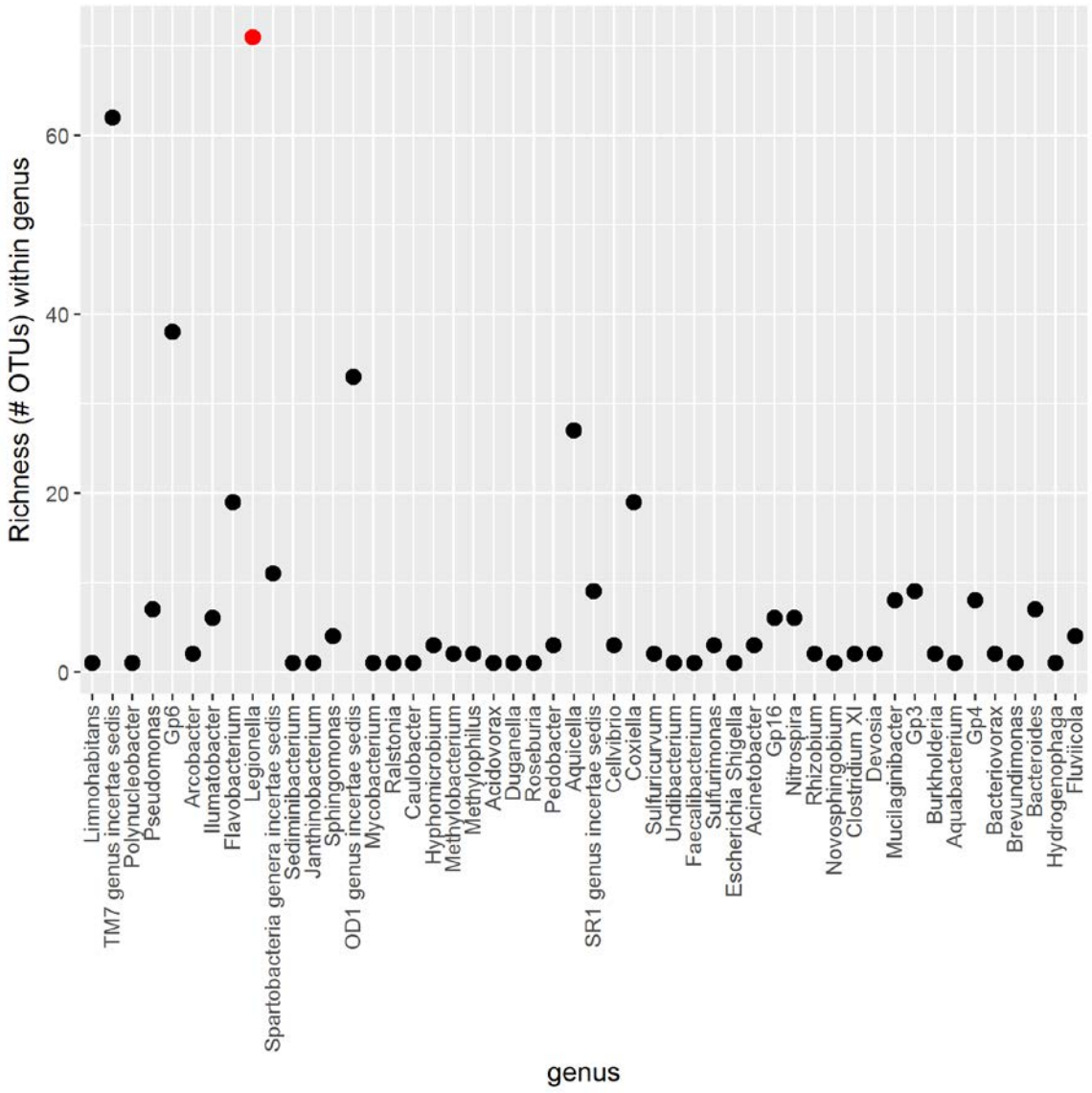

Figure 3.17 Richness (number of OTUs) for the top 50 most abundant genera in the watershed dataset.

Genera are plotted in order from most abundant (*Limnohabitans*) to 50th most abundant (*Fluviicola*). *Legionella* has the highest richness. *Legionella* is shown in red.

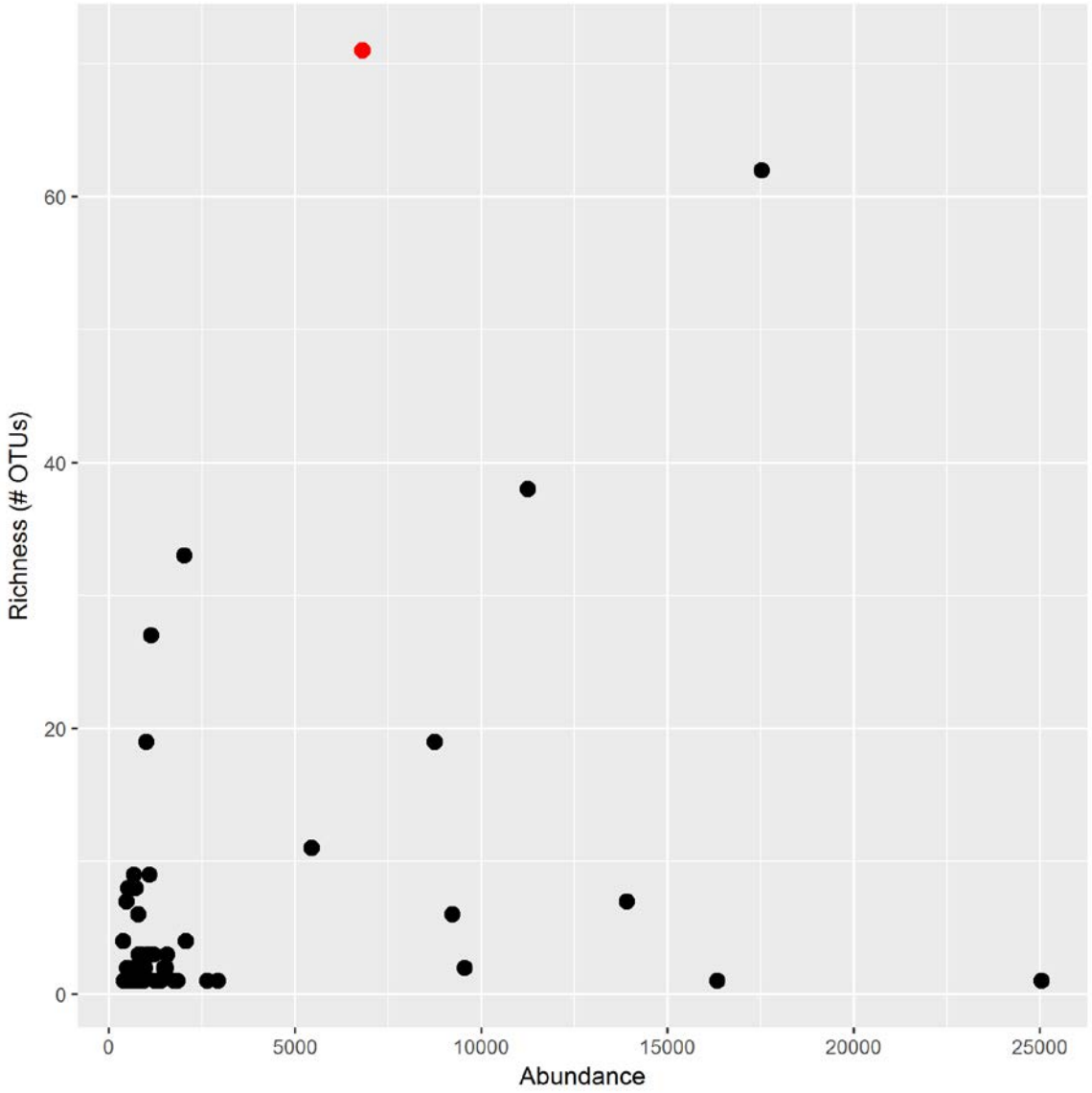

Figure 3.18 Richness versus abundance of top 50 most abundant genera.
Legionella has the highest richness (71), shown in red, which is substantially higher than the mean (8.1) or median (2). Abundance is the sum of reads assigned to the genus across all samples.

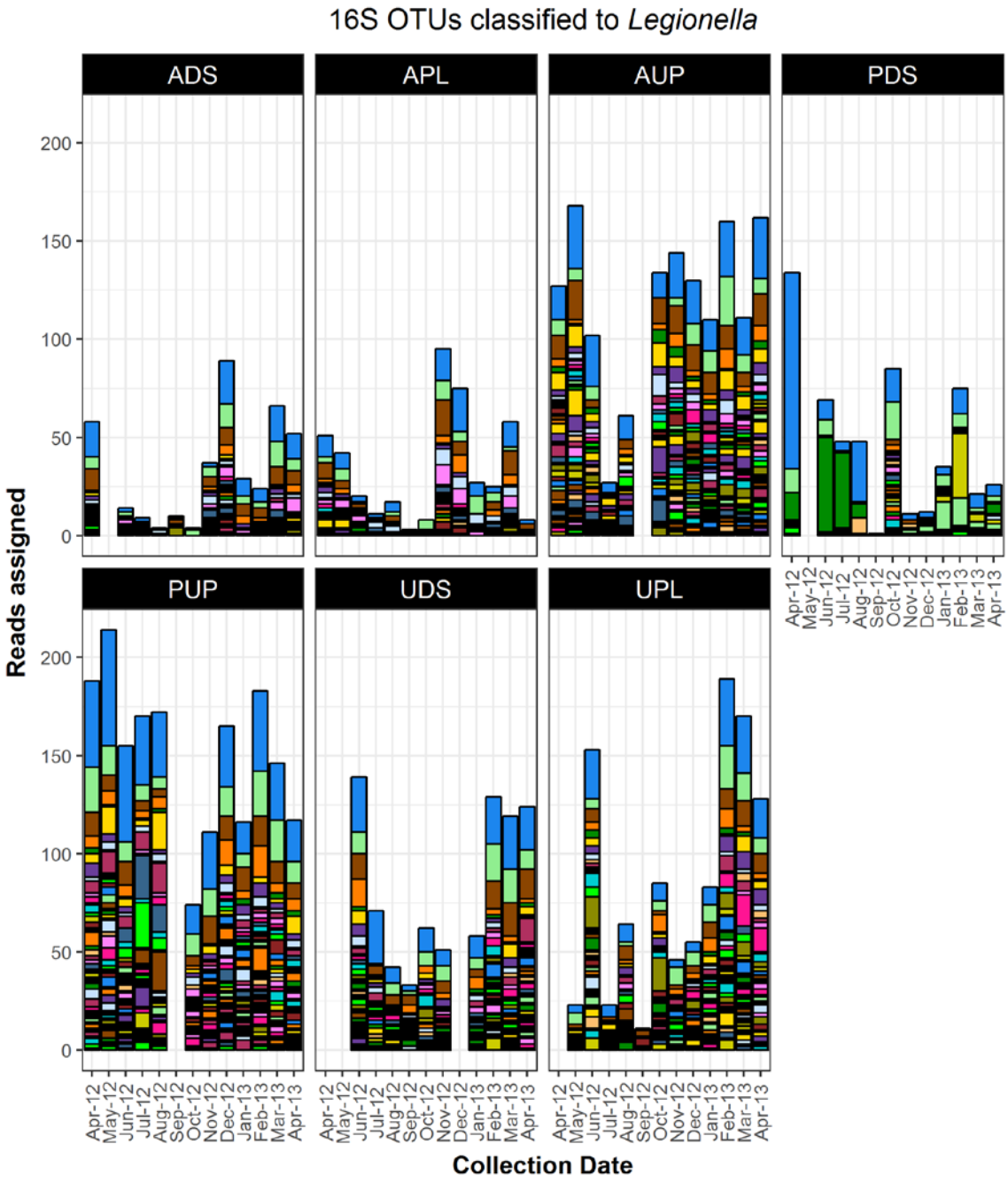

Figure 3.19 Distribution of *Legionella* OTUs derived from 16S rRNA amplicon analysis across sampling sites and date.

There are many OTUs assigned to *Legionella*, and the spread of abundance is relatively even. The number of 16S rRNA amplicon sequence reads assigned to Legionella OTUs are denoted along the y-axis. Each OTU is denoted by a different colored shading.

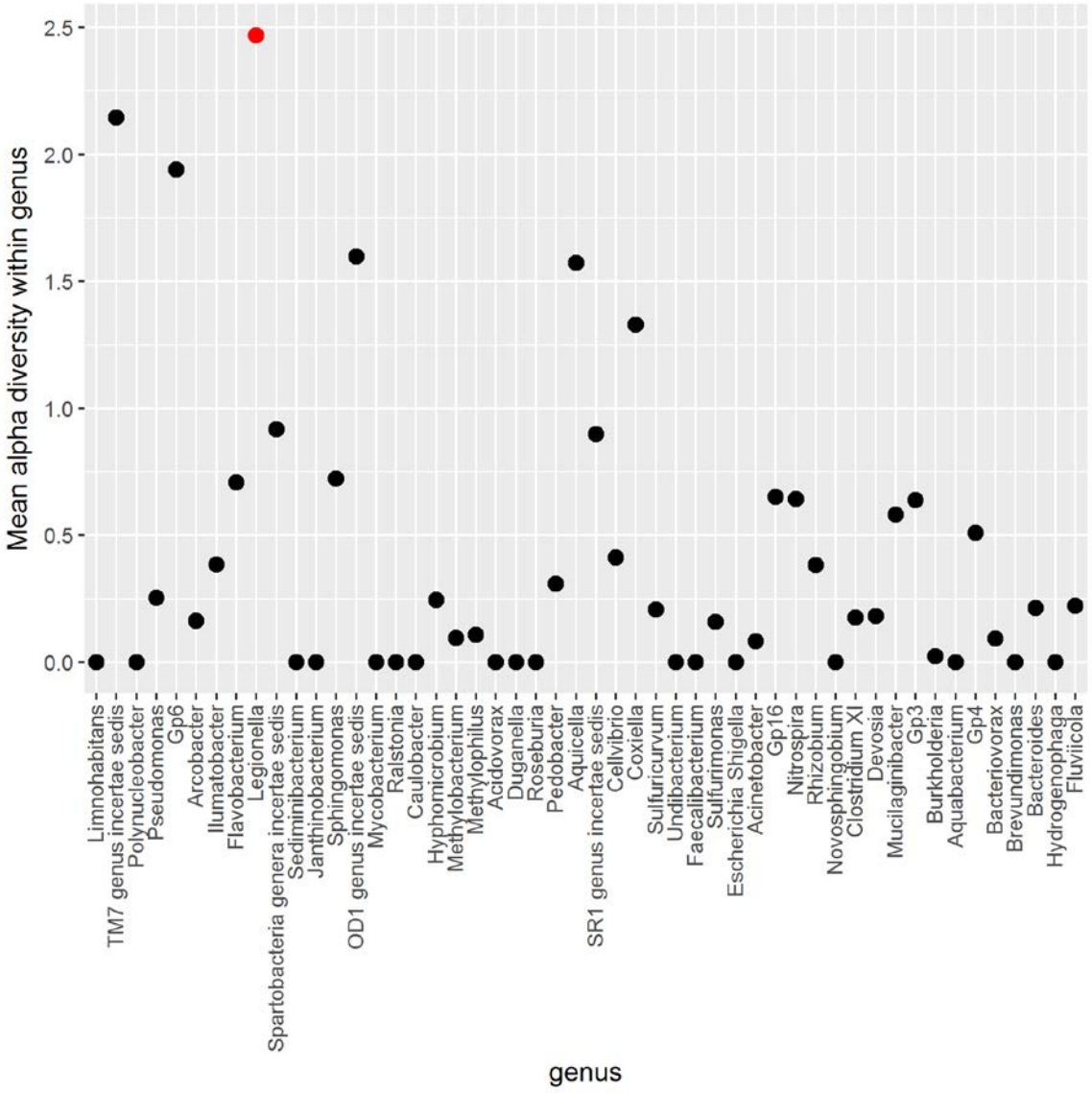

Figure 3.20 Mean alpha diversity for the top 50 most abundant genera in the watershed dataset.

Genera are plotted in order from most abundant (*Limnohabitans*) to 50th most abundant (*Fluviicola*). Alpha diversity was calculated based on the OTUs which were assigned to each genus. *Legionella* is shown in red.

Comparison of *Legionella* and Amoebozoa abundances

Analysis of the 18S rRNA dataset revealed a wide distribution in the abundance of the phylum Amoebozoa among the sample sites (Fig. 3.21). A statistically significant ($q < 0.05$) difference was observed between the ADS site and the PUP and UDS sites (Table 3.7). The more pristine sites within the agricultural (AUP) and protected (PUP) watersheds displayed a wider distribution of Amoebozoa abundances than the more

impacted sites (ADS and PDS). The most pronounced seasonal effect was observed in the UPL site where samples collected between July and September had the highest abundance of Amoebozoa.

Platyamoeba was the most abundant (mean abundance over all samples ~0.3%) genus represented among the Amoebozoa (data not shown). *Acanthamoeba*, a well-known host of *Legionella*, had a mean abundance nearly 100-fold lower than *Platyamoeba*. *Naegleria*, an amoebal member of a separate phylum which can also host *Legionella,* was not detected in any samples.

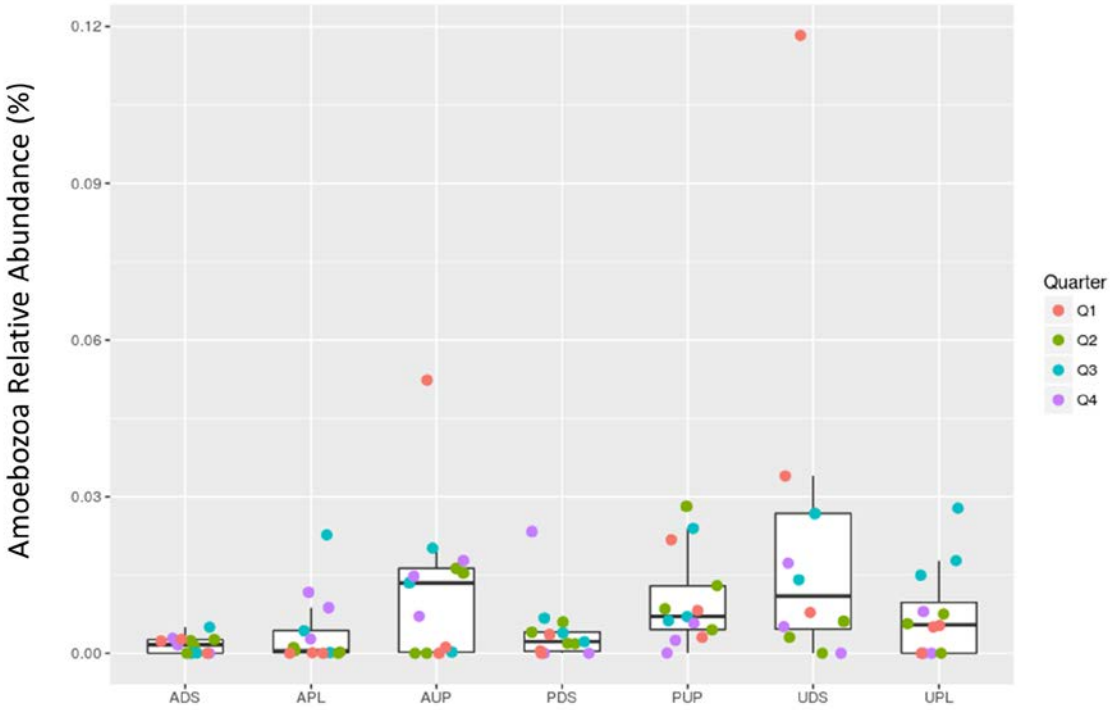

Figure 3.21 Relative abundance of Amoebozoa at watershed sites.
Boxplots represent the relative abundances of Amoebozoa present for all samples obtained from the indicated site determined by 18S rRNA sequencing. Each circle represents the relative abundance of Amoebozoa in a specific sample. Circles are color coded by the date of sample collection where quarters of the year are as follows: Q1, January, February, March; Q2, April, May, June; Q3, July, August, September; and Q4, October, November, December. Boxes are bounded by the 25th and 75th percentile and the middle line represents the median relative abundance.

Table 3.7 Dunn's test results comparing relative abundance of Amoebozoa based on 18S rRNA gene-sequencing data between watershed sites

Site 1	Site 2	q-value
APL: Agri-Pollution	ADS: Agri-Downstream	0.2895
APL: Agri-Pollution	AUP: Agri-Upstream	0.1124
APL: Agri-Pollution	PDS: Protected-Downstream	0.4331
APL: Agri-Pollution	PUP: Protected	0.0411*
APL: Agri-Pollution	UDS: Urban-Downstream	0.0530
APL: Agri-Pollution	UPL: Urban-Pollution	0.2303
ADS: Agri-Downstream	AUP: Agri-Upstream	0.0445*
ADS: Agri-Downstream	PDS: Protected-Downstream	0.2582
ADS: Agri-Downstream	PUP: Protected	0.0117*
ADS: Agri-Downstream	UDS: Urban-Downstream	0.0178*
ADS: Agri-Downstream	UPL: Urban-Pollution	0.0988
AUP: Agri-Upstream	PDS: Protected-Downstream	0.1569
AUP: Agri-Upstream	PUP: Protected	0.2929
AUP: Agri-Upstream	UDS: Urban-Downstream	0.2782
AUP: Agri-Upstream	UPL: Urban-Pollution	0.3178
PDS: Protected-Downstream	PUP: Protected	0.0519
PDS: Protected-Downstream	UDS: Urban-Downstream	0.0472*
PDS: Protected-Downstream	UPL: Urban-Pollution	0.2759
PUP: Protected	UDS: Urban-Downstream	0.4435
PUP: Protected	UPL: Urban-Pollution	0.1952
UPL: Urban-Pollution	UDS: Urban-Downstream	0.1740

* Indicates a q-value < 0.05.

3.3.5. Discussion

The robust methodology described by Uyaguari-Diaz et al. (2016) to separate various microbial components (eukaryotic, bacterial and viral) in natural water samples using both amplicon (16S and 18S) as well as metagenomic sequencing was used to characterize the composition of water samples from various watersheds in British Columbia. Analysis of the data collected from this year-long study of BC watersheds showed a seasonal shift in the microbial composition of some of the sites when looking at the family level (section 3.1) or k-mer composition (Van Rossum et al., 2015). Here,

we examined the dataset for the presence of freshwater bacterial pathogen genera, and among other findings, detected *Legionella* at all sampling sites and collection dates.

Unlike the seasonal patterns observed when examining the entire bacterial community composition in section 3.1, the relative abundance of *Legionella* varied considerably throughout the year, without any discernible seasonal patterns, reaching as high as 2% of the bacterial taxa present. Notably, these values do not provide a measure of absolute abundance of the organism in these samples. Other researchers have detected *Legionella* using culture or quantitative PCR in natural water sources. In a recent study, investigators found ~10^4 - 10^5 cells/L of *Legionella* spp. in some Taiwanese river water samples using real-time PCR (Kao et al., 2013), while a previous study of marine and freshwater sites in Puerto Rico demonstrated an abundance of *L. pneumophila* of 10^4 cells/mL using direct florescence-antibody (DFA) testing (Ortiz-Roque and Hazen, 1987). These findings suggest that the quantity of *Legionella* in the natural environment may be highly variable.

The relative abundance of *Legionella* in our study was highest in sites with limited land use: from the site upstream of agricultural activity (AUP) and from a river that empties into a drinking water reservoir (PUP). In both cases, these sources feed downstream sites where the abundance of *Legionella* is lower. There are several possible explanations for this decrease in *Legionella* abundance. Downstream sites may contain contaminants or lack specific nutrients for *Legionella* growth. There may also be an increase in certain non-*Legionella* genera in these downstream sites, resulting in a lower relative abundance of *Legionella* in the community. Alternatively, *Legionella* may become associated with biofilms, decreasing its abundance in the surface water collected in this study. Notably, the water collected at the PDS site travels [from the protected watershed (PUP)] through a nearly 9 km pipe where biofilm may be present, trapping *Legionella*. *Legionella* can also survive and grow within various amoeba species. Similar to the pattern observed with *Legionella* abundances, the lowest mean abundances of Amoebozoa were found at the ADS, APL, and PDS sites, supporting the possibility that *Legionella* abundance is amplified by the presence of amoeba in these natural water sources.

This study revealed that *Legionella* spp. present in the examined watersheds are incredibly diverse. More than 70 OTUs were detected using 16S amplicon sequencing.

The metagenomic sequence analysis used in this study demonstrated that *L. pneumophila* was the most common species represented. Similarly, Fliermans et al. (1981) detected *L. pneumophila* by DFA in nearly all concentrated water samples collected from 67 natural water sources in North Carolina, South Carolina, Georgia, Florida, Alabama, Indiana, and Illinois (Fliermans et al., 1981). Various *Legionella* spp. are frequently detected in studies of natural water sources (Carvalho et al., 2007; Kao et al., 2013; Ortiz-Roque and Hazen, 1987; Parthuisot et al., 2010). Notably, sequence analysis of the most common *Legionella* 16S rRNA-based OTU amplified from water samples along a Brazilian river were associated with unknown/uncultured bacteria (Carvalho et al., 2007). The high diversity of *Legionella* spp. among these sources may have implications for clinical disease since several non-*pneumophila Legionella* spp. are associated with clinical disease (including pneumonia) especially among immunosuppressed populations (Muder and Yu, 2002). A study of natural water sources in the Mount Saint Helens (Washington, USA) blast zone was conducted after researchers exposed to lakes and streams in the region reported symptoms consistent with Pontiac fever in the early 1980's (Tison et al., 1983). Various known *Legionella* spp. were detected in the Saint Helens study with higher organism abundances found in water samples taken within the blast zone and lakes receiving water from hydrothermal seeps compared to those sites outside the blast zone. A novel species (*L. sainthelensi*) was isolated from water samples collected around Mount Saint Helens (Campbell et al., 1984) and this species was subsequently found to be associated with clinical disease (Benson et al., 1990).

Metagenomics classification programs such as MEGAN6 used in the current study may over-classify reads to incorrect species if the matching species is not present in the database (Peabody et al., 2015). More specifically, the program might assign reads to the most closely related species in the database. There are currently over 500 *L. pneumophila* genome sequences in the NCBI database but only a few representatives are present for other *Legionella* spp. A wide range of alignment identities observed with the MEGAN6 analysis further suggests that novel or uncharacterized *Legionella* may be present in the samples. Notably, alignment of the shotgun metagenomic reads with the *Legionella mip* gene also uncovered a large number of *Legionella* spp. (>35) among the watershed samples but sequencing coverage of this gene may be limited. Nonetheless,

the alignment identity of these matches was low (typically < 80%) further suggesting the presence of additional novel *Legionella* species in these watersheds.

While the presence of *Legionella* spp. in natural water samples is not on its own a significant public health concern, these organisms may seed man-made water systems. In turn, these systems could become sources of *Legionella* dissemination under permissive conditions. Understanding the diversity of organisms present in the natural aquatic environment and factors that may contribute to increased abundance of specific *Legionella* spp. in these environments may help public health workers identify potential new threats to human health and respond quickly to Legionnaires' Disease (LD) using improved diagnostic and typing assays. This study demonstrates that natural aquatic environments including watersheds likely harbor previously unrecognized *Legionella* spp. As culture-independent diagnostic tests for LD become more commonly utilized, it is important to evaluate the ability of these assays to detect new and emerging *Legionella* spp. and assess their potential to cause disease.

Chapter 4.

PSORTm: PSORTb for metagenomics datasets

Chapter 4 introduces a version of PSORTb, a bacterial and archaeal subcellular localization predictor, that works with metagenomics datasets. The development and evaluation of this software is described, followed by applying the software to the analysis of the watershed samples from the Applied Metagenomics of the Watershed Microbiome project.

I completed all work presented in this chapter with the following exceptions: Gemma Hoad assisted in the development of PSORTm, including setting up the website and Docker version, and Thea Van Rossum assisted in the statistical analysis of differential functional categories.

4.1. Abstract

Although many methods for microbial protein subcellular localization (SCL) prediction exist, there are not currently any readily available that are designed to work with metagenomics sequence data, despite the interest from researchers studying microbial communities (for example, for development of ELISA-based diagnostics). Thus, PSORTb, one of the most precise bacterial and archaeal subcellular localization predictors, was modified to accommodate metagenomics sequences, and the utility of this modified PSORTb (PSORTm) was demonstrated through an analysis of freshwater samples collected monthly from a pristine, an urban, and an agricultural watershed over a one-year period. An evaluation using 5-fold cross validation with *in silico* fragmented sequences showed that PSORTm maintains high precision, and that sensitivity increases along with increased input sequence fragment length. Running PSORTm on the watershed samples revealed the importance of normalization, and showed that taxonomic profiles derived from the subset of sequences predicted to be exposed (extracellular, outer membrane, or cell wall) are similar to those derived from the full set of sequences. It also identified potential protein function categories that could act as biomarkers of water quality. PSORTm has many potential applications, such as in the identification of cell-surface based biomarkers for protein-based diagnostic tests, and may be useful in a wide range of studies examining microbial communities for medical, environmental, and agricultural applications.

4.2. Introduction

Ever since PSORTb was first introduced in 2003, it has remained one of the most precise subcellular localization predictors available (Gardy et al., 2003, 2005; Yu et al., 2010b). The initial version of PSORTb predicted the subcellular localization of Gram-negative bacteria. PSORTb 2.0 extended the capabilities of PSORTb to allow prediction of both Gram-negative and Gram-positive bacteria. The most recent version of PSORTb, PSORTb 3.0, generates predictions for all the types of prokaryotic cell structures that are known: Archaea, traditional Gram-positive (Gram-positive without an outer membrane), traditional Gram-negative (Gram-negative with an outer membrane), Gram-positive with an outer membrane, and Gram-negative without an outer membrane. PSORTb 3.0 added new localization subcategories (host-associated, type III secretion, fimbrial,

flagellar, and spore), and was the first SCL predictor to do so. Before version 3, PSORTb had previously been limited to predictions of the major cellular compartments: cytoplasmic, inner membrane, periplasmic, outer membrane, and extracellular in Gram-negative bacteria, and cytoplasmic, cytoplasmic membrane, cell wall, extracellular in Gram-positive bacteria (and Archaea).

In addition to PSORTb, there are a variety of other SCL prediction tools that have been developed. A distinction can be made between tools that perform specialized predictions of one of a few SCLs, such as tools that make predictions for whether proteins are secreted or not based on the presence of specific features, or tools that are more general and make predictions to multiple categories covering all possible subcellular localizations. For example, SignalP (Petersen et al., 2011), PRED-SIGNAL (Bagos et al., 2009), and CW-PRED (Litou et al., 2008) detect for the presence of specific features such as signal peptides or cell wall sorting signals. Other programs are broader in scope, predicting various properties of proteins including SCL, such as the web server ProteomeAnalyst (Szafron et al., 2004) which predicts GO molecular function and subcellular localization of proteins within a proteome using Naïve Bayes classifiers. Some methods are hybrid based approaches which combine two different classifiers to improve performance: these methods include PSLpred, which relies on PSI-BLAST and SVM modules; PSL101, which combines a one-versus-one SVM model and a structural homology approach; and PSLDoc, which combines SVM classifiers with probabilistic latent semantic analysis (Bhasin et al., 2005; Chang et al., 2008; Su et al., 2007). Other predictors take this one step further, combining three or more classifiers such as PSORTb, SubcellPredict, and HensBC (Bulashevska and Eils, 2006; Niu et al., 2008; Yu et al., 2010b). MetaLocGramN is a meta-predictor, combining the predictions of four other methods, some of which themselves take an ensemble approach (Magnus et al., 2012).

Other methods for subcellular localization distinguish themselves through other characteristics. Several methods are able to deal with multiple location proteins; some examples are Gneg-mPLoc (Shen and Chou, 2010) and Gpos-mPLoc (Shen and Chou, 2009), which perform predictions through an ensemble approach on a training dataset with reduced redundancy. These methods were improved in Gpos-ECC-mPLoc and Gneg-ECC-mPLoc (Wang et al., 2015), which are multi-label predictors using an ensemble of classifier chains. There are other methods that focus exclusively on

predicting subcellular localization for a specific taxa, such as TBPred (Rashid et al., 2007) for predicting subcellular location of mycobacterial proteins, and FGsub (Sun et al., 2010), which although for a particular fungal pathogen rather than bacteria, predicts the subcellular localization of *Fusarium graminearum* proteins.

Although PSORTb makes predictions for all prokaryotes, most tools are specific for certain types of microbes; for example, it is common for tools to be developed for either Gram-positive or Gram-negative bacteria. Methods that are specific for Gram-positive bacteria include Locate P (Zhou et al., 2008), an unnamed method that takes a secretomics-based strategy and utilizes decision trees (Renier et al., 2012), and Augur (Billion et al., 2006). Many others, such as SOSUI-GramN (Imai et al., 2008) and CELLO (Yu et al., 2004, 2006), are specific for Gram-negative bacteria.

Computational prediction is a relatively rapid and inexpensive alternative to experimental methods for microbial protein SCL. Determination of the SCL of a protein aids in identification of protein function and annotation of genomes. Furthermore, there is interest in the identification of cell surface/secreted proteins for applications such as the development of ELISA-based tests or identification of drug or vaccine targets. Despite the number of bacterial SCL prediction methods that have been developed, there is a notable lack of methods designed specifically to work with metagenomics sequences. Although one method exists, MetaP (Luo et al., 2009), it assumes all sequences are from Gram-negative organisms, and is not made readily available for online use or download except for potentially through contact with the authors. Thus, PSORTb, a precise bacterial and archaeal subcellular localization program, was modified to enable the classification of metagenomics sequences. This predictor is called PSORTm, and because it is a complicated package, a Docker image has been made available (available at https://hub.docker.com/r/brinkmanlab/psortm/); we are also developing an online version. PSORTm is shown to maintain high precision, while increasing sensitivity as input fragment lengths get larger. Finally, PSORTm was applied to the analysis of the watershed samples, providing insights into how PSORTm works on a real metagenomics dataset, and identifying potential gene functional category biomarkers of water quality.

4.3. Methods

4.3.1. Implementation changes to software

PSORTm is a modified version of PSORTb 3 (Yu et al., 2010b). The different modules used in PSORTb 3, and whether they are incorporated and/or modified in PSORTm, are listed in Table 4.1. All of the modules are used except for the signal peptide module, which in PSORTb 3 returns a prediction of non-cytoplasmic if an N-terminal signal peptide is identified. However, with predicted protein sequences derived from metagenomics sequences, the predicted sequence may start anywhere within the protein, so the first amino acids within the sequence are not necessarily the N-terminal amino acids, hence the module would not be effective and was not incorporated. All of the other modules were incorporated, and the SCL-BLAST module was the only incorporated module modified. In PSORTb 3, the SCL-BLAST module had a length restriction such that the query sequence must be within 80-120% of the length of the subject protein, which was in place to reduce errors due to the domain nature of proteins. This would have been too restrictive for metagenomics fragments, many of which would not meet the 80% length cut-off, so the length restriction was removed.

Table 4.1 List of modules used in PSORTb 3, and whether they were incorporated or modified in PSORTm

Module	Features used for prediction	SCLs predicted	Incorporated	Modified
Signal peptide prediction	N-terminal signal peptide	Non-cytoplasmic	No	-
SVMs	Frequent subsequences within protein sequences	All SCLs	Yes	No
ModHMM	Transmembrane α-helices	CM	Yes	No
PROSITE motifs	Motifs associated with specific SCLs	All SCLs	Yes	No
PROFILE motifs	Motifs associated with specific SCLs	All SCLs	Yes	No
Outer membrane motifs	Motifs associated with β-barrel OM proteins	OM	Yes	No
SCL-BLAST	Homology	All SCLs	Yes	Yes

An additional tool was also incorporated into PSORTm to sort input sequences according to type of organism and cell envelope. PSORTb 3 has the following 5 input options: Archaea, Gram-negative, Gram-positive, Gram-negative without an outer membrane, and Gram-positive with an outer membrane. This tool requires the user to provide an input file of reads along with their associated taxonomic classification, and will provide output files of the reads sorted into the 5 aforementioned organism/cell envelope categories, as well as an additional file of reads that could not be categorized (for example, the reads that the taxonomic classification program used was not able to classify). The categorization scheme is borrowed from PSORTdb, which uses taxonomy (along with an OMP85 marker) to categorize newly sequenced bacterial and archaeal genomes into organism/cell envelope categories so that the appropriate input option for PSORTb could be chosen.

4.3.2. Software evaluation

Five-fold cross validation was performed on the Gram-positive, Gram-negative, and archaeal datasets that were used to evaluate PSORTb 3. This dataset present in ePSORTdb was previously published (Yu et al., 2011), and consists of protein subcellular localization data obtained from Swiss-Prot version 49, as well as SCL data manually retrieved from the literature, the EcoSal database, and the *Pseudomonas* Genome Database (Winsor et al., 2016). In total, the dataset is comprised of 8230 Gram-negative proteins, 2652 Gram-positive proteins, and 810 archaeal proteins. To simulate metagenomics fragments, in the five-fold cross validation the test sequences were randomly fragmented *in silico* from lengths 60 to 450 in increments of 30. For each of these fragment lengths, fragments were generated 10 times.

4.3.3. Analysis of watershed microbiomes

Watershed sample collection and processing, and metagenomics sequencing and quality control of the sequence reads, were performed as described in section 3.1.3. Following quality control, reads were taxonomically classified by first being aligned to the nr database with RAPSearch2 (Zhao et al., 2012), followed by classification with DiScRIBinATE (Ghosh et al., 2010), as described in section 3.1.3. Metagenomics gene prediction was also run on the same reads to produce putative amino acid sequences using MetaProdigal with default parameters (Hyatt et al., 2012). Together, the files of

taxonomic classification from DiScRIBinATE and predicted amino acid sequences from MetaProdigal were then run on PSORTm to obtain predicted SCLs.

Functional classification of sequence reads and comparisons were performed as described previously (Van Rossum et al., 2015): shotgun metagenomic reads were functionally classified to SEED subsystems (Overbeek et al., 2005) using MEGAN 5.10 (Huson et al., 2011) following alignment to the nr database with RAPSearch2 (Zhao et al., 2012), gene functional category profiles were normalized by average genome size using MicrobeCensus (Nayfach and Pollard, 2015), and the Wilcox test was used for analyzing differential abundance of gene group profiles after removing low abundance features (mean abundance <0.01% in all samples).

4.4. Results

4.4.1. Five-fold cross validation

Five-fold cross validation results comparing the latest unmodified version of PSORTb (PSORTb 3.0.2) versus PSORTm (with the signal peptide module removed and the SCL-BLAST module modified) are shown in Figure 4.1 (sensitivity) and Figure 4.2 (precision). PSORTm shows substantially higher sensitivity than PSORTb 3.0.2, which data suggests is due to the removal of the length restriction on the SCL-BLAST module. Sensitivity tends to increase with increasing fragment length, whereas precision tends to stay consistently high and not show a clear trend in relation to fragment length.

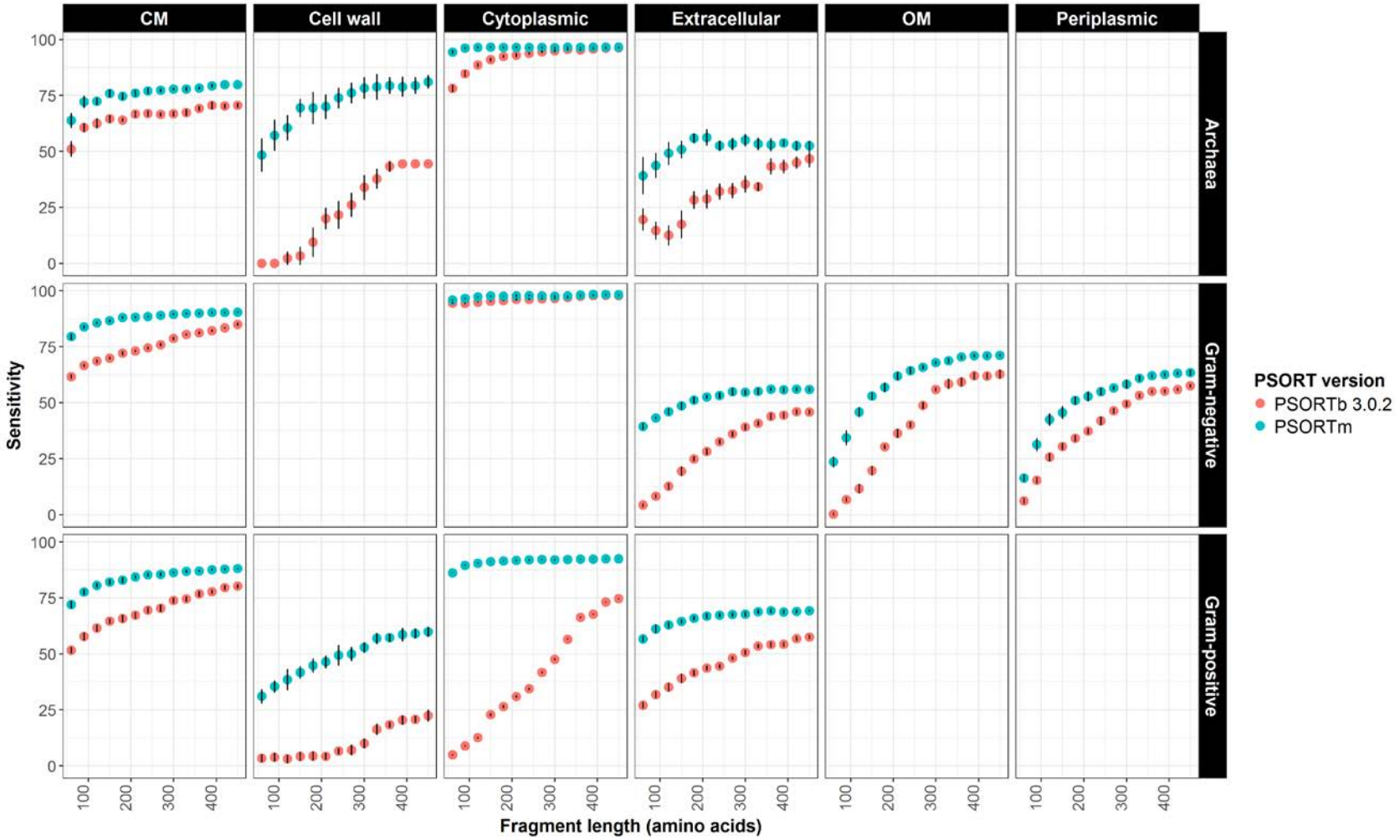

Figure 4.1 Five-fold cross validation of PSORTm sensitivity over differing taxa, SCL, and sequence fragment length.
PSORTm has higher sensitivity than PSORTb 3.0.2, and sensitivity tends to increase with increasing fragment length.
Error bars show standard deviation, fragment lengths were subsampled 10 times.
CM = Cytoplasmic Membrane; OM = Outer Membrane

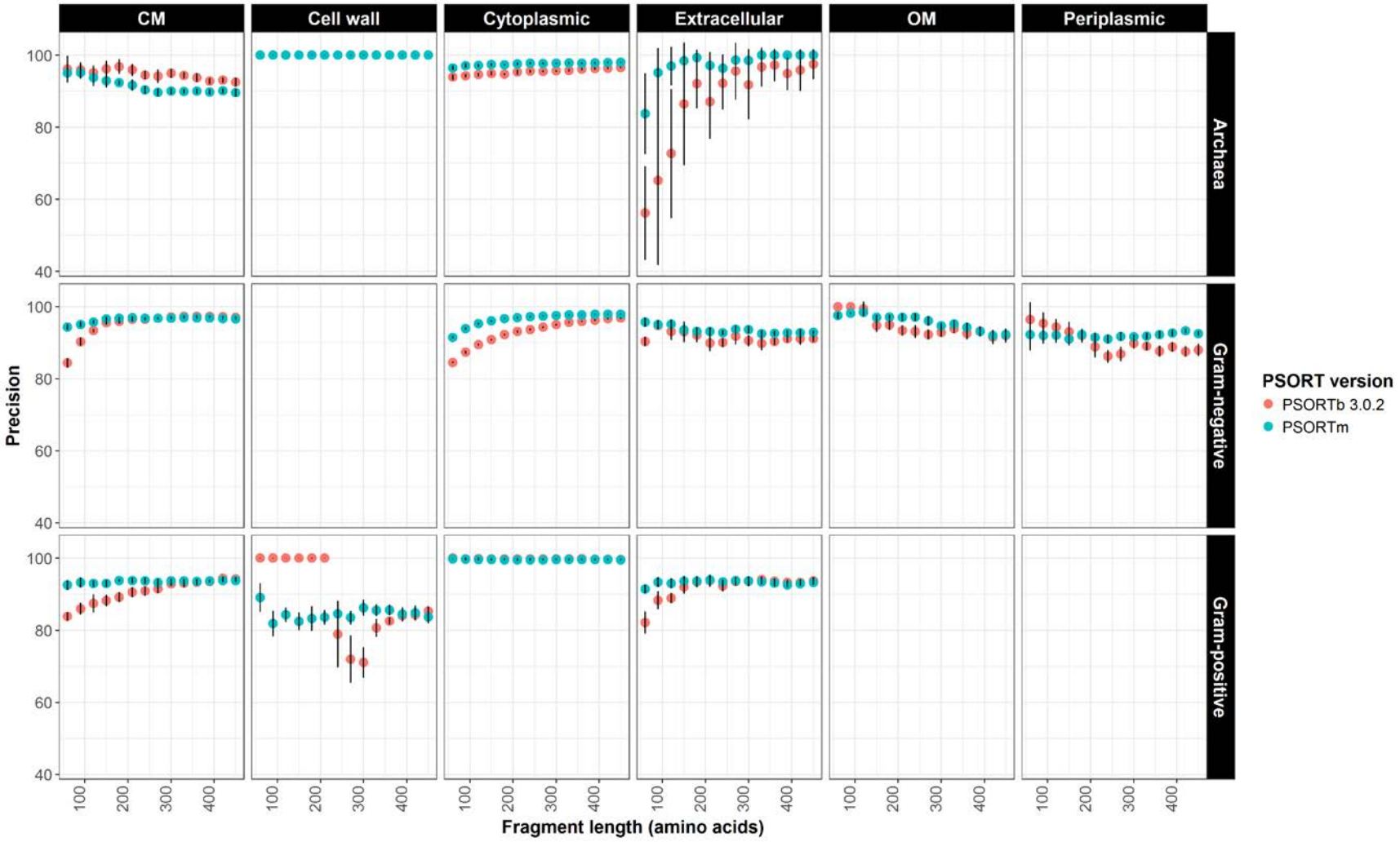

Figure 4.2 Five-fold cross validation of PSORTm precision over differing taxa, SCL, and sequence fragment length.
Precision remains consistently high in PSORTm. Error bars for some categories are larger due to the smaller number of these proteins in the test dataset. Error bars show standard deviation, fragment lengths were subsampled 10 times. CM = Cytoplasmic Membrane; OM = Outer Membrane

4.4.2. Analysis of watershed microbiomes

The taxonomic classification files and predicted protein/amino acid sequence files were input into PSORTm for each watershed sample, and the average number of sequences assigned to each organism type for each of the watershed sites are shown in Figure 4.3. The agricultural ADS and APL sites have more sequences that are characterized to an organism type than the other sites (i.e. they have the least uncategorized sequences). All of the other sites have on average more sequences that could not be categorized than sequences that could categorized to any organism type. This is similar to what was found in section 3.1, where more sequences could be assigned to the family level in the ADS and APL sites relative to the other sites, and a large proportion of sequenced could not be categorized to the family level.

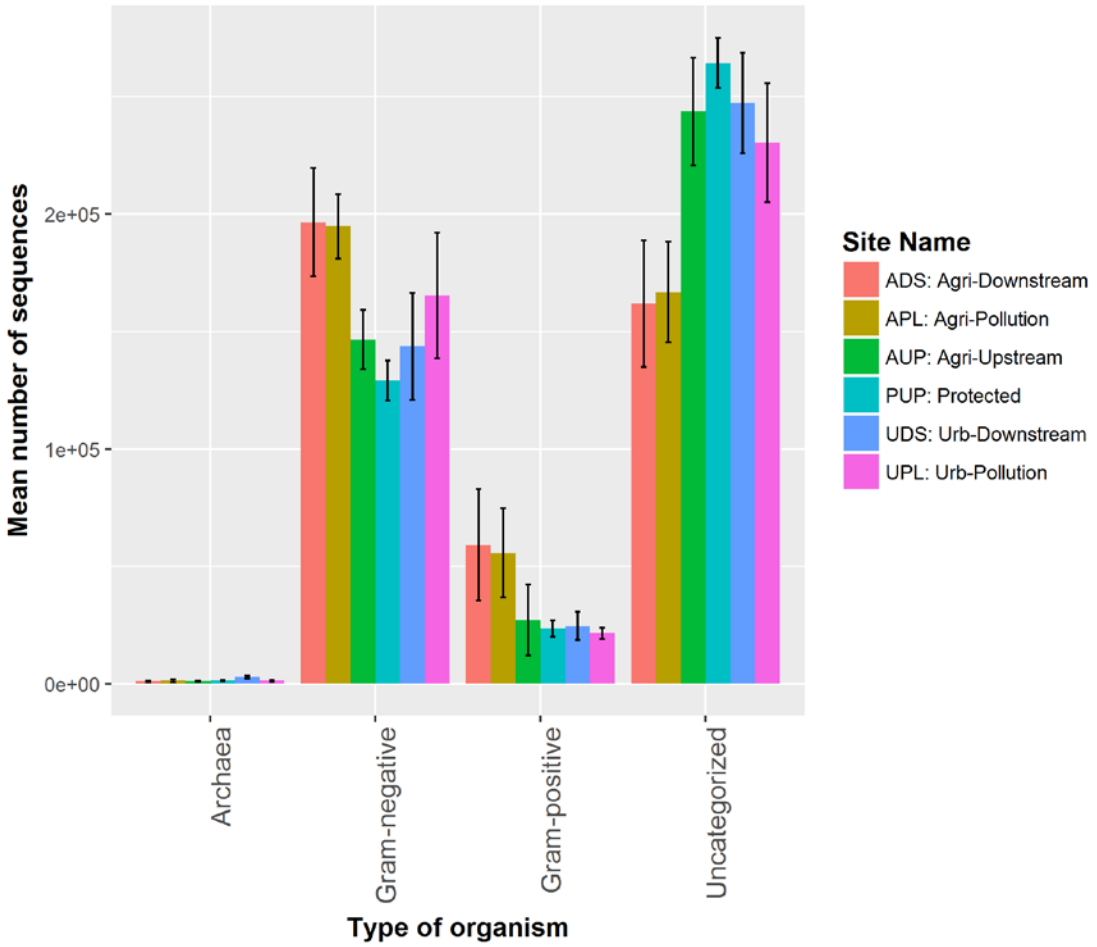

Figure 4.3 Mean number of sequences with predicted organism type for each watershed site.

More sequences are assigned to an organism type from the APL and ADS sites relative to the other watershed sites, although all sites show a relatively large proportion of sequences that were not categorized to an organism type.
Gram-negative without an outer membrane and Gram-positive with an outer membrane are not shown, as very few sequences are assigned to these categories (mean < 100). Bars represent 95% confidence intervals.

Next, the sequences categorized to the different organism/cell envelope categories were run through the second part of the pipeline of PSORTm, producing predictions of the subcellular localization of the protein sequences. More sequences were assigned to the Gram-negative organism category than to any of the other organism categories. Thus, for sequences categorized to Gram-negative, the mean number of sequences predicted to be localized to each subcellular localization, including those with predictions of unknown, is shown in Figure 4.4. Although it appears that the agricultural APL and ADS sites have significantly more sequences assigned to the cytoplasm and cytoplasmic membrane, these sites also had a greater total number of

sequences run through the second part of the PSORTm pipeline (assigning subcellular localization), due to the greater number of sequences that were assigned to an organism type in the first step of the pipeline. Thus, the results were normalized according to the number of sequences run through the second part of the pipeline, or the number of sequences predicted to a subcellular localization category other than unknown (Figure 4.5). Just as more sequences from APL and ADS could be classified to an organism type, those sequences classified to an organism type were also more likely to be classified to a subcellular localization (i.e. not unknown). When normalizing according to the number of sequences assigned to any localization, the proportion of sequences assigned to each localization is very similar among watershed sites, with perhaps a slightly higher number of sequences with a predicted localization of cytoplasmic and a slightly lower number of sequences with a predicted localization of cytoplasmic membrane in the APL and ADS sites relative to the other sites (Figure 4.5B).

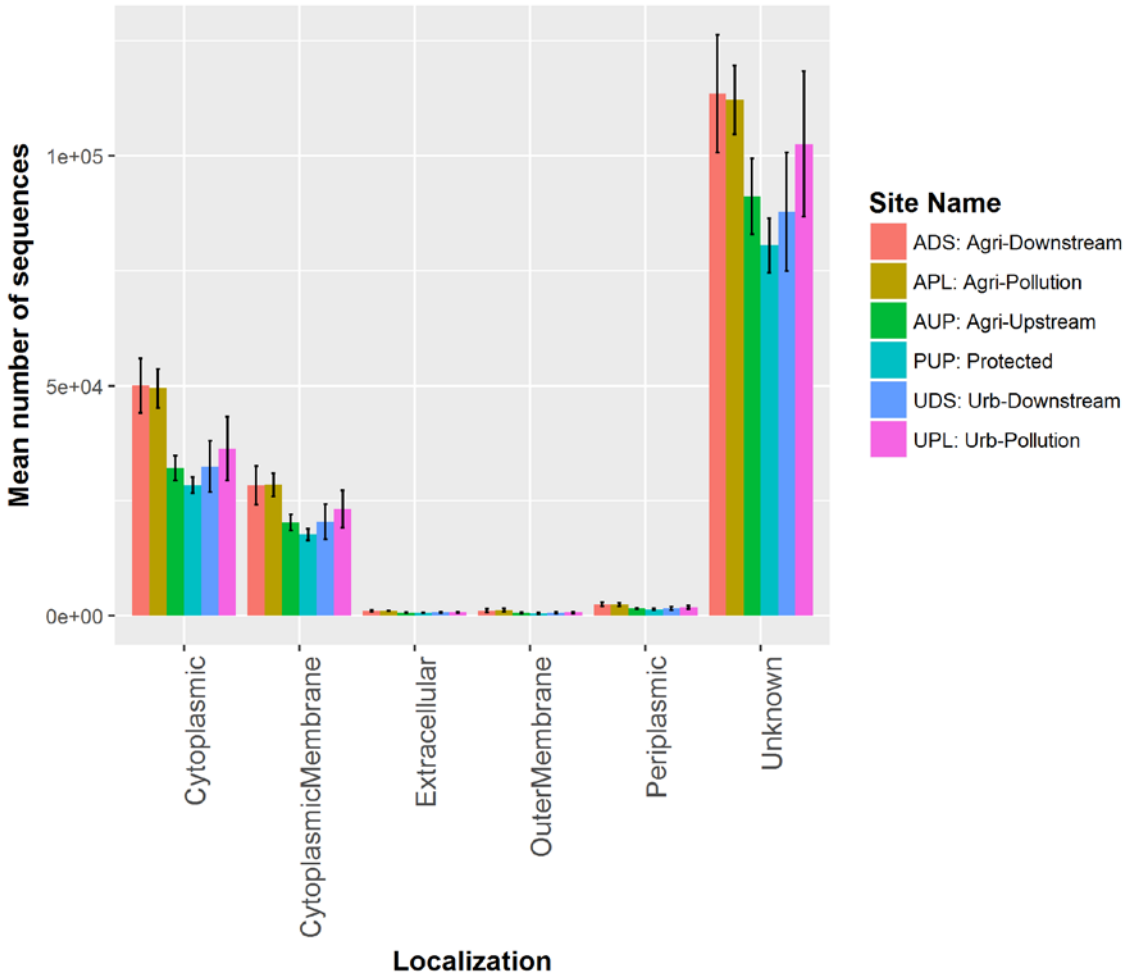

Figure 4.4 Mean number of proteins with predicted subcellular localizations for each watershed site (from sequences categorized as from Gram-negative organisms).

The ADS and APL sites appear to have more proteins localized to the cytoplasm and the cytoplasmic membrane than the other sites.

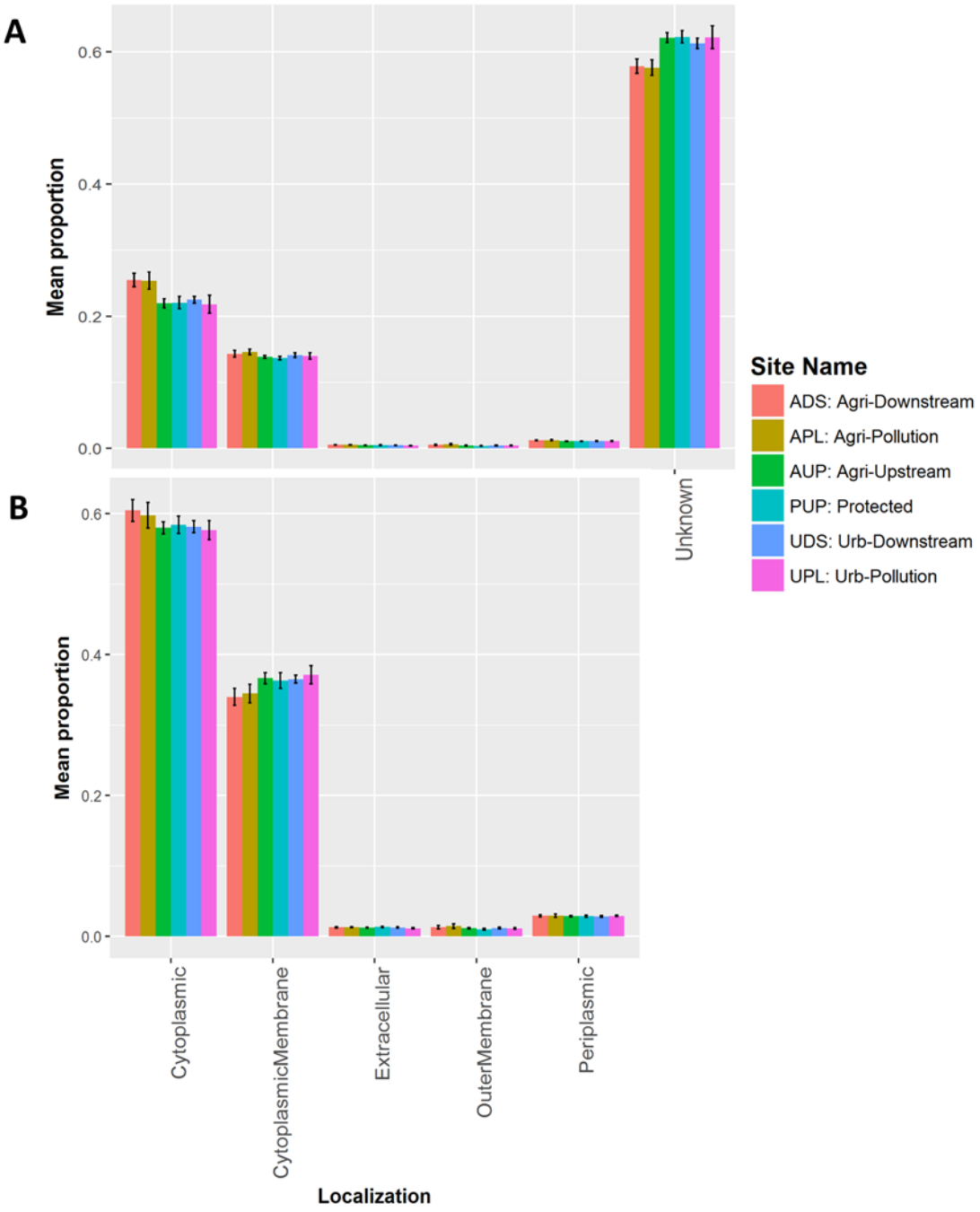

Figure 4.5 Mean number of proportion of proteins with predicted subcellular localizations for each watershed site when including predictions of unknown (A) and normalizing after excluding predictions of unknown (B).

Different approaches to normalization may lead to different interpretations of the same data. For example, the APL and ADS sites show a lower proportion of proteins with a predicted localization of cytoplasmic membrane when normalized by the total number of reads assigned to a localization (B). This was not seen when looking at raw counts (Figure 4.4) or normalized by the number of input sequences (A).

Researchers may be interested in exposed (cell surface or secreted) proteins for a variety of reasons, such as their potential use as vaccine targets or diagnostics. To see whether the microbial profile of the exposed subset of sequences—the sequences with predicted subcellular localization of extracellular, outer membrane, or cell wall—were comparable to the microbial profiles from the entire set of sequences in a sample, the microbial profiles of this exposed subset versus the original (entire set) of sequences in the agricultural watershed sites were plotted (Figure 4.6). The microbial profiles are similar, although certain taxa are either overrepresented (Bacillacaea) or underrepresented (Conexibacteraceae) in the exposed subset of sequences.

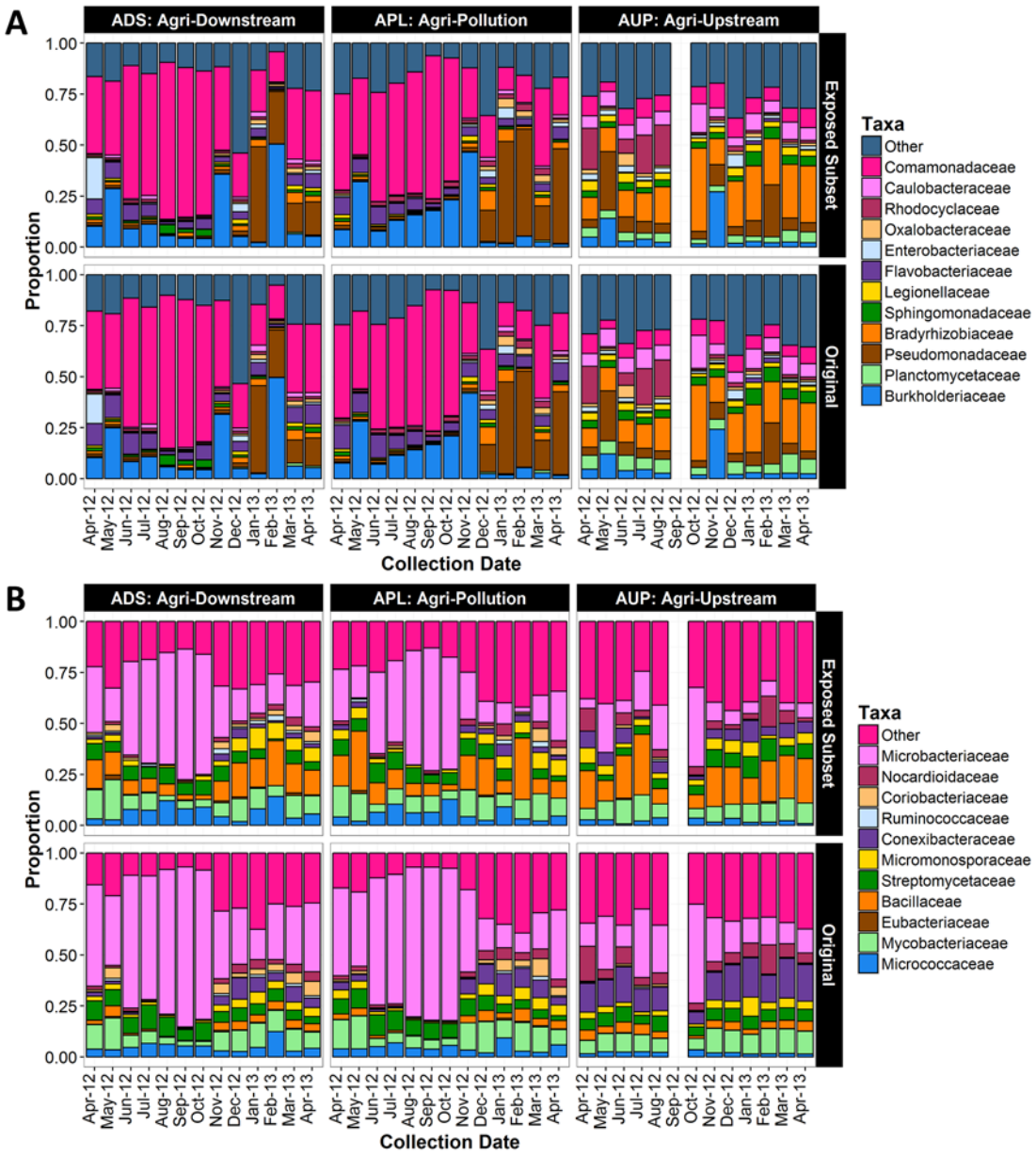

Figure 4.6 Most abundant predicted taxa over time in the agricultural watershed on the exposed proteins predicted by PSORTm (Exposed Subset) versus the full set of data (Original) for Gram-negative (A) and Gram-positive (B) organisms.

The microbial profiles are similar, although certain taxa are either overrepresented (e.g. Bacillacaea) or underrepresented in the exposed subset of sequences.
Exposed subset: reads assigned to localizations of extracellular, outer membrane, or cell wall.
Most abundant taxa include the union of the top 10 most abundant families in the exposed subset and the top 10 most abundant families in the original subset.
Original: all reads.

In a previous study, samples in the agricultural watershed had been separated into two groups: less affected samples (agricultural upstream site samples and agricultural polluted and downstream site samples collected in the drier months, May to October) and more affected samples (agricultural polluted and downstream site samples collected in the wetter months, November to April) (Van Rossum et al., 2015). The samples had been categorized using the Canadian Council of Ministers of the Environment (CCME) Water Quality Index (WQI), which is a framework to evaluate surface water quality for the protection of aquatic life (Canadian Council of Ministers of the Environment, 2007). The more affected site samples had CCME WQI ratings of “marginal” or “poor”, while less affected site samples had ratings of “fair”, “good”, or “excellent” water quality based on guidelines for ammonia, chloride, dissolved oxygen, nitrate, pH, and orthophosphate. Here, we compare the results of this previous analysis to the results when the analysis is re-run on the subset of sequences predicted to be exposed by PSORTm (Table 4.2). Of the 12 differential gene function categories identified with a fold change of greater than 2.5 in the more affected samples versus the less affected samples, only 3 were also identified in the subset of sequences predicted to be exposed by PSORTm (highlighted in Table 4.2). Some of the gene function categories with the highest fold change in the exposed subset are unsurprisingly related to proteins expected to be exposed/found in membranes such as “sodium hydrogen antiporter” and “bacterial chemotaxis”.

Table 4.2 Comparison of differentially abundant gene functional categories between samples with higher and lower water quality in the agricultural watershed for the full set of reads versus the subset of reads predicted to be exposed by PSORTm

Differential SEED subsystem	Seed class	q-value[1]	Fold change[2]
Original analysis on full set of reads			
Malonate decarboxylase	Carbohydrates	0.0045	4.6
Nitrosative stress	Nitrogen metabolism	0.0045	3.7
Denitrification	Nitrogen metabolism	0.0065	3.5
Phage capsid proteins	Phages, prophages, transposable elements	0.0045	3.5
Na(+)-translocating NADH-quinone oxidoreductase and rnf-like group of electron transport complexes	Respiration	0.0074	2.9
Lysine degradation	Amino acids and derivatives	0.0115	2.8
Pyruvate ferredoxin oxidoreductase	Carbohydrates	0.0065	2.8
Bacterial hemoglobins	Stress response	0.0125	2.7
D-galactarate, D-glucarate, and D-glycerate catabolism	Carbohydrates	0.0065	2.6
D-galactonate catabolism	Carbohydrates	0.0098	2.6
Pyrimidine utilization	Nucleosides and nucleotides	0.0065	2.6
RNA 3' terminal phosphate cyclase	RNA metabolism	0.0065	2.5
Analysis on subset of reads predicted to be exposed by PSORTm (extracellular, outer membrane, or cell wall)			
Sodium hydrogen antiporter	Membrane transport	0.0025	6.3
Bacterial chemotaxis	Motility and chemotaxis	0.0025	4.5
Denitrification	Nitrogen metabolism	0.0025	4.2

Arginine and Ornithine Degradation	Amino acids and derivatives	0.0053	4.2
Terminal cytochrome O ubiquinol oxidase	Respiration	0.0047	3.8
D-ribose utilization	Carbohydrates	0.0047	3.6
Bacterial hemoglobins	Stress response	0.0048	3.4
Terminal cytochrome oxidases	Respiration	0.0047	3.3
Polyamine Metabolism	Amino acids and derivatives	0.0047	2.9
Ribitol, Xylitol, Arabitol, Mannitol, and Sorbitol utilization	Carbohydrates	0.0047	2.9
Purine Utilization	Nucleosides and nucleotides	0.0047	2.8
Methionine Biosynthesis	Amino acids and derivatives	0.0032	2.8
Methionine Degradation	Amino acids and derivatives	0.0033	2.8
Na(+)-translocating NADH-quinone oxidoreductase and rnf-like group of electron transport complexes	Respiration	0.0053	2.8
Alkanesulfonates Utilization	Sulfur metabolism	0.0142	2.7
Glycine and Serine Utilization	Amino acids and derivatives	0.0048	2.7
Lipid A modifications	Cell wall and capsule	0.0047	2.5
Terminal cytochrome d ubiquinol oxidases	Respiration	0.0047	2.5

Only three of the top differential gene functional categories are the same for the full set of reads versus the exposed subset (cell surface or secreted).
Differential gene functional categories with fold change greater than 2.5 are listed and sorted by fold change, and the differential gene functional categories that were identified by both the entire set of sequences and the subset of sequences with a predicted exposed localization are highlighted.
[1]Corrected for FDR using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995)
[2]Median abundance in lower quality water divided by median abundance in higher quality water.

4.5. Discussion

We have developed PSORTm, a modified version of PSORTb 3.0.2 that can predict the subcellular localization of proteins encoded by metagenomics sequences. PSORTm maintains a high level of precision over a range of fragment lengths. Sensitivity also generally remains high, tending to show a modest increase as input fragment length increases. However, there are certain categories of localizations that benefit much more from increased fragment lengths, such as outer membrane proteins in Gram-negative bacteria. The sensitivity for these outer membrane fragments increases from 25% to almost 75% as input fragment length increases from 60 to 450 amino acids, emphasizing the importance of longer metagenomics sequence lengths.

Applying PSORTm to the metagenomics sequences from the Applied Metagenomics of the Watershed Microbiome project demonstrated that a large proportion of metagenomics sequences could not be assigned to a subcellular localization site, and were thus given an unknown prediction. This is not surprising, as the microbial communities inhabiting fresh water are not well characterized. The sampling sites where more sequences could be taxonomically assigned (APL and ADS), were also the sites with a larger proportion of sequences predicted to a specific localization. Thus, it is likely that a greater proportion of metagenomics sequences will be assigned a subcellular localization from microbial communities from better characterized environments such as the human gut (Human Microbiome Jumpstart Reference Strains Consortium et al., 2010).

The importance of normalization was also demonstrated by applying PSORTm to the metagenomics sequences from watershed samples. When looking at raw numbers, the APL and ADS sites had a greater number of predicted cytoplasmic membrane proteins, but when normalizing by the number of sequences assigned to a subcellular localization, the proportion of predicted cytoplasmic membrane sequences was actually lower in APL and ADS relative to the other watershed sites. There is likely no best way to analyze the data, but researchers should be aware that the choices they are making (such as to look at the raw number of reads assigned to a category, or to normalize based on the number of reads assigned to any category but unknown) may affect the interpretation of the results.

As researchers may be interested in exposed (cell surface or secreted) proteins due to their potential industrial applications or use as vaccine targets or diagnostics, taxonomic and gene function analyses were carried out comparing the profiles derived from the subset of sequences predicted to be exposed (extracellular, outer membrane, or cell wall localizations) versus the profiles derived from the entire set of sequences.

The taxonomic profile comparison revealed that the taxonomic profiles derived from the subset of sequences with a predicted exposed localization were similar to those derived from the entire set of sequences. However, one of the taxa, Bacillacaea was found to be overrepresented in the profiles from the exposed subset (Figure 4.6). Notably, Bacillaceae contains the genus *Bacillus*, of which many species secrete numerous proteins into the environment (Simonen and Palva, 1993).

Bacillus species are of great interest to researchers in academia and industry, in part due to the large number of proteins they secrete. For example, *Bacillus amyloliquefaciens* produces several proteins of interest such as natural antibiotics (e.g. barnase and plantazolicin), the protease subtilisin used in detergents, and alpha amylase used in starch hydrolysis (Deb et al., 2013; Molohon et al., 2011; Vasantha et al., 1984). *Bacillus licheniformis* secretes an α-amylase used at high temperatures for liquefying starches, and *Bacillus subtilis*, which is considered the best studied Gram-positive organism, is involved in the production of the traditional fermented soybean dish natto (Schallmey et al., 2004). Other species of *Bacillus* are pathogens, causing anthrax (*Bacillus anthracis*) and food poisoning (*Bacillus cereus*) (Guinebretière et al., 2002; Spencer, 2003).

Bacillacaea, a family containing species of interest in large part due to the relatively large numbers of proteins secreted, was identified as a taxonomic group overrepresented in the microbial profile of the exposed subset of sequences versus the full set of sequences. Therefore, a strategy for identifying taxonomic groups that may contain microbes that are potential pathogens or producers of products of industrial interest, may be to compare taxonomic profiles from an entire set of sequences versus the subset that are predicted by PSORTm to be localized to the cell surface or extracellularly and look for taxa overrepresented in the exposed group.

The comparison of differential gene functional categories between the more affected and less affected agricultural samples demonstrated that the functional profiles derived from the subset of sequences with a predicted exposed localization were, unsurprisingly, substantially different from those derived from the entire set of sequences. The differential functional categories identified only by the exposed set of sequences seemed to be focused more on membrane associated proteins, such as the categories “sodium hydrogen antiporter”, “bacterial chemotaxis”, and “lipid A modifications”. These functional categories may have potential as biomarkers for water quality, and, if they are cell surface or secreted proteins, would be targetable with ELISA-based tests.

PSORTm is the first readily available protein subcellular localization predictor for metagenomics sequences with open source code freely available, a Docker image (available at https://hub.docker.com/r/brinkmanlab/psortm/), and plans for an online version. It has been developed and shown to maintain high precision across a wide range of sequence lengths. This tool has many potential applications, such as in the identification of cell-surface based biomarkers for protein-based diagnostic tests, or identification of taxa that may be of interest due to a relatively higher number of cell surface or secreted proteins. PSORTm may be useful not only in studies of freshwater samples, but also in a wide range of studies involving microbial communities.

Chapter 5.

PSORTdb: expanding the Bacteria and Archaea protein subcellular localization database to better reflect diversity in cell envelope structures

Chapter 5 describes the identification of markers for bacteria with atypical cell envelopes, which builds on previous work identifying OMP85 as a marker for outer membrane containing bacteria. This work was then utilized in the expansion of PSORTdb to improve protein subcellular localization data from bacteria with atypical cell envelopes. Portions of this chapter have been previously published in the article "PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures" by M.A. Peabody, M.R. Laird, C. Vlasschaert, R. Lo, and F.S.L. Brinkman in Nucleic Acids Research, 44 (D1). © 2015 Peabody et al; licensee Oxford University Press.

I completed all work presented in this chapter with the following exceptions: Caitlyn Vlasschaert, an undergraduate student whom I supervised, performed the identification of markers for atypical cell envelopes described in Section 5.1; Matthew Laird implemented the computational predictor for identifying atypical cell envelope structures into cPSORTdb; and Raymond Lo was responsible for the manual curation of new proteins in ePSORTdb.

5.1. Identification of markers for atypical cell envelopes

5.1.1. Abstract

Systematic *in silico* determination of prokaryotic cell envelope structure can be achieved through identification of structure-specific protein markers. While Omp85/BamA distinguishes classical outer membranes, reliable markers for the classification of atypical diderm structures had previously remained elusive. We generated lists of markers specific to Thermotogae, Deinococci and Corynebacteriales (formerly Corynebacterineae), prominent atypical diderm phyla. Three proteins uniquely and ubiquitously define all 17 sequenced members of Thermotogae. The two sets of Deinococci markers, one representing Deinococcus-Thermus and the other restricted to Deinococcales, contain seven and four highly conserved markers, respectively. Notably, an essential outer membrane lipase, is specific to all 96 complete and nearly all 119 incomplete Corynebacteriales genomes. These atypical membrane marker lists have been greatly refined in comparison with earlier studies owing to the exponential increase in sequenced genomes, which underscores the importance of routine re-verification of previously identified markers such as Omp85. Employing protein markers for rapid, accurate bacterial membrane typing would further broaden the predictive capacities of our precise prokaryotic protein subcellular localization computational tool, PSORTb, and can have other applications in phylogenetics, metagenomics, diagnostics and gene ontology.

5.1.2. Introduction

Bacteria cell envelopes can be categorized according to whether they have one (monoderm) or two (diderm) cellular membranes. Classical monoderms have a relatively thick peptidoglycan layer in their cell wall and retain Gram staining, while diderms, which have a thin peptidoglycan layer, stain Gram negative. Furthermore, diderms typically contain lipopolysaccharide (LPS) and lipoproteins such as Omp85/BamA in their outer membrane. However, the layers that compose bacterial cell envelopes vary more extensively than the above broad classification would suggest: monoderm and diderm are not entirely synonymous with Gram-positive and Gram-negative (Sutcliffe, 2010). Some bacterial clades have evolved atypical outer membranes that cannot be properly classified based on peptidoglycan thickness (Figure 5.1).

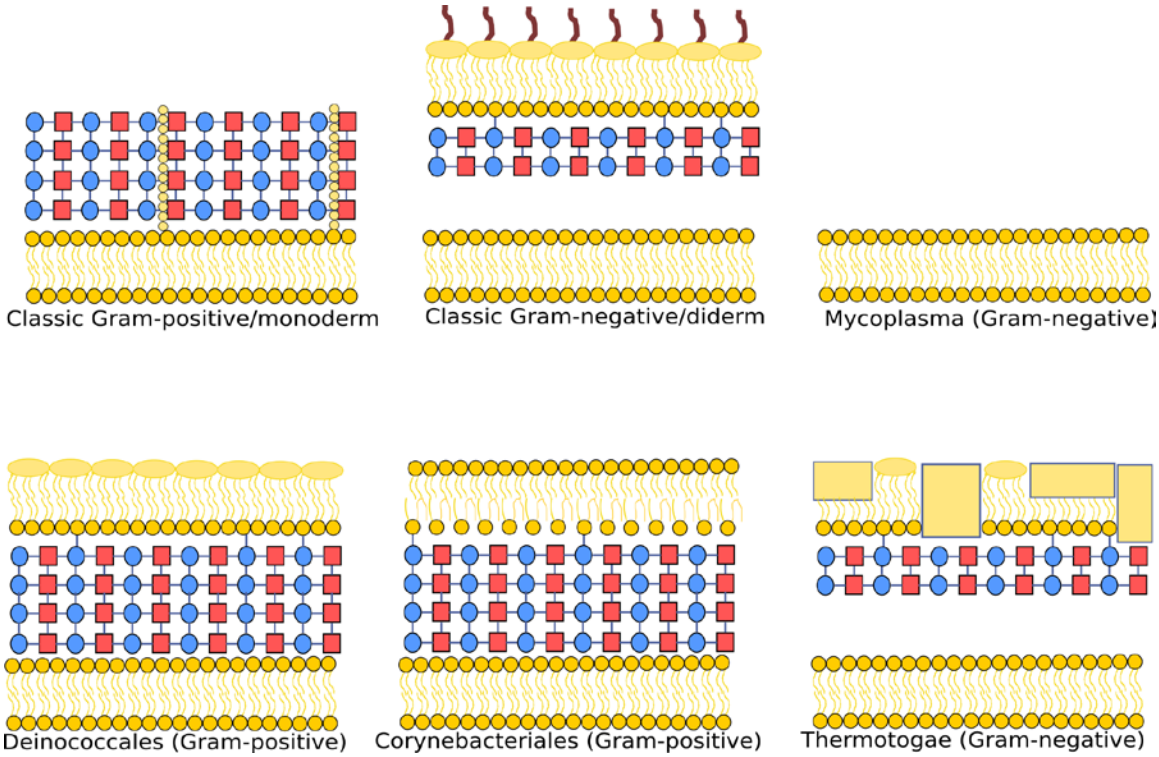

Figure 5.1 Schematic illustrating the diversity of arrangements for bacterial cell envelopes, with selected examples.

Classic Gram-positive bacteria are monoderms (contain one cell membrane in their cell envelope) with a thick peptidoglycan layer, and classic Gram-negative bacteria are diderms (enveloped by two membranes) with a thin peptidoglycan layer between the inner and outer membranes. Examples of bacteria with other cell envelope structures include *Mycoplasma* (monoderms without a peptidoglycan layer), Deinococcales (diderms with a thick peptidoglycan layer that lacks lipopolysaccharide), Corynebacteriales (diderms with a thick peptidoglycan layer and unique outer membrane containing mycolic acids also known as a mycomembrane), and Thermotogae (diderms with a unique outer membrane, called a toga, which is very rich in proteins). These schematic diagrams do not show certain components such as S-layers. Red squares represent *N*-acetylglucosamine, blue circles represent *N*-acetylmuramic acid, beige squares represent proteins, and lipopolysaccharides are shown in brown.

Cell-type specific proteins are regularly used as markers in both eukaryotes and prokaryotes. Omp85/BamA, for example, is found in bacteria with outer membranes and has been used for sorting bacterial cell wall types into one of two prototypical categories, either Gram-negative ($Omp85^{+}$) or Gram-positive ($Omp85^{-}$) (Gentle et al., 2004). However, there do not exist reliable protein markers for many bacterial clades with atypical cell envelopes, such as extremophiles Deinococci, Thermotogae, Dictyoglomi (Francke et al., 2011), and the Corynebacteriales (Burkovski, 2013) which contain notable pathogens. Limited genomic sequence availability has until recently impeded Thermotogae and Dictyoglomi marker elucidation altogether, while previous marker lists

for Deinococci and Corynebacteriales should be re-verified using the greater number of genomes now available to determine whether these remain reliable markers.

Members of the early-branching phylum Deinococcus-Thermus are considered atypical in that their cell envelopes include features of both Gram-positive and -negative bacteria. They contain a multi-layered cell envelope with peptidoglycan and an outer membrane containing Omp85, however, the lipid-rich outer layer is devoid of LPS (Figure 5.1). There are two main groups within the phylum Deinococcus-Thermus, the Deinococcales and Thermales, both of which belong to the class Deinococci though they differ by Gram staining and radiosensitivity. Deinococcales retain Gram staining (Albuquerque et al., 2005), possibly due to greater peptidoglycan thickness, and exhibit radioactivity resistance (Daly, 2009). The Thermales are not universally regarded as atypical bacteria due to their radiosensitivity and greater morphological resemblance to traditional Gram-negative bacteria (Albuquerque et al., 2005). The last formal assessment of Deinococci markers compared the three genomes available at the time and listed 65 signature proteins (Griffiths and Gupta, 2007).

Next, each affiliate of the Thermotogae phylum has a unique outer membrane-like structure, the “toga”, which consists of an orderly matrix of proteins enclosing the periplasm (Robert Huber, 1986). The toga likely confers their characteristic hyperthermophilicity, akin to the radio-protective sheath of Deinococcales, and contains Omp85 (Sutcliffe, 2010). Contrary to Deinococcales, however, Thermotogae cell walls stain Gram-negative (Robert Huber, 1986). Nevertheless, they can be unambiguously characterized as atypical because their distinctive outer membrane structure permits one of the highest growth temperatures among bacteria (65-90˚C) (Conners et al., 2006), it consists primarily of proteins rather than lipids in contrast to traditional diderms (Gupta, 2011), their peptidoglycan layer contains novel cross-linking patterns (Boniface et al., 2009), and appear to lack LPS (Sutcliffe, 2010). Previous efforts have not successfully identified signature proteins for this phylum (Gupta and Bhandari, 2011), and markers for the close phylogenetic relatives of the Thermotogae (Nishida et al., 2011), Dictyoglomi, have never been investigated.

Corynebacteriales, formerly Corynebacterineae (Whitman et al., 2012), is a suborder of the traditionally Gram-positive phylum Actinobacteria which contains the *Mycobacterium*, and are distinct in that they possess a waxy outermost layer that serves

as a proper though non-traditional outer membrane lacking Omp85 (Gentle et al., 2004), which makes the cells impervious to Gram staining. Mycolic acid variants often comprise the major constituents of these outer membranes (Gebhardt et al., 2007). Six protein markers specific for Corynebacteriales have been previously suggested in 2012 (Gao and Gupta, 2012).

To provide updated lists of markers for clades with atypical membrane structures, we used the ortholog database OrtholugeDB (Whiteside et al., 2013) to identify initial lists of proteins that had orthologs within several members in the clade of interest, but were not found in members of closely related clades. These potential markers were then verified with using BLASTP searches against the nr database of proteins from bacterial genomes to ensure that the markers were specific and thorough: they are found in all members of the clade they represent (the *in-group*) while absent from all other species (the *out-group*). We identified three markers for Thermotogae and 55 for Dictyoglomi, clades for which no previous marker lists have been determined. However, for Dictyoglomi there are only two available genomes, so we suspect that only a subset (if any) of these are truly phylum-thorough. We updated previous lists for Deinococci, from 65 proteins down to eleven (seven for all Deinococci, four for the Deinococcales), and reduced a previous list of six protein markers for Corynebacteriales down to only one. We also performed substitution rate analysis on the hypothetical Deinococci and Thermotogae markers to determine the markers that are most conserved. Routine re-evaluation of previously reported bacterial clade markers is imperative as the number of publicly available genomes increases exponentially. In turn, specific and thorough markers can be particularly useful as automatic classifiers for newly sequenced genomes.

5.1.3. Methods

Species genome list

The NCBI FTP website was used to compile a list of all available complete and draft bacterial genomes for Corynebacteriales, Deinococci and Thermotogae on 06-23-2014 (Appendix C, lists C1-C3).

Identification of candidate markers

OrtholugeDB (Whiteside et al., 2013), a database of orthology predictions for completely sequenced bacterial and archaeal genomes, was utilized to produce an initial list of candidate markers to be individually verified by iterative BLAST searches. Index genomes used as input were *Deinococcus radiodurans R1, Thermotoga lettingae TMO* and *Mycobacterium tuberculosis H37Rv* for Deinococci, Thermotogae and Corynebacteriales in-groups, respectively. One limitation in using OrtholugeDB (v. 76 & 77) is its restriction to 10 comparison genomes. We used as input different combinations of species that broadly represented the in-group in question, for which we selected the "Ortholog is present" option, as well as some near and distant relatives of the out-group ("Ortholog is absent"). The results of several successive OrtholugeDB inquiries were compiled to generate a reasonably small subset of candidate markers for each in-group. OrtholugeDB version 76 was accessed throughout July and August 2014 for the analyses. Observations were verified using a newer version 77, containing 2083 genomes, on its release date (09-17-2014).

Verification of candidate markers

Iterative BLASTP searches against the non-redundant (nr) protein database (Pruitt et al., 2007) were conducted for all candidate markers generated from OrtholugeDB. First, marker in-group specificity was evaluated using cut-off criteria of E-value < 1e-10 and query coverage >80%. Then, we verified marker thoroughness, defined as their presence in every complete genome of the in-group, and specificity, defined as their absence in every genome not in the in-group.

Optimal query sequence (OQS) derivation

OQSs are defined here as modified consensus sequences that permit enhanced distinction between in-group and out-group. DAMBE v. 5.3.57 (Xia, 2013) was used to align each set of in-group marker sequences and derive position weight matrices (PWMs), scoring matrices from which consensus sequences can be derived. Inferred indels (sites that were conserved in <50% of in-group) were eliminated. The cutinase domain of Rv3802c, the single protein marker found for Corynebacteriales, is found in various bacterial and fungal genomes, and thus was largely removed from its OQS to eliminate all false positives. One limitation of systematically deriving OQSs includes the inherent phylogenetic bias: consensus sequences are reflective of the compositional

bias of the input. This resulted in higher E-values for certain underrepresented and/or outlier species, such as *Truepera radiovictrix* for Deinococcales. To correct for this, *T. radiovictrix* was given a higher weighting (4:1) in PWM derivation. Also, only one representative sequence was used for species with multiple strains (e.g. *Mycobacterium tuberculosis*). OQSs were not obtained for Tlet_0241 or Tlet_1043 because of Thermotogae phylum divergence: the *T. lettingae* orthologs were more representative of the median than the consensus sequences and generated lower E-values. Because of its shorter length and greater similarity amongst Thermotogae, a Tlet_0008 OQS was successfully produced.

Substitution rate analysis

The hypothetical protein markers identified for Thermotogae and Deinococci were subject to substitution rate analysis to identify which markers had been under selective pressure and thus were most likely to serve as reliable markers. Hypothetical marker orthologous sequences were retrieved from each available in-group genome through a BLASTP search using the hypothetical proteins as queries. Phylogenetic trees were built from aligned codon sequences and the ratio of the number of nonsynonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks) was calculated for each branch using DAMBE v.5.3.57 (Xia, 2013).

Omp85 re-verification

A small re-assessment was performed to verify that the four Omp85 markers used in PSORTdb (Yu et al., 2011) still specifically and thoroughly represents all 22 currently named phyla of Gram-negative bacteria. The four Omp85 markers that are used in PSORTdb (Yu et al., 2011) were used as queries in a BLASTP search specifically against each of the 22 phyla of Gram-negative bacteria and the Gram-positive Firmicutes and Actinobacteria. Bacterial genomes were considered to have Omp85 present if there was a hit with an E-value <= 1e-3.

5.1.4. Results

Deinococci

We've identified seven proteins that are shared by all currently available complete genome sequences of the Deinococcus-Thermus phylum (N = 17; Appendix C): DR_0042, DR_1021, DR_1474, DR_2007, DR_2136, DR_2156 and DR_2318 (*Deinococcus radiodurans* orthologs), a list greatly pared down from the previously identified 65 proteins deduced in 2007 by Griffiths and Gupta from the sparse Deinococci sequences available at the time (Griffiths and Gupta, 2007). Of note, DR_0042 is a putative cytosolic metallophosphoesterase (Appendix C, Table C2). A high manganese(II) to iron(II) ratio is essential for the characteristic radio-resistance of Deinococci (Daly, 2009). Specialized proteins such as Dps1 (Nguyen and Grove, 2012) and perhaps DR_0042 regulate this ratio through sequestration of the metal species. Another promising marker candidate, DR_2007 virtually consists of a single domain, DUF3248 (Finn et al., 2010). Only one DUF3248 domain is found per complete Deinococcus-Thermus proteome, as determined by BLASTP analysis, and with very high sequence similarity among species, rendering it an ideal candidate as a marker. In addition, we've identified four orthologous proteins that are widely distributed solely among members of the Deinococcales branch: DR_0638, DR_0889, DR_2001 and DR_2271 (Appendix C, Table C3). These are labeled as hypothetical proteins for which, by definition, functional roles have not been assigned, though limited insight may be gained from their known functional domains (Finn et al., 2010).

For optimal implementation of these orthologous proteins as markers, we've derived modified consensus sequences that permit maximal in-group vs. outgroup separation (largest ΔE-value). We will term these "optimal query sequences", or OQSs. While simply using an ortholog (e.g. DR_2007) as a marker still permits this differentiation, OQSs are more representative of the phylum as a whole and thus might be more apt in detecting orthologs in new genomes.

OQSs were obtained by aligning orthologs from all complete genomes and creating a position-weight matrix for sites present in >50% of these. Portions of some conserved domains were eliminated to prevent false positives (see Methods section for details). The OQSs for all Deinococci markers are listed in Appendix C, Table C5.

Thermotogae

We set out to determine phylum-specific markers for Thermotogae (N = 17 complete genomes; Appendix C) that would mark the ubiquitous presence of their atypical sheathed membranes. Whereas previous attempts have failed to derive any signature proteins for Thermotogae (Gupta and Bhandari, 2011), we found three candidate marker proteins using our OrtholugeDB-based method: Tlet_0008, Tlet_0231 and Tlet_1043 (*Thermotoga lettingae* orthologs).

While Tlet_0008 and Tlet_1043 are hypothetical proteins, a conserved domain in Tlet_0231 classifies it within the LptC superfamily (Finn et al., 2010) involved in outer membrane lipopolysaccharide (LPS) transport (Appendix C, Table C4). This is notable since the outer membranes of Thermotogae, like Deinococci, appear to lack LPS (Sutcliffe, 2010). The LptC-like domain could be reflective of a convergent or, more likely, divergent evolution of Tlet_1043, in either case rendering it a Thermotogae-specific protein. Functional and molecular characterization of these three conserved proteins of undefined function would further support our postulation of these as reliable markers. Thermotogae marker OQSs can be found in Appendix C, Table C5.

Dictyoglomi

When this work was performed, only two complete Dictyoglomi genomes were available. Using our OrtholugeDB-based method, we detected 55 proteins that are found only in both *D. thermophilum H-6-12* and *D. turgidum DSM 6724* (Appendix C, Table C6). However, due to the limited number of Dictyoglomi genomes available, it would be inappropriate to use any of these as markers at this time. It is likely that this list will be further refined in coming years, similar to what has occurred for Deinococci. Dictyoglomi protein OQSs were thus not computed.

Corynebacteriales

We refined the six protein markers specific for Corynebacteriales previously suggested in 2012 (Gao and Gupta, 2012) down to a single protein, Rv3802c. Rv3802c is a cutinase enzyme, and orthologous sequences are found only in all 96 Corynebacteriales complete genomes (representing 46 species; Appendix C List C1) and 80% of the 119 species with incomplete genomes (Appendix C Table C1). Cutin is a waxy fatty acid polymer found in the protective cuticle (epidermis) of plants that bears

structural resemblance to mycolic acids (Riederer and Schönherr, 1988). Pathogenic fungi invade their botanical hosts by employing cutinases (cutin hydrolases) to damage plant wall integrity (Ettinger et al., 1987). Within the bacterial realm, the cutinase functional domain is found almost exclusively in Rv3802c and its orthologs (some non-compromising exceptions are reviewed below). Rv3802c most likely functions as a mycosyltransferase enzyme in the Corynebacteriales mycolic acid biosynthesis pathway (Takayama et al., 2005). Rv3802c would thus be more accurately referred to as both a lipase and an esterase, instead of a cutinase, reflective of its substrates (West et al., 2009).

Corynebacteriales Rv3802c orthologs can be easily distinguished from other cutinase domain proteins due to high dissimilarity of the N- and C-terminal regions. It is thus possible to distinguish Rv3802c from other non-Corynebacteriales cutinase proteins by rendering compulsory the recognition of the extremity sequences. Elimination of false positives can conceivably be accomplished deriving an Rv3802c OQS that focuses on the N- or C-terminus. The C-terminal sequence is more highly conserved among different Corynebacteriales genera than the N-terminal sequence. We've determined an OQS for Rv3802c that eliminates all Type I errors (false positives), composed of a portion of the cutinase domain and the major part of the C-terminal sequence (Appendix C, Table C5).

Of note, the N-terminus contains a ~25 residue N-terminal hydrophobic region. While the sequence of this region varies among species, its presence is noted in all Rv3802c orthologs. Akin to cytoplasmic membrane proteins, proteins anchored in the fatty acid-rich outer membrane of Corynebacteriales would encompass a hydrophobic stretch enabling membrane attachment (Marchand et al., 2012). Indeed, previous experiments have identified Rv3802c as an outer membrane protein (Hoffmann et al., 2008; Niederweis et al., 2010; Song et al., 2008), the ideal protein localization for a marker detecting organisms with Corynebacteriales-type atypical membranes.

Rv3802c has been demonstrated to be a vital gene in *M. tuberculosis* (Meniche et al., 2009; Sassetti et al., 2003), and is the only cutinase-domain protein conserved in the minimal genome *M. leprae* (West et al., 2009). Rv3802c is found within the mycolic acid synthesis gene cluster (Rv3799-Rv3807) (Ramulu et al., 2006) and the protein localizes to the extracellular face of the cell wall where it presumably regulates the

synthesis of the mycolic acid-rich waxy outer layer of *Mycobacteria* and other Corynebacteriales (Parker et al., 2009; Takayama et al., 2005; West et al., 2009). In fact, we found a novel correlation between mycolic acid chain length and active site size: Rv3802c is a traditional serine protease characterized by a catalytic triad (Ser175, Asp268, His299). West et al. pointed out that the active site residues of Rv3802c, in comparison with the other Culp (cutinase) proteins of *M. tuberculosis*, were located further apart (West et al., 2009). Specifically, the distance between Asp268 and His299 of *M. tuberculosis* Rv3802c is 19 residues, longer than in any of its semi-conserved paralogs. Rv3802c accommodates longer substrates than the other (mostly cytoplasmic) Culps, which we believe might have been achieved by a selection for a larger Asp-His distance. Since mycolic acid length differs for each Corynebacteriales genera, we then questioned whether there was a correlation between the Asp-His distance and average mycolic acid length. Notably, we found that when genera are grouped according to the length of their major mycolic acids (short vs. medium vs. long), there is a direct correlation with the Asp-His distance (Figure 5.2).

Due to its ubiquity and uniqueness to Corynebacteriales, and vital role in *M. tuberculosis* survival and outer membrane cellular localization, we propose that Rv3802c could serve as a specific structural marker for Corynebacteriales.

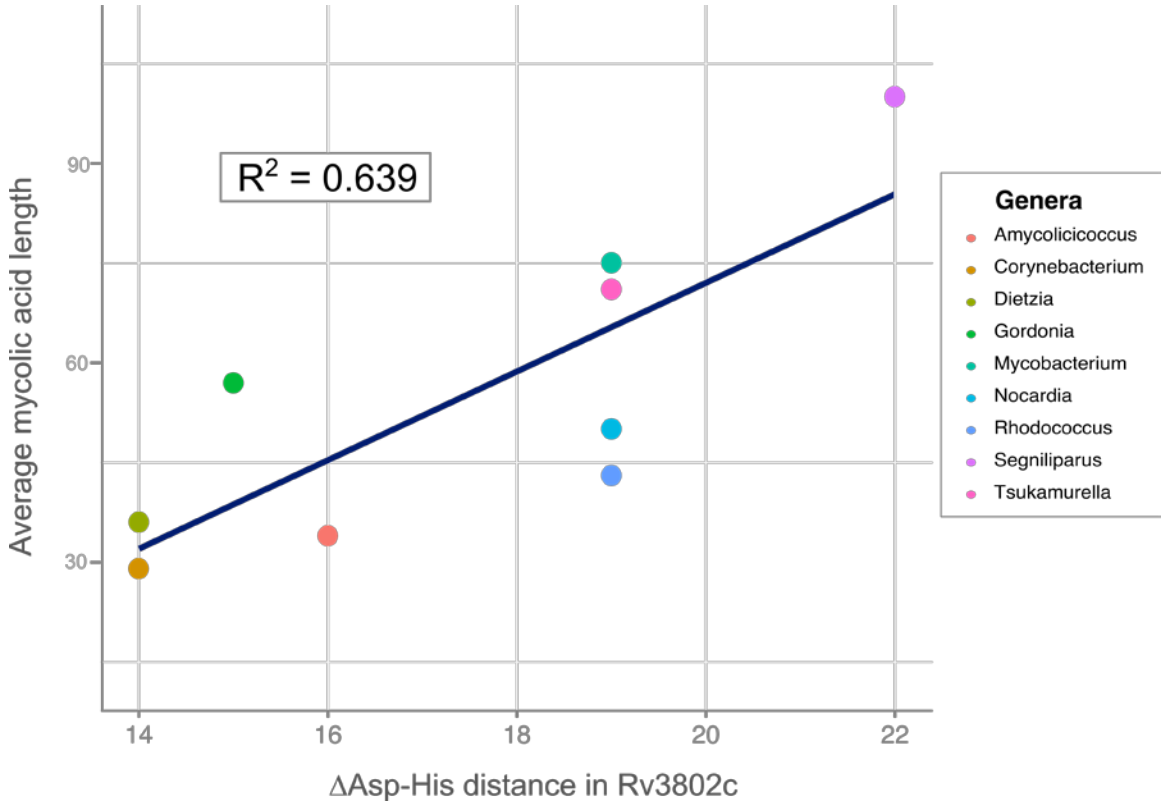

Figure 5.2 Relationship between mycolic acid length and size of cutinase active site in several genera of Corynebacteriales.
Average mycolic acid length versus Asp-His distance in the active site of cutinase of several genera of Corynebacteriales. The general trend is Asp-His distance positively correlates with mycolic acid length.

Substitution rate analysis

Since all identified Deinococci and Thermotogae markers are exclusively hypothetical proteins, we opted to assess their durability *in silico* by evaluating the evolutionary rates as well as degrees of protein conservation relative to other known markers. First, phylogenetic trees were constructed from aligned codon sequences and Ka/Ks ratio analyses were performed. Among hypothetical protein markers for Deinococci and Thermotogae, a few exhibit low Ka/Ks: 1 for Deinococcus-Thermus (DR_2156), 2 for Deinococcales (DR_0638 and DR_0889) and 2 for Thermotogae (Tlet_0008 and Tlet_0231) (Figure 5.3).

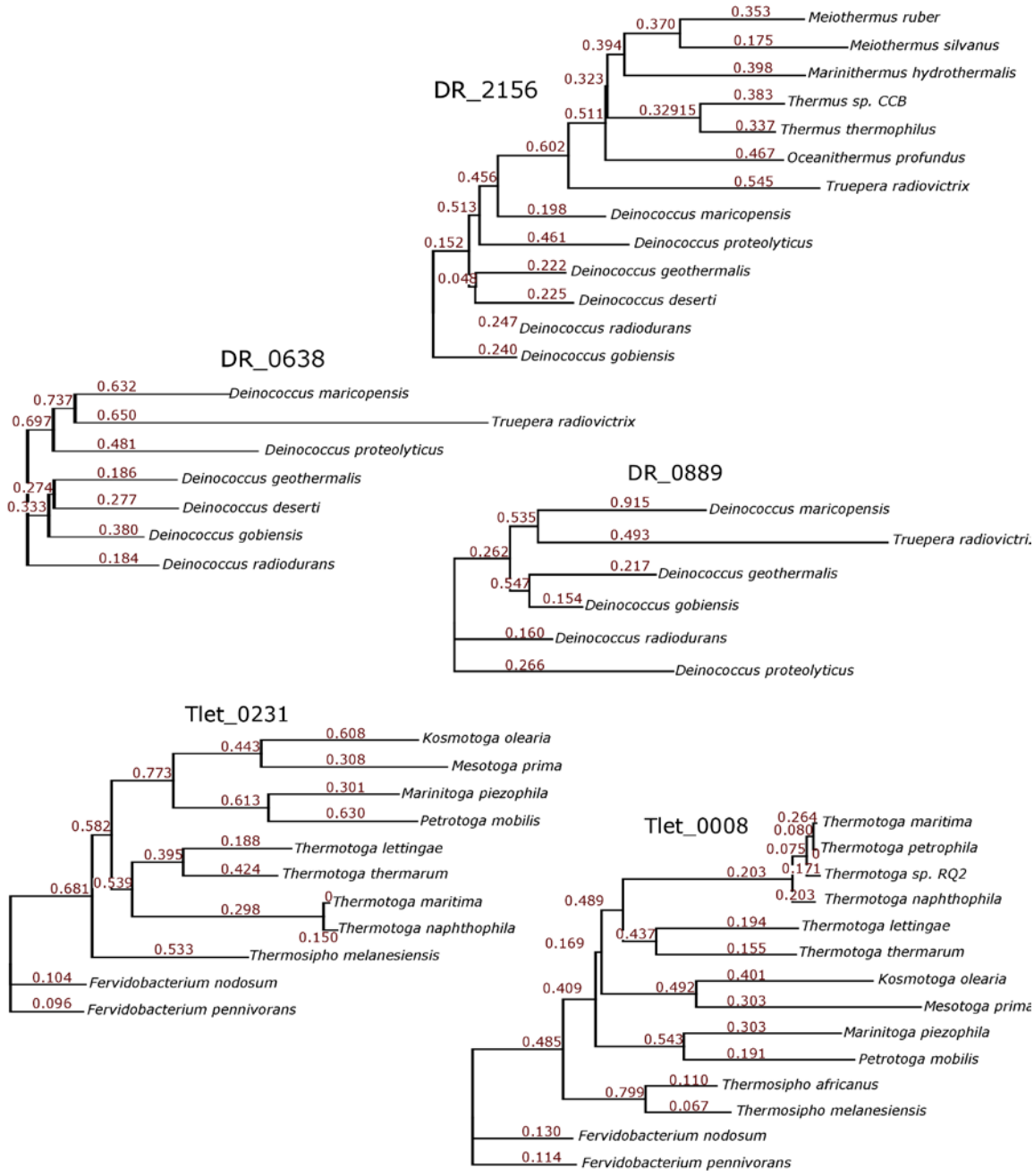

Figure 5.3 Phylogenetic trees with Ka/Ks ratios for hypothetical proteins from Deinococci and Thermotogae marker lists.

These are the well-conserved markers: 1 for Deinococcus-Thermus (DR_2156), 2 for Deinococcales (DR_0638 and DR_0889) and 2 for Thermotogae (Tlet_0008 and Tlet_0231). These markers have a low Ka/Ks ratio throughout the phylogenetic trees, indicating they are well-conserved.

Proposed markers for Deinococci, Theromotogae, and Corynebacteriales

Substitution rate analysis refined our list of proposed markers for Deinococci and Thermotogae down to DR_2156 for Deinococcus-Thermus, DR_0638 and DR_0889 for Deinococcales, and Tlet_0008 and Tlet_0231 for Thermotogae. Only a single marker, Rv3802c, was found uniquely within all complete Corynebacteriales genomes, and thus is the proposed marker for Corynebacteriales. Figure 5.4 outlines a schematic of how to run genomes using these markers, along with the known marker Omp85, to identify the type of cell envelope of a bacterium based on its genome.

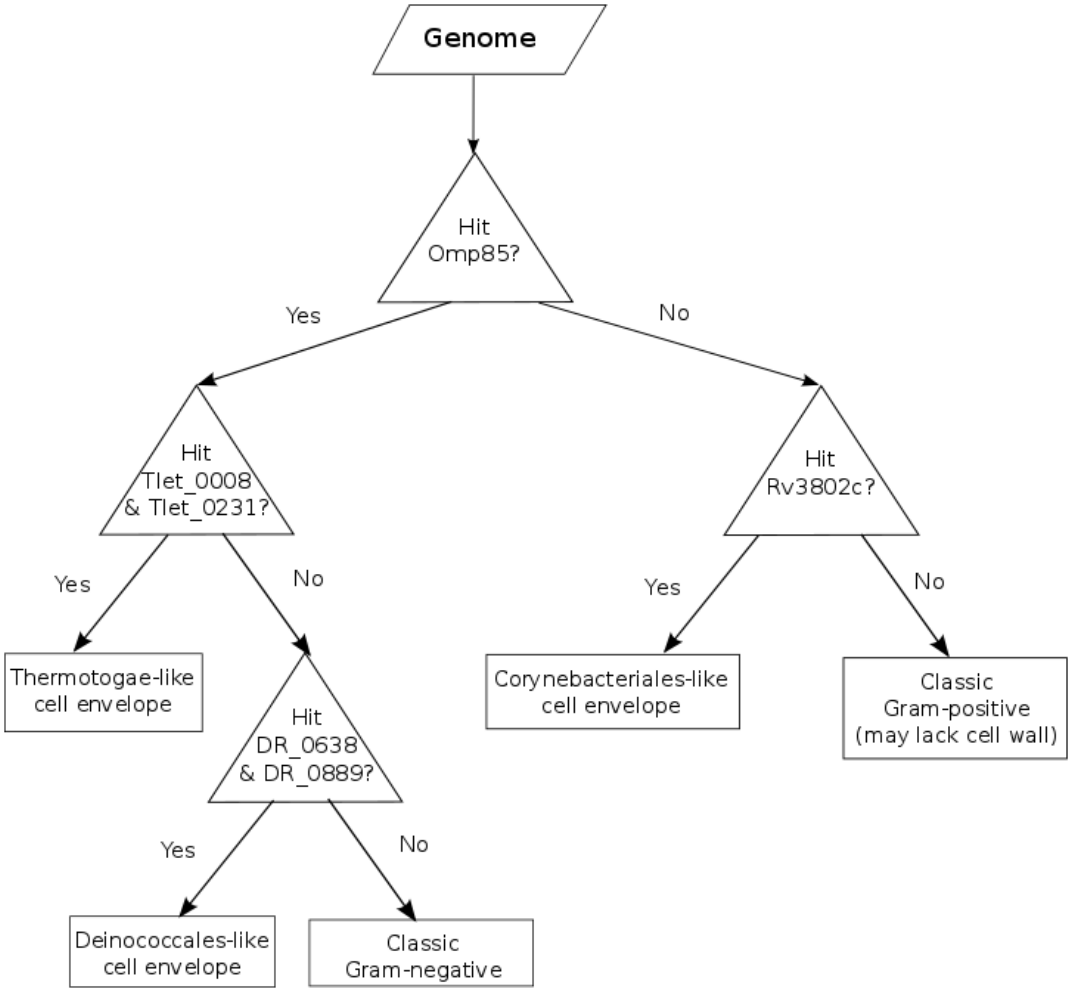

Figure 5.4 Proposed strategy to run new genomes to identify their type of cell envelope.

The presence or absence of Omp85 will separate genomes into classic Gram-positive and Gram-negative cell-envelope types, but first, the markers identified in this paper can be used to identify any atypical cell envelopes.

Omp85 re-verification

At least one of the four Omp85 markers that are used in PSORTdb was detected in all Gram-negative phyla. However, in performing a BLASTP of Omp85 against Gram-positive phyla, we uncovered several false positives: *Enterococcus gallinarum EGD-AAK12* (AccessionID: PRJNA211613), *Mumia flava* MUSC 201 genome (PRJNA261102), *Virgibacillus halodenitrificans S2* (AccessionID: PRJEB5861).

5.1.5. Discussion

Identifying precise protein markers for different bacterial membrane types becomes both increasingly important and feasible with the exponential influx of sequenced genomes.

There exists a negative relationship between the number of representative in-group sequences studied and the number of proteins specific to this in-group. For example, 55 potential markers were found for the clade for which there were the fewest genomes available (two Dictyoglomi genomes), while only 1 was found in the clade with the greatest number of genomes available (96 Corynebacteriales genomes). When fewer genomes are available for analysis, identified signature proteins may not necessarily represent phylum-wide markers. It is thus critical to periodically re-evaluate markers for continued in-group specificity and thoroughness. We have herein created and refined the lists of existing markers for atypical diderm bacterial membranes, namely those of Deinococci, Thermotogae, Dictyoglomi and Corynebacteriales, to reflect the current status of genome availability. We derived OQSs for these, which may be more useful in applications where greater precision is required. The standard marker for typical diderm membranes, Omp85, has been extensively re-verified in recent years and will be briefly reviewed below, while markers for monoderm membranes will be addressed in future studies.

Bacterial structural markers can have many applications. First, microbial databases such as PSORTdb (Yu et al., 2011) employ markers such as Omp85 for direct assignment of cell structure types from genomic sequences. However, existing databases currently lack rigorous markers for atypical membrane structures and thus often misclassify non-traditional bacteria. This poses a hurdle for certain bioinformatics tools that rely on the automated classifications generated by these databases, like

PSORTb (Yu et al., 2010b). While PSORTb 3 is one of the most precise bacterial protein subcellular localization tools at present, atypical membrane protein localizations are often undefined or, albeit less commonly, wrongly predicted by the program (for instance, it predicts a periplasmic localization for the mycolic acid outer membrane protein Rv3802c). Subsequent versions will make use of protein markers to enable proper recognition of, and thus prediction to, atypical outer membranes.

Nevertheless, the PSORTb suite is optimized for most bacteria and readily distinguishes traditional Gram-negative from Gram-positive using four divergent Omp85 sequences as outer membrane markers. In view of our emphasis on periodic marker re-assessment, we verified that 1) Omp85 still specifically and thoroughly represents all 22 phyla of Gram-negative bacteria and 2) the four Omp85 proteins used in PSORTdb are sufficient in detecting all of these phyla. As expected, Omp85 remains a reliable outer membrane marker and the Omp85 sequences in PSORTb are sufficient.

In performing a BLASTP of Omp85 against bacterial genomes classified as Gram-positive we uncovered several false positives, highlighting a second advantage of bacterial marker elucidation: reliable markers can detect phylogenetic misclassifications and contamination. As previously reported, Omp85-presenting *Enterococcus gallinarum EGD-AAK12* (AccessionID: PRJNA211613) is in fact a highly Klebsiella-contaminated sample as further evidenced by proteome homology (Pible et al., 2014). By a similar logic, we postulate that the *Mumia flava* MUSC 201 genome (PRJNA261102) may likewise be contaminated by *Burkholderia contaminans.* The proteobacterial Omp85 also identifies *Virgibacillus halodenitrificans S2* (a Gram-positive Firmicute). The S2 strain draft genome sequence (AccessionID: PRJEB5861) contains the information for two full proteomes, those of *Virgibacillus halodenitrificans 1806* and *Chromohalobacter salexigens,* a proteobacteria. There are two different CarA, HisS, 50S L14, etc. sequences in the S2 "genome" (one corresponding to *V. halodenitrificans* and the other to *C. salexigens*). Our Omp85 marker is thus identifying the *Chromohalobacter* Omp85 in *V. halodenitrificans S2*. Similarly, detection of our cutinase marker, Rv3802c, in two non-Corynebacteriales Actinobacterial genomes enabled the proper re-classification of *Pilimelia anulata* and *Microbispora sp. NRRL B-24597* genome sequences.

Markers that are well conserved throughout an in-group suggests that they play important cellular roles. Rv3802c, for instance, is vital for *Mycobacterium tuberculosis*

survival and has thus been suggested as an accessible cell-surface drug target (Parker et al., 2009; Saravanan et al., 2012; West et al., 2011). Derivation of functional importance imparts confidence in a protein marker's durability and supports its implementation as a biomarker. Since all identified Deinococci and Thermotogae markers are exclusively hypothetical proteins, we opted to assess their durability *in silico* by evaluating the evolutionary rates. A few of these markers appeared to be under purifying selective pressure throughout the clade, suggesting the markers may be robust. We used these markers, along with Rv3802c and Omp85, to propose a strategy for identifying the type of cell envelope a bacterium has by searching for these markers in the organism's genome (Figure 5.4). Of course, it is likely that there exist other bacteria with atypical cell envelopes that have not been identified yet. Markers for these new bacterial clades can later be incorporated into our classification strategy. Furthermore, the Mollicutes, which include the *Mycoplasma* species (Bové, 1993), are atypical in that they lack a cell wall, but are currently not captured by our strategy, and future studies could look at identifying robust markers for this group.

Automated methods for the identification and characterization of organisms become increasingly important as the cost of genome sequencing continues to plummet. Researchers may begin to sequence organisms without thoroughly studying their characteristics, such as cell envelope type, and will thus rely on automated methods which can predict these characteristics. In other settings, such as genomes assembled from metagenomes, researchers may not even be able to culture and study the characteristics of the organism whose genome they are studying. Thus, a greater reliance will be placed on automated methods, such as the strategy of classification of cell envelope structures for bacteria proposed here. These automated methods will need to be continuously refined as new organisms are studied and sequenced to verify they continue to accurately characterize new genomes.

5.1.6. Conclusions

We have identified and refined marker lists for bacterial clades with atypical cell envelope structures: Thermotogae, Dictyoglomi, Deinococci, and Corynebacteriales. These marker lists will allow identification and characterization of organisms in these clades, particularly for newly sequenced genomes or genomes assembled from metagenomes. Use of these markers may be useful for certain software tools, such as

PSORTb (Yu et al., 2010b), for rapid identification of organisms with atypical cell envelope structures. The exponential increase in the number of publicly available genomes will allow the identification of markers for clades in which there were previously too few genomes, but also make it imperative to re-evaluate previously reported markers.

5.2. Expanding PSORTdb utilizing identified markers for atypical diderm membranes

5.2.1. Abstract

Protein subcellular localization (SCL) is important for understanding protein function, genome annotation, and has practical applications such as identification of potential vaccine components or diagnostic/drug targets. PSORTdb (http://db.psort.org) comprises manually curated SCLs for proteins which have been experimentally verified (ePSORTdb), as well as pre-computed SCL predictions for deduced proteomes from bacterial and archaeal complete genomes available from NCBI (cPSORTdb).

We now report PSORTdb 3.0. It features improvements increasing user-friendliness, and further expands both ePSORTdb and cPSORTdb with a focus on improving protein SCL data in cases where it is most difficult – proteins associated with non-classical Gram-positive/Gram-negative/Gram-variable cell envelopes. ePSORTdb data curation was expanded, including adding in additional cell envelope localizations, and incorporating markers for cPSORTdb to automatically computationally identify whether new genomes to be analyzed fall into certain atypical cell envelope categories (i.e. Deinococcus-Thermus, Thermotogae, Corynebacteriales/Corynebacterineae, including *Mycobacteria*). The number of predicted proteins in cPSORTdb has increased from 3,700,000 when PSORTdb 2.0 was released to over 13,000,000 currently. PSORTdb 3.0 will be of wider use to researchers studying a greater diversity of monoderm or diderm microbes, including medically, agriculturally, and industrially important species that have non-classical outer membranes or other cell envelope features.

5.2.2. Introduction

Identification of protein subcellular localization (SCL) aids in understanding the function of proteins, determining their potential interaction partners, and identifying cell surface-exposed components. Bacterial and archaeal proteins can exist freely in cellular spaces such as the cytoplasm or periplasmic space, anchored in cytoplasmic or outer membranes, excreted into the extracellular space, or even injected directly into eukaryotic/host cells. The determination or prediction of SCL to any outer membrane, S-layer (surface layer), or extracellular environment (secreted), is of particular interest. These proteins may be more accessible to the immune system, so be of interest as potential vaccine components, and their accessibility also makes them more attractive as potential drug targets and for use in microbial diagnostics with medical or non-medical applications (Gardy and Brinkman, 2006; Rodríguez-Ortega et al., 2006; Romero-Saavedra et al., 2014; Severin et al., 2007).

Determination of the SCL of proteins through low-throughput laboratory experiments is accurate, providing high-quality localization information, but is laborious and expensive. High-throughput laboratory methods, involving subcellular fractionation and proteomics, are rapid and relatively cost-effective. However, they are notably less accurate, in particular due to cross-contamination of cellular sub-fractions (Huber et al., 2003; Rey et al., 2005b). Computational/*in silico* SCL prediction methods require only the genome (or gene) sequence, and high-precision computational SCL predictors, such as PSORTb (Gardy et al., 2003, 2005; Yu et al., 2010b), have been shown to exceed the accuracy of common high-throughput laboratory approaches (Rey et al., 2005b).

Although there are a variety of cellular envelope structures, most SCL predictors have focused on predictions for just the two most common types of cell envelope arrangements – the classic Gram-positive monoderms (one cell membrane) and Gram-negative diderms (enveloped by two cell membranes; Figure 5.1) (Gardy and Brinkman, 2006). Classic Gram-positive bacteria comprise primarily the cytoplasm, the cytoplasmic membrane, and a cell wall surrounding the cytoplasmic membrane containing a thick layer of peptidoglycan. Many of the most well-studied Archaea contain these same basic components as classic Gram-positive bacteria. Classic Gram-negative bacteria, however, comprise the cytosol, cytoplasmic membrane, an additional outer membrane,

and between the two membranes a periplasmic space that contains a thin cell wall composed of peptidoglycan.

Although many of the well-studied bacteria conform to the classic Gram-positive monoderm and Gram-negative diderm cell types, there are notable exceptions (Sutcliffe, 2010). Corynebacteriales, which includes notable pathogens *Mycobacterium tuberculosis* and *Mycobacterium leprae* (Burkovski, 2013), have a completely different type of outer membrane, composed of mycolic acids that stain Gram-negative or Gram-variable, even though they possess a thick peptidoglycan layer (Gebhardt et al., 2007). Some “atypical Gram-positives” stain Gram-positive due to a thick peptidoglycan layer, but they also have an outer membrane, such as Deinococcales which include *Deinococcus spp.* (Thompson and Murray, 1981). There are also atypical Gram-negatives that stain Gram-negative due to the reduced/lack of peptidoglycan cell wall, but they also have no outer membrane, such as Mollicutes which include pathogens within the *Mycoplasma spp.* (Miyata and Ogaki, 2006). There are additionally atypical Gram-negative bacteria which have a non-classical Gram-negative outer membrane. For example, the Thermotogae have a unique outer membrane, also known as a toga, which is very different from the classic Gram-negative outer membrane. This toga is likely responsible for their hyperthermophilicity (Sutcliffe, 2010), and is very rich in proteins (Gupta, 2011). A schematic overview of these different types of cell envelopes is shown in Figure 5.1.

There exist several databases containing prokaryotic SCL information. Some of these databases contain general protein annotations which include SCL, such as UniProt (UniProt Consortium, 2015). Several others, such as CoBaltDB (Goudenège et al., 2010), incorporate predictions from multiple SCL tools. There also exist databases that are targeted towards specific interests. For example, DBMLoc (Zhang et al., 2008) is a database specific for proteins with multiple SCLs. Other SCL databases are specific for certain types of bacteria, such as the LocateP database (Zhou et al., 2008) specific to Gram-positive bacteria, or ClubSub-P which contains predictions only for Gram-negative bacteria and Archaea (Paramasivam and Linke, 2011).

Although there are many databases containing SCL information, there is a need to improve support for prokaryotes with diverse cell envelope structures. PSORTdb (Rey et al., 2005a), first released in 2005, is a bacterial and archaeal SCL database

comprising two components: ePSORTdb, which contains experimentally determined, manually curated protein SCLs, and cPSORTdb, which contains computationally predicted protein SCLs derived from the SCL predictor PSORTb (Gardy et al., 2003, 2005; Yu et al., 2010b). The original version included predictions only for Gram-positive and Gram-negative bacteria, but PSORTdb 2.0 published in 2011 (Yu et al., 2011) expanded to include all predictions made by a new version of PSORTb that included Archaea, and also was set up to have automatic updates. This database has continued to be updated over the years; however, the coverage of these bacteria with diverse cell envelopes was relatively limited, for example, initially containing only 44 replicons belonging to bacteria with atypical cell envelope structures in cPSORTdb. Organisms such as medically important *Mycobacteria* did not have certain key localizations predicted, such as the mycobacterial outer membrane proteins. There was a high need to improve the database to better handle the diversity of SCLs that may be present in the Bacteria and Archaea.

Here we describe PSORTdb 3.0 (http://db.psort.org/), an expanded database that better reflects the diversity in cell envelope structures, and ensures that certain proteins of high medical interest are predicted appropriately. It builds upon PSORTdb 2.0 by including some additional user friendly features, additional manually curated annotations of proteins from bacteria with atypical cellular envelopes in ePSORTdb, and updated computationally predicted SCLs in cPSORTdb, utilizing the new ePSORTdb protein annotations and expanded SCL prediction categories. Furthermore, we improved our automated update system, adding in the computational predictor required to automatically detect bacteria with certain atypical cell envelope structures. This database will be of particular interest to researchers studying microbes outside those with the classical Gram-negative diderm and Gram-positive monoderm cell envelope structures, including medically relevant species such as *Mycobacterium tuberculosis* and *Mycoplasma pneumoniae* (Smith, 2003; Waites and Talkington, 2004), agriculturally relevant species such as *Spiroplasma citri* (Shi et al., 2014), and industrially relevant species such as *Thermotoga maritima* (Conners et al., 2006).

5.2.3. Expanded database and features of PSORTdb

User-friendly database features, and expanded subcellular localizations to better reflect bacteria with non-classical bacterial and archaeal subcellular localization

To better reflect bacterial diversity we have incorporated additional subcellular localization categories. In particular, the “toga” subcategorization/secondary localization was to highlight the notably unique structure which is a particularly protein rich envelope layer found in Thermotogae. We also specify when proteins are predicted to be in the S-layer, which is an additional subcategorization. However, note that we have continued to maintain the main localization sites used in previous versions of this database: cytoplasmic, cytoplasmic membrane, periplasm, cell wall, outer membrane and extracellular. This is critical to ensure analyses of SCL maintain stability across database versions, and enable appropriate comparisons across both PSORTdb versions and with other SCL databases. Any expanded subcellular localizations are classified under a subcategory system, to provide finer resolution predictions and highlight notable differences. For example, proteins in the Thermotogae outer membrane type structure with the subcategory name “toga” help the PSORTdb database user appreciate that the Thermotogae does not have a classical outer membrane, but rather has a specialized, unique, “toga” one. Note though, that some organisms, such as Deinococcales, have outer membranes that are simply referred to as such (even though they are not classical Gram-negatives), as that is what they are commonly referred to as.

In addition to these notable subcategory SCLs that have been added, more user-friendly features have been incorporated into the database. We have made changes to the site to help clarify certain points such as providing educational material on the different types of cell envelope structures and localizations, separating out the cPSORTdb and ePSORTdb search pages for a more customized search environment for each, and clarifying ways to search (i.e. cPSORTdb searches can be made by just genome, or more sophisticated searches on various fields, but one should be aware that multiple genomes can have the same strain name now in NCBI, and so an advanced search with a strain name can return results from multiple proteomes. In cases where one wants to identify, for example, outer membrane proteins associated with a particular strain, a genome-specific search, which can then be limited to different organisms, may be preferred).

Additionally, there is now greatly increased ease of local installation, should one want to locally run genomes not available in our automated updates. We created a Docker installation of PSORTb which can be found at: https://github.com/brinkmanlab/psortb-docker (previously, running PSORTb locally required multiple dependencies and was distinctly not a user-friendly installation). We also increased the user-friendliness of certain Ajax-based searches of the database, either by protein or by genome, to better allow more complex queries.

Expanded ePSORTdb database of proteins with experimentally determined SCL, with a focus on key proteins found in bacteria with atypical cell envelope structures

We have expanded the curated ePSORTdb database with additional entries for bacteria with atypical cell envelope structures, which came from manual literature search. In addition to more classical organisms, there are now 143 entries covering key proteins known in Gram-negative bacteria without an outer membrane, 55 entries reflecting targeted proteins found in Gram-positive bacteria with an outer membrane, and 56 entries for Thermotogae which have an additional atypical outer membrane. These entries will be of interest to researchers studying these bacteria with diverse cellular envelopes, as well as for bioinformaticians interested in training data for developing SCL predictors.

Incorporation of a more flexible computational predictor for identifying atypical cell envelope structure for cPSORTdb update computations

In PSORTdb version 2, a computational outer membrane detection procedure was introduced relying on Omp85, the only essential outer membrane protein found in all classic Gram-negative bacteria (Voulhoux et al., 2003; Yu et al., 2011). This greatly enabled automatic updating of SCL predictions from deduced proteomes of complete genomes, by enabling automated prediction of which microbial genomes should be run as a “Gram-positive” and which should be run with the “Gram-negative” designation for PSORTb. However, this procedure did not detect all outer membranes such as the unusual outer membrane of Corynebacteriales (including medically relevant *Mycobacteria*), nor did it distinguish between other outer membranes, such as the toga of Thermotogae. Thus, we incorporated markers for automated detection of such atypical outer membrane/cell envelope types, so the deduced proteomes from microbial

genomes may be correctly assigned the right PSORTb modules, enabling correct SCL predictions.

Corynebacteriales are found in the traditionally Gram-positive phylum Actinobacteria, and contain a non-traditional waxy outer membrane. This outer membrane resists Gram staining or is Gram-variable / acid-fast, even though they contain a thick peptidoglycan cell wall. This very different outer membrane does not contain Omp85, which is not surprising, as Corynebacteriales is believed to be the result of convergent evolution of an outer membrane (Houben et al., 2014). This is believed not only because the Corynebacteriales are within the phylum Actinobacteria where the rest of the known microbes have classic Gram-positive cell envelopes, but also because the composition and biogenesis of the outer membrane is completely different from traditional classic Gram-negative outer membranes (Houben et al., 2014). We have identified a signature marker for Corynebacteriales, a cutinase (West et al., 2009), which was found to be conserved uniquely within Corynebacteriales. This marker has been incorporated into PSORTdb, which acts alongside the Omp85 outer membrane detector. First, the Omp85 detector is run to identify whether the organism contains a classical outer membrane, and if it does not, the cutinase detector is run to identify whether the organism contains a Corynebacteriales mycolic acid containing outer membrane. This predictor may be useful for rapidly identifying the membrane structure of newly sequenced bacterial genomes.

Although we identified markers for automated detection of other atypical cell envelope types such as those of the Thermotogae and Deinococcales (see section 5.1), we were not highly confident that the markers covered these groups of bacteria at this time. Therefore, we decided to continue to rely on the NCBI taxonomy, which is in most cases a good indicator of cell envelope structure, though it does not contain any such data to identify microbial cell envelope types. Previously, the phylum of a genome was used as the indicator of cell envelope structure. However, not all members of a phylum necessarily share the same cell envelope structure. For example, the Deinococcales stain Gram-positive but have an outer membrane, while the Thermales within the same phylum Deinococcus-Thermus stains Gram-negative. We have updated our method that classifies cell envelope structure based on taxonomy, so that it is more flexible and can now classify based on any level of the taxonomy, allowing different cell envelope structure classifications within the same phylum. We aim to continue to expand this

database feature, to accommodate additional bacterial and archaea with atypical cell envelope structures as they are discovered, including the use of markers for rapid automatic identification of specific types of atypical cell envelope structures.

Expanded cPSORTdb database for all bacteria and archaea that have complete genome sequences, incorporating expanded localization subcategories.

cPSORTdb has been updated as more bacterial and archaeal genome sequences have become available through NCBI's microbial genome database and our associated MicrobeDB (Langille et al., 2012). As part of this process, MicrobeDB needed to be extensively updated to reflect changes in NCBI database structure which it would access. There are now SCL predictions for the deduced proteomes from over 1917 Gram-positive replicons, 5042 Gram-negative replicons (including 31 Thermotogae), 245 archaeal replicons, 123 replicons that belong to Gram-negative bacteria without an outer membrane, and 287 replicons that belong to Gram-positive bacteria with an outer membrane. In total, there are over 13,000,000 proteins with predicted SCL, with the total number of proteins predicted for each SCL summarized in Table 5.1. By examining the number of predictions (everything but unknown) relative to the total number of proteins, we can identify the coverage of proteins, or proportion of proteins that get a predicted SCL by cPSORTdb. For Gram-positive organisms, we have 2,956,985 proteins with a predicted SCL out of 3,658,804 total proteins, for an average coverage of 80.8%. There is 71.7% coverage for Gram-negative bacteria (6,181,044/ 8,619,991), 86.8% coverage for Archaea (327,046/376,790), 65.9% coverage for Gram-negative bacteria without an outer membrane (60,948/92,512), and 68.4% coverage for Gram-positive bacteria with an outer membrane (510,605/746,232).

The lower overall average coverage for bacteria with atypical cell envelopes indicates that, even with SCL entries in ePSORTdb which were incorporated into the predictor for cPSORTdb, there is still a need for additional work to be done to improve predictions for these less well studied bacteria. This may be done by gathering more experimental data, incorporating additional proteins of known SCL into the training sets of these predictors, or by the design of computational predictors (or modules of predictors) that specialize in predicting the SCL of proteins from bacteria with atypical cell envelope structures.

Table 5.1 Total number of proteins for each computationally predicted SCL site currently in the cPSORTdb dataset, grouped by type of microbe

SCL site	Gram-positive	Gram-negative[1]	Archaea	Advanced-[2]	Advanced+[3]
Cytoplasmic	1 852 434	3 792 447	249 648	37 926	333 168
Cytoplasmic Membrane	1 004 086	1 882 975	72 249	21 326	153 925
Cell wall	39 244	-	1 702	-	-
Extracellular	61 221	100 158	3 447	1 696	9 420
Outer Membrane	-	175 500	-	-	1 950
Periplasmic	-	229 964	-	-	12 142
Unknown	701 819	2 438 947	49 744	31 564	235 627

[1] Classic Gram-negative, plus those with atypical outer membranes, such as the toga in Thermotogae, which are denoted by additional subcategorization as mentioned in the text.
[2] Gram-negative without an outer membrane.
[3] Gram-positive or Gram-variable, with an outer membrane.

5.2.4. Conclusion

Genome sequencing has become substantially cheaper and quicker over the past few years, with the number of sequences deposited in publicly available databases increasing exponentially. Thus, the use of computational methods for annotation of genomes, such as predictors of SCL, has become increasingly important. This is even more pressing as new computational approaches for development of vaccines and diagnostics depend on accurate SCL prediction. This paper describes a new version of PSORTdb that has been developed which incorporates a more flexible computational predictor for identifying a wider variety of atypical cell envelope structures, an expanded data set of experimentally verified SCLs found in bacteria with atypical cell envelope structure, some improved database interface features, and an up-to-date data set of computationally predicted SCLs with expanded localization subcategories. This update, which expands the capabilities of PSORTdb to predict the protein SCLs from bacteria with more diverse cell envelope structures, will be of wide interest to researchers studying a diversity of microbes, including notable medically important microbes belonging to the Mycobacteriaceae, and organisms of industrial interest within the Thermotogae. As more diverse cell envelope structures are identified, including additional outer membranes anticipated that likely have evolved by convergent evolution, the database is well structured now to balance both predicting primary SCLs, as well as

handling more diverse or specific subcategory localizations in keeping with more diverse cell structures.

Chapter 6. Concluding Remarks

In this thesis, I first performed an evaluation of metagenomics sequence classification methods using both *in silico* and *in vitro* simulated communities. This evaluation was undertaken because when this research began, no comprehensive evaluation of metagenomics sequence classification methods had yet been published, and I wanted to determine the best methods to use for our watershed project. In order to resemble the common situation in metagenomics studies where sequenced microbial genomes have varying levels of relatedness to those in reference databases, a clade exclusion approach was taken. I found that methods vary widely in their performance, with some methods falsely predict tens to hundreds of species, even when the simulated community is composed of only 11 species. The variability in runtimes of the different methods was striking. This emphasizes the importance of the development of improved algorithms and methods, particularly as the throughput of DNA sequencing continues to increase and the total amount of sequencing data for projects gets ever larger. I found that there was no single best method – that methods each had their own strengths and weaknesses. Researchers were advised to take these strengths and weaknesses into account when choosing a classification method best suited for their purposes. Another issue that became apparent in my evaluation is that although many methods for taxonomic classification of metagenomics sequences have been developed, few of these are updated or maintained. Web servers may no longer function, and standalone software is often not updated to incorporate newly sequenced microbial genomes (only 2 out of 12 methods evaluated in Chapter 2 have been updated in the past 2 years).

With its focus on clade exclusion and how methods vary depending on the rank of clade exclusion, this evaluation complements others, such as the CAMI (Critical Assessment of Metagenome Interpretation) analysis, that have been released since this evaluation was published. The CAMI initiative has many positive characteristics, such as their community approach whereby anyone can submit their method to be evaluated. However, it would be beneficial to incorporate aspects of other successful large scale evaluations, such as the Critical Assessment of protein Function Annotation (CAFA) evaluation (Radivojac et al., 2013). For example, CAFA examines predictor performance on distinct categories of targets, such as dividing sequences into easy versus difficult target categories. Evaluations like CAMI could take a similar approach, placing

sequences into categories with differing degrees of relatedness to genomes in reference databases.

Following the evaluation of metagenomics sequence classification methods, I utilized the knowledge gained to inform an analysis of the microbial community composition of freshwater over time and space, associated with the Genome Canada Watershed Project (http://www.watersheddiscovery.ca). Samples were collected monthly (with additional select hourly sampling) over a one-year period from three watersheds (an agricultural, an urban, and a protected watershed) located in British Columbia, Canada. This analysis of watersheds revealed that sites affected by agricultural activity had lower alpha diversity than unaffected sites, that these affected sites exhibited a dynamic change in bacterial community composition from the dry to the rainy season, and that 16S rRNA and metagenomic sequencing methods show similar overall results such as similar ecological trends, with differences at the level of individual taxa. This was the first analysis of its kind of watershed/river microbiomes over such a time scale, providing valuable baseline data regarding microbial dynamics in rivers.

We hypothesized that such data may aid the identification of additional markers of water quality, since current water quality tests such as the fecal coliform test are not accurate (not all fecal coliforms are pathogens, and many waterborne pathogens are not fecal coliforms). As a first step towards the development of improved water quality tests, biomarkers were identified that could distinguish agriculturally polluted and downstream sites from an upstream site. Notably, an examination of specific potential pathogens in the watershed dataset identified *Legionella* in all watersheds and sampling sites. This was the first time that Legionella had been examined so extensively over time and space in a natural water environment (as opposed to the built environment). Together, these analyses of the watershed samples in Chapter 3 improve our understanding of freshwater microbial communities, particularly the presence and diversity of *Legionella* in natural freshwater communities, and demonstrate the feasibility of identifying biomarkers which may be used to develop a molecular based test from metagenomics sequences.

While a PCR-based diagnostic test is of primary interest, there is also an interest in the development of rapid “dip stick” or ELISA-based tests. One of the most frequent requests we got from users of the PSORTb software for protein localization prediction was for a metagenomics-compatible version, as it appears increasingly researchers wish

to identify cell surface and secreted protein biomarkers in their various samples (clinical and environmental) that have been studied using metagenomics. An ELISA-based water quality test, based on cell surface/secreted proteins, may be another option so I therefore developed PSORTm, the first readily available protein subcellular localization predictor for metagenomics sequences. PSORTm was developed by modifying the Brinkman lab's PSORTb software, one of the most precise bacterial and archaeal subcellular localization predictors. The evaluation of PSORTm using 5-fold cross validation, with sequences that were fragmented *in silico,* demonstrated that PSORTm maintains high precision similar to PSORTb, and that sensitivity is improved as input sequence fragment length is increased.

An analysis of the watershed project data with PSORTm identified sequences encoding for predicted exposed (cell surface and secreted) proteins. Notably, taxonomic profiles derived from this exposed subset were similar to those derived from the full set of data. Specific taxa overrepresented in the profiles of these "exposed" (cell surface and secreted) subset of sequences may represent potential taxa of interest. The gene function profiles derived from the "exposed" subset allowed the identification of potential exposed proteins or protein categories that may act as biomarkers of water quality. There are many potential applications for PSORTm, in addition to the prediction of cell surface or secreted proteins for the development of ELISA-based diagnostics, such as the identification of vaccine components or more accessible drug targets.

Finally, as a component of this PSORT development, markers for atypical diderm structures were identified, to enable analyses of microbes with more diverse membrane structural components, and a database of protein subcellular localization associated with PSORTb was expanded to better reflect diversity in cell envelope structures. This will act as a useful resource, not only for those wishing to identify the localization of a protein, but also for those interested in predictions of the cell envelop structures of more diverse organisms which is of interest for antimicrobial development, and other industrial uses of microbes.

This work suggests several avenues for future research. The evaluation of metagenomics sequence classification methods pointed out areas of importance and possible improvement in such methods. For example, methods need to be relatively efficient and not take too long to run if they are to be used for large projects. Many

methods tended to over classify reads if the actual species were not in the database (e.g. classify a read to an incorrect species within the correct genus), so methods may be improved by simply being more conservative in the rank of their taxonomic classifications. Further details about the strengths and drawbacks of various methods are described in Chapter 2.

The study of the microbial communities in the watershed project over space and time generated key insights regarding microbial dynamics in freshwater communities. Future work could expand many of the findings in this thesis. For example, studies could examine the variability over shorter timescales, further sampling hourly, daily, or weekly. Longer studies over multiple years could examine whether the temporal trends identified, such as the shift in community composition from the dry to rainy season, are seen over multiple years, and explore the implications of these seasonal trends. Additional watersheds should be examined, identifying microbes that are common amongst watersheds with similar features such as land use. The study of a large number of samples from a variety of watersheds will be crucial to the testing of any markers proposed for use in water quality testing. Regardless, it may be metagenomics of water could play a role in future investigations directly – in cases where water quality is a concern – using such data to aid source attribution of water pollution.

As more microbial genome sequences become available in public databases, datasets including the watershed project data should be re-examined. This is particularly true for environments where many of the microorganisms are not well characterized, such as freshwater. As shown in section 3.1, there are samples where fewer than 10% of sequence reads can be assigned to the family level. The importance of re-examining datasets with updated databases is underscored by the analysis of *Legionella*. The taxonomic classification of metagenomics sequences for the watershed project was undertaken in April of 2014. When the dataset was examined for the presence of freshwater bacterial pathogens, and *Legionella* was identified in all watershed sampling sites, I noticed that only 6 complete *Legionella* genomes were available in the nr database. I decided to re-run the samples on an updated database, and found that in slightly less than 3 years later, over 40 complete *Legionella* genomes were available. Further investigating the degree in which expanded databases of microbial genomic sequences aid metagenomics analyses should be further explored.

The finding of the ubiquity and diversity of *Legionella* in the watershed sampling sites was notable, as *Legionella* is not well characterized in the natural environment. Since this is an important human pathogen, it would seem prudent to study this further – how and under what conditions *Legionella* in the natural environment seed man-made water systems, and which *Legionella* spp. have the potential to cause disease. However, caution must be taken in such analyses to not simply assume many of the *Legionella* are pathogenic. Clarifying further when a taxon present in a metagenomics analysis is truly of medical interest will require more research and better sampling and studying our natural environment is a key step as part of such investigations.

In conclusion, the work in this thesis improves our understanding of metagenomics software classification methods and their performance, provides key insights into freshwater microbial communities in watersheds affected by varying land use, and outlines the development of a new method for the protein subcellular localization prediction from metagenomics sequences. Together, this work may be useful not only in water quality analysis and test development, but also in a wide range of studies involving other microbial communities.

References

Acinas, S.G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., and Polz, M.F. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. Appl. Environ. Microbiol. *71*, 8966–8969.

Albuquerque, L., Simões, C., Nobre, M.F., Pino, N.M., Battista, J.R., Silva, M.T., Rainey, F.A., and da Costa, M.S. (2005). Truepera radiovictrix gen. nov., sp. nov., a new radiation resistant species and the proposal of Trueperaceae fam. nov. FEMS Microbiol. Lett. *247*, 161–169.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

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

Ames, S.K., Hysom, D.A., Gardner, S.N., Lloyd, G.S., Gokhale, M.B., and Allen, J.E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. Bioinforma. Oxf. Engl. *29*, 2253–2260.

Amodeo, M.R., Murdoch, D.R., and Pithie, A.D. (2010). Legionnaires' disease caused by Legionella longbeachae and Legionella pneumophila: comparison of clinical features, host-related risk factors, and outcomes. Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. *16*, 1405–1407.

Ander, C., Schulz-Trieglaff, O.B., Stoye, J., and Cox, A.J. (2013). metaBEETL: high-throughput analysis of heterogeneous microbial populations from shotgun DNA sequences. BMC Bioinformatics *14 Suppl 5*, S2.

Apprill, A., McNally, S., Parsons, R., and Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat. Microb. Ecol. *75*, 129–137.

Atlas, R.M., and Hazen, T.C. (2011). Oil Biodegradation and Bioremediation: A Tale of the Two Worst Spills in U.S. History. Environ. Sci. Technol. *45*, 6709–6715.

Bagos, P.G., Tsirigos, K.D., Plessas, S.K., Liakopoulos, T.D., and Hamodrakas, S.J. (2009). Prediction of signal peptides in archaea. Protein Eng. Des. Sel. PEDS *22*, 27–35.

Barbaree, J.M., Breiman, R.F., and Dufour, A.P. (1993). Legionella: Current Status and Emerging Perspectives (American Society for Microbioloy).

Bazinet, A.L., and Cummings, M.P. (2012). A comparative evaluation of sequence classification programs. BMC Bioinformatics *13*, 92.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. *57*, 289–300.

Benson, R.F., Thacker, W.L., Fang, F.C., Kanter, B., Mayberry, W.R., and Brenner, D.J. (1990). Legionella sainthelensi serogroup 2 isolated from patients with pneumonia. Res. Microbiol. *141*, 453–463.

Berendzen, J., Bruno, W.J., Cohn, J.D., Hengartner, N.W., Kuske, C.R., McMahon, B.H., Wolinsky, M.A., and Xie, G. (2012). Rapid phylogenetic and functional classification of short genomic fragments with signature peptides. BMC Res. Notes *5*, 460.

Bernhard, A.E., and Field, K.G. (2000). A PCR assay To discriminate human and ruminant feces on the basis of host differences in Bacteroides-Prevotella genes encoding 16S rRNA. Appl. Environ. Microbiol. *66*, 4571–4574.

Bhasin, M., Garg, A., and Raghava, G.P.S. (2005). PSLpred: prediction of subcellular localization of bacterial proteins. Bioinforma. Oxf. Engl. *21*, 2522–2524.

Biddle, A., Stewart, L., Blanchard, J., and Leschine, S. (2013). Untangling the Genetic Basis of Fibrolytic Specialization by Lachnospiraceae and Ruminococcaceae in Diverse Gut Communities. Diversity *5*, 627–640.

Billion, A., Ghai, R., Chakraborty, T., and Hain, T. (2006). Augur—a computational pipeline for whole genome microbial surface protein prediction and classification. Bioinformatics *22*, 2819–2820.

Bodelier, P. (2011). Toward Understanding, Managing, and Protecting Microbial Ecosystems. Front. Microbiol. *2*.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., et al. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. *13*, R122.

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

Bolhuis, H., Poele, E.M.T., and Rodriguez-Valera, F. (2004). Isolation and cultivation of Walsby's square archaeon. Environ. Microbiol. *6*, 1287–1291.

Boniface, A., Parquet, C., Arthur, M., Mengin-Lecreulx, D., and Blanot, D. (2009). The elucidation of the structure of Thermotoga maritima peptidoglycan reveals two novel types of cross-link. J. Biol. Chem. *284*, 21856–21862.

Borella, P., Guerrieri, E., Marchesi, I., Bondi, M., and Messi, P. (2005). Water ecology of Legionella and protozoan: environmental and public health perspectives. Biotechnol. Annu. Rev. *11*, 355–380.

Bork, P., Bowler, C., Vargas, C. de, Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara Oceans studies plankton at planetary scale. Science *348*, 873–873.

Bové, J.M. (1993). Molecular features of mollicutes. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. *17 Suppl 1*, S10-31.

Brady, A., and Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nat. Methods *8*, 367–367.

Brady, A., and Salzberg, S.L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat. Methods *6*, 673–676.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., and Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U. S. A. *99*, 14250–14255.

Brenner, D.J., Steigerwalt, A.G., and McDade, J.E. (1979). Classification of the Legionnaires' disease bacterium: Legionella pneumophila, genus novum, species nova, of the family Legionellaceae, familia nova. Ann. Intern. Med. *90*, 656–658.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M.J., Wrighton, K.C., Williams, K.H., and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. Nature *523*, 208–211.

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

Bulashevska, A., and Eils, R. (2006). Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics *7*, 298.

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S., and Thomas, T. (2011). Bacterial community assembly based on functional genes rather than species. Proc. Natl. Acad. Sci. U. S. A. *108*, 14288–14293.

Burkovski, A. (2013). Cell envelope of corynebacteria: structure and influence on pathogenicity. ISRN Microbiol. *2013*, 935736.

Burns, D.G., Camakaris, H.M., Janssen, P.H., and Dyall-Smith, M.L. (2004). Cultivation of Walsby's square haloarchaeon. FEMS Microbiol. Lett. *238*, 469–473.

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

Campbell, B.J., and Kirchman, D.L. (2013). Bacterial diversity, community structure and potential growth rates along an estuarine salinity gradient. ISME J. *7*, 210–220.

Campbell, J., Bibb, W.F., Lambert, M.A., Eng, S., Steigerwalt, A.G., Allard, J., Moss, C.W., and Brenner, D.J. (1984). Legionella sainthelensi: a new species of Legionella isolated from water near Mt. St. Helens. Appl. Environ. Microbiol. *47*, 369–373.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., and Knight, R. (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*, 4516–4522.

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

Carvalho, F.R.S., Vazoller, R.F., Foronda, A.S., and Pellizari, V.H. (2007). Phylogenetic study of legionella species in pristine and polluted aquatic samples from a tropical Atlantic forest ecosystem. Curr. Microbiol. *55*, 288–293.

Castellarin, M., Warren, R.L., Freeman, J.D., Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R.A., et al. (2012). Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. Genome Res. *22*, 299–306.

de Champdoré, M., Staiano, M., Rossi, M., and D'Auria, S. (2007). Proteins from extremophiles as stable tools for advanced biotechnological applications of high social interest. J. R. Soc. Interface *4*, 183–191.

Chang, J.-M., Su, E.C.-Y., Lo, A., Chiu, H.-S., Sung, T.-Y., and Hsu, W.-L. (2008). PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. Proteins *72*, 693–710.

Chien, M., Morozova, I., Shi, S., Sheng, H., Chen, J., Gomez, S.M., Asamani, G., Hill, K., Nuara, J., Feder, M., et al. (2004). The Genomic Sequence of the Accidental Pathogen Legionella pneumophila. Science *305*, 1966–1968.

Cirillo, J.D., Falkow, S., and Tompkins, L.S. (1994). Growth of Legionella pneumophila in Acanthamoeba castellanii enhances invasion. Infect. Immun. *62*, 3254–3261.

Clarke, G., Stilling, R.M., Kennedy, P.J., Stanton, C., Cryan, J.F., and Dinan, T.G. (2014). Minireview: Gut microbiota: the neglected endocrine organ. Mol. Endocrinol. Baltim. Md *28*, 1221–1238.

Conners, S.B., Mongodin, E.F., Johnson, M.R., Montero, C.I., Nelson, K.E., and Kelly, R.M. (2006). Microbial biochemistry, physiology, and biotechnology of hyperthermophilic Thermotoga species. FEMS Microbiol. Rev. *30*, 872–905.

Cook, C., Prystajecky, N., Feze, I.N., Joly, Y., Dunn, G., Kirby, E., Özdemir, V., and Isaac-Renton, J. (2013). A comparison of the regulatory frameworks governing microbial testing of drinking water in three Canadian provinces. Can. Water Resour. J. Rev. Can. Ressour. Hydr. *38*, 185–195.

Corso, P.S., Kramer, M.H., Blair, K.A., Addiss, D.G., Davis, J.P., and Haddix, A.C. (2003). Cost of illness in the 1993 waterborne Cryptosporidium outbreak, Milwaukee, Wisconsin. Emerg. Infect. Dis. *9*, 426–431.

Daly, M.J. (2009). A new perspective on radiation resistance based on Deinococcus radiodurans. Nat. Rev. Microbiol. *7*, 237–245.

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A., Bik, H.M., and Eisen, J.A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ *2*, e243.

Davenport, C.F., Neugebauer, J., Beckmann, N., Friedrich, B., Kameri, B., Kokott, S., Paetow, M., Siekmann, B., Wieding-Drewes, M., Wienhöfer, M., et al. (2012). Genometa--a fast and accurate classifier for short metagenomic shotgun reads. PloS One *7*, e41224.

David, S., Mentasti, M., Tewolde, R., Aslett, M., Harris, S.R., Afshar, B., Underwood, A., Fry, N.K., Parkhill, J., and Harrison, T.G. (2016). Evaluation of an Optimal Epidemiological Typing Scheme for Legionella pneumophila with Whole-Genome Sequence Data Using Validation Guidelines. J. Clin. Microbiol. *54*, 2135–2148.

Deb, P., Talukdar, S.A., Mohsina, K., Sarker, P.K., and Sayem, S.A. (2013). Production and partial characterization of extracellular amylase enzyme from Bacillus amyloliquefaciens P-001. SpringerPlus *2*.

Declerck, P. (2010). Biofilms: the environmental playground of Legionella pneumophila. Environ. Microbiol. *12*, 557–566.

Delmont, T.O., Robe, P., Cecillon, S., Clark, I.M., Constancias, F., Simonet, P., Hirsch, P.R., and Vogel, T.M. (2011). Accessing the soil metagenome for studies of microbial diversity. Appl. Environ. Microbiol. *77*, 1315–1324.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.-U., Martinez, A., Sullivan, M.B., Edwards, R., Brito, B.R., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science *311*, 496–503.

Diaz, N.N., Krause, L., Goesmann, A., Niehaus, K., and Nattkemper, T.W. (2009). TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics *10*, 56.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science *308*, 1635–1638.

Eddy, S.R. (2008). A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS Comput. Biol. *4*, e1000069.

Edwards, R.A., Olson, R., Disz, T., Pusch, G.D., Vonstein, V., Stevens, R., and Overbeek, R. (2012). Real time metagenomics: using k-mers to annotate metagenomes. Bioinforma. Oxf. Engl. *28*, 3316–3317.

Ettinger, W.F., Thukral, S.K., and Kolattukudy, P.E. (1987). Structure of cutinase gene, cDNA, and the derived amino acid sequence from phytopathogenic fungi. Biochemistry (Mosc.) *26*, 7883–7892.

Euser, S.M., Pelgrim, M., and den Boer, J.W. (2010). Legionnaires' disease and Pontiac fever after using a private outdoor whirlpool spa. Scand. J. Infect. Dis. *42*, 910–916.

Fang, G.D., Yu, V.L., and Vickers, R.M. (1989). Disease due to the Legionellaceae (other than Legionella pneumophila). Historical, microbiological, clinical, and epidemiological review. Medicine (Baltimore) *68*, 116–132.

Fenwick, A. (2006). Waterborne infectious diseases--could they be consigned to history? Science *313*, 1077–1081.

Fields, B.S., Benson, R.F., and Besser, R.E. (2002). Legionella and Legionnaires' disease: 25 years of investigation. Clin. Microbiol. Rev. *15*, 506–526.

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

Fierer, N., Leff, J.W., Adams, B.J., Nielsen, U.N., Bates, S.T., Lauber, C.L., Owens, S., Gilbert, J.A., Wall, D.H., and Caporaso, J.G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci. U. S. A. *109*, 21390–21395.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., et al. (2010). The Pfam protein families database. Nucleic Acids Res. *38*, D211-222.

Fliermans, C.B., Cherry, W.B., Orrison, L.H., Smith, S.J., Tison, D.L., and Pope, D.H. (1981). Ecological distribution of Legionella pneumophila. Appl. Environ. Microbiol. *41*, 9–16.

Fortunato, C.S., Eiler, A., Herfort, L., Needoba, J.A., Peterson, T.D., and Crump, B.C. (2013). Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. ISME J. *7*, 1899–1911.

Francke, C., Groot Kormelink, T., Hagemeijer, Y., Overmars, L., Sluijter, V., Moezelaar, R., and Siezen, R.J. (2011). Comparative analyses imply that the enigmatic Sigma factor 54 is a central controller of the bacterial exterior. BMC Genomics *12*, 385.

Fraser, D.W., Tsai, T.R., Orenstein, W., Parkin, W.E., Beecham, H.J., Sharrar, R.G., Harris, J., Mallison, G.F., Martin, S.M., McDade, J.E., et al. (1977). Legionnaires' disease: description of an epidemic of pneumonia. N. Engl. J. Med. *297*, 1189–1197.

Freitas, T.A.K., Li, P.-E., Scholz, M.B., and Chain, P.S.G. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic Acids Res. *43*, e69–e69.

Frith, M.C., Hamada, M., and Horton, P. (2010). Parameters for accurate genome alignment. BMC Bioinformatics *11*, 80.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Fuka, M.M., Wallisch, S., Engel, M., Welzl, G., Havranek, J., and Schloter, M. (2013). Dynamics of Bacterial Communities during the Ripening Process of Different Croatian Cheese Types Derived from Raw Ewe's Milk Cheeses. PLoS ONE *8*.

Fukushima, M., Kakinuma, K., and Kawaguchi, R. (2002). Phylogenetic Analysis of Salmonella, Shigella, and Escherichia coli Strains on the Basis of the gyrB Gene Sequence. J. Clin. Microbiol. *40*, 2779–2785.

Gao, B., and Gupta, R.S. (2012). Phylogenetic framework and molecular signatures for the main clades of the phylum Actinobacteria. Microbiol. Mol. Biol. Rev. MMBR *76*, 66–112.

García-Armisen, T., İnceoğlu, Ö., Ouattara, N.K., Anzil, A., Verbanck, M.A., Brion, N., and Servais, P. (2014). Seasonal variations and resilience of bacterial communities in a sewage polluted urban river. PloS One *9*, e92579.

Garcia-Etxebarria, K., Garcia-Garcerà, M., and Calafell, F. (2014). Consistency of metagenomic assignment programs in simulated and real data. BMC Bioinformatics *15*, 90.

Gardy, J.L., and Brinkman, F.S.L. (2006). Methods for predicting bacterial protein subcellular localization. Nat. Rev. Microbiol. *4*, 741–751.

Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnády, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., et al. (2003). PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res. *31*, 3613–3617.

Gardy, J.L., Laird, M.R., Chen, F., Rey, S., Walsh, C.J., Ester, M., and Brinkman, F.S.L. (2005). PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics *21*, 617–623.

Gebhardt, H., Meniche, X., Tropis, M., Kramer, R., Daffe, M., and Morbach, S. (2007). The key role of the mycolic acid content in the functionality of the cell wall permeability barrier in Corynebacterineae. Microbiology *153*, 1424–1434.

Gentle, I., Gabriel, K., Beech, P., Waller, R., and Lithgow, T. (2004). The Omp85 family of proteins is essential for outer membrane biogenesis in mitochondria and bacteria. J. Cell Biol. *164*, 19–24.

Gerlach, W., and Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Res. *39*, e91.

Gerlach, W., Jünemann, S., Tille, F., Goesmann, A., and Stoye, J. (2009). WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. BMC Bioinformatics *10*, 430.

Ghai, R., Rodriguez-Valera, F., McMahon, K.D., Toyama, D., Rinke, R., Cristina Souza de Oliveira, T., Wagner Garcia, J., Pellon de Miranda, F., and Henrique-Silva, F. (2011). Metagenomics of the water column in the pristine upper course of the Amazon river. PloS One *6*, e23785.

Ghosh, T.S., Monzoorul Haque, M., and Mande, S.S. (2010). DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. BMC Bioinformatics *11 Suppl 7*, S14.

Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbrück, L., Reeder, J., Temperton, B., Huse, S., McHardy, A.C., Knight, R., Joint, I., et al. (2012). Defining seasonal marine microbial community dynamics. ISME J. *6*, 298–308.

Gilbert, J.A., Jansson, J.K., and Knight, R. (2014). The Earth Microbiome project: successes and aspirations. BMC Biol. *12*, 69.

Glick, T.H., Gregg, M.B., Berman, B., Mallison, G., Rhodes, W.W., and Kassanoff, I. (1978). Pontiac fever. An epidemic of unknown etiology in a health department: I. Clinical and epidemiologic aspects. Am. J. Epidemiol. *107*, 149–160.

Goldstein, S.T., Juranek, D.D., Ravenholt, O., Hightower, A.W., Martin, D.G., Mesnik, J.L., Griffiths, S.D., Bryant, A.J., Reich, R.R., and Herwaldt, B.L. (1996). Cryptosporidiosis: an outbreak associated with drinking water despite state-of-the-art water treatment. Ann. Intern. Med. *124*, 459–468.

Goudenège, D., Avner, S., Lucchetti-Miganeh, C., and Barloy-Hubler, F. (2010). CoBaltDB: Complete bacterial and archaeal orfeomes subcellular localization database and associated resources. BMC Microbiol. *10*, 88.

Graham, J.E., Clark, M.E., Nadler, D.C., Huffer, S., Chokhawala, H.A., Rowland, S.E., Blanch, H.W., Clark, D.S., and Robb, F.T. (2011). Identification and characterization of a multidomain hyperthermophilic cellulase from an archaeal enrichment. Nat. Commun. *2*, 375.

Griffiths, E., and Gupta, R.S. (2007). Identification of signature proteins that are distinctive of the Deinococcus-Thermus phylum. Int. Microbiol. Off. J. Span. Soc. Microbiol. *10*, 201–208.

Guinebretière, M.-H., Broussolle, V., and Nguyen-The, C. (2002). Enterotoxigenic profiles of food-poisoning and food-borne Bacillus cereus strains. J. Clin. Microbiol. *40*, 3053–3056.

Gupta, R.S. (2011). Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes. Antonie Van Leeuwenhoek *100*, 171–182.

Gupta, R.S., and Bhandari, V. (2011). Phylogeny and molecular signatures for the phylum Thermotogae and its subgroups. Antonie Van Leeuwenhoek *100*, 1–34.

Hahn, M.W., Kasalický, V., Jezbera, J., Brandt, U., Jezberová, J., and Šimek, K. (2010). Limnohabitans curvus gen. nov., sp. nov., a planktonic bacterium isolated from a freshwater lake. Int. J. Syst. Evol. Microbiol. *60*, 1358–1365.

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. Microbiol. Mol. Biol. Rev. *68*, 669–685.

Harwood, V.J., Levine, A.D., Scott, T.M., Chivukula, V., Lukasik, J., Farrah, S.R., and Rose, J.B. (2005). Validity of the Indicator Organism Paradigm for Pathogen Reduction in Reclaimed Water and Public Health Protection. Appl. Environ. Microbiol. *71*, 3163–3170.

Harwood, V.J., Staley, C., Badgley, B.D., Borges, K., and 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.

van der Heijden, M.G.A., Bardgett, R.D., and van Straalen, N.M. (2008). The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol. Lett. *11*, 296–310.

Henson, S.J., Majowicz, S.E., Masakure, O., Sockett, P.N., MacDougall, L., Edge, V.L., Thomas, M.K., Fyfe, M., Kovacs, S.J., and Jones, A.Q. (2008). Estimation of the costs of acute gastrointestinal illness in British Columbia, Canada. Int. J. Food Microbiol. *127*, 43–52.

Hoffmann, C., Leis, A., Niederweis, M., Plitzko, J.M., and Engelhardt, H. (2008). Disclosure of the mycobacterial outer membrane: cryo-electron tomography and vitreous sections reveal the lipid bilayer structure. Proc. Natl. Acad. Sci. U. S. A. *105*, 3963–3967.

Horwitz, M.A. (1984). Phagocytosis of the Legionnaires' disease bacterium (Legionella pneumophila) occurs by a novel mechanism: engulfment within a pseudopod coil. Cell *36*, 27–33.

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

Huber, L.A., Pfaller, K., and Vietor, I. (2003). Organelle proteomics: implications for subcellular fractionation in proteomics. Circ. Res. *92*, 962–968.

Hudault, S., Guignot, J., and Servin, A.L. (2001). Escherichia coli strains colonising the gastrointestinal tract protect germfree mice against Salmonella typhimurium infection. Gut *49*, 47–55.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. J. Bacteriol. *180*, 4765–4774.

Human Microbiome Jumpstart Reference Strains Consortium, Nelson, K.E., Weinstock, G.M., Highlander, S.K., Worley, K.C., Creasy, H.H., Wortman, J.R., Rusch, D.B., Mitreva, M., Sodergren, E., et al. (2010). A catalog of reference genomes from the human microbiome. Science *328*, 994–999.

Huson, D.H., and Xie, C. (2014). A poor man's BLASTX—high-throughput metagenomic protein database search using PAUDA. Bioinformatics *30*, 38–39.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome Res. *17*, 377–386.

Huson, D.H., Richter, D.C., Mitra, S., Auch, A.F., and Schuster, S.C. (2009). Methods for comparative metagenomics. BMC Bioinformatics *10 Suppl 1*, S12.

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Res. *21*, 1552–1560.

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Comput. Biol. *12*, e1004957.

Hyatt, D., LoCascio, P.F., Hauser, L.J., and Uberbacher, E.C. (2012). Gene and translation initiation site prediction in metagenomic sequences. Bioinforma. Oxf. Engl. *28*, 2223–2230.

Imai, K., Asakawa, N., Tsuji, T., Akazawa, F., Ino, A., Sonoyama, M., and Mitaku, S. (2008). SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in gram-negative bacteria. Bioinformation *2*, 417–421.

Isberg, R.R., O'Connor, T., and Heidtman, M. (2009). The Legionella pneumophila replication vacuole: making a cozy niche inside host cells. Nat. Rev. Microbiol. *7*, 13–24.

Ishii, S., Ksoll, W.B., Hicks, R.E., and Sadowsky, M.J. (2006). Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds. Appl. Environ. Microbiol. *72*, 612–621.

Janda, J.M., and Abbott, S.L. (2007). 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. J. Clin. Microbiol. *45*, 2761–2764.

J. R. Bray, J.T.C. (1957). An ordination of the upland forest communities of southern Wisconsin. *27*, 325–349.

Jiang, H., An, L., Lin, S.M., Feng, G., and Qiu, Y. (2012). A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. PloS One *7*, e46450.

Jones, D.S., Podolsky, S.H., and Greene, J.A. (2012). The burden of disease and the changing task of medicine. N. Engl. J. Med. *366*, 2333–2338.

Kagan, J.C., and Roy, C.R. (2002). Legionella phagosomes intercept vesicular traffic from endoplasmic reticulum exit sites. Nat. Cell Biol. *4*, 945–954.

Kao, P.-M., Tung, M.-C., Hsu, B.-M., Chiu, Y.-C., She, C.-Y., Shen, S.-M., Huang, Y.-L., and Huang, W.-C. (2013). Identification and quantitative detection of Legionella spp. in various aquatic environments by real-time PCR assay. Environ. Sci. Pollut. Res. Int. *20*, 6128–6137.

Karlsson, F.H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C.J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. Nature *498*, 99–103.

Kasalický, V., Jezbera, J., Hahn, M.W., and Šimek, K. (2013). The Diversity of the Limnohabitans Genus, an Important Group of Freshwater Bacterioplankton, by Characterization of 35 Isolated Strains. PLOS ONE *8*, e58209.

Kerepesi, C., Bánky, D., and Grolmusz, V. (2014). AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. Gene *533*, 538–540.

Klingenberg, H., Aßhauer, K.P., Lingner, T., and Meinicke, P. (2013). Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. Bioinforma. Oxf. Engl. *29*, 973–980.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J.A., Hugenholtz, P., van der Lelie, D., Meyer, F., Stevens, R., et al. (2012). Unlocking the potential of metagenomics through replicated experimental design. Nat. Biotechnol. *30*, 513–520.

Korajkic, A., McMinn, B.R., Shanks, O.C., Sivaganesan, M., Fout, G.S., and Ashbolt, N.J. (2014). Biotic interactions and sunlight affect persistence of fecal indicator bacteria and microbial source tracking genetic markers in the upper Mississippi river. Appl. Environ. Microbiol. *80*, 3952–3961.

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

Krause, L., Diaz, N.N., Goesmann, A., Kelley, S., Nattkemper, T.W., Rohwer, F., Edwards, R.A., and Stoye, J. (2008). Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. *36*, 2230–2239.

Kusters, J.G., van Vliet, A.H.M., and Kuipers, E.J. (2006). Pathogenesis of Helicobacter pylori Infection. Clin. Microbiol. Rev. *19*, 449–490.

Lane, D.J., Pace, B., Olsen, G.J., Stahl, D.A., Sogin, M.L., and Pace, N.R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Natl. Acad. Sci. U. S. A. *82*, 6955–6959.

Langille, M.G.I., Laird, M.R., Hsiao, W.W.L., Chiu, T.A., Eisen, J.A., and Brinkman, F.S.L. (2012). MicrobeDB: a locally maintainable database of microbial genomic sequences. Bioinformatics *28*, 1947–1948.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. *31*, 814–821.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Layton, B.A., Cao, Y., Ebentier, D.L., Hanley, K., Ballesté, E., Brandão, J., Byappanahalli, M., Converse, R., Farnleitner, A.H., Gentry-Shields, J., et al. (2013). Performance of human fecal anaerobe-associated PCR-based assays in a multi-laboratory method evaluation study. Water Res. *47*, 6897–6908.

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., et al. (2013). Richness of human gut microbiome correlates with metabolic markers. Nature *500*, 541–546.

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

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

Lin, J., and Ganesh, A. (2013). Water quality indicators: bacteria, coliphages, enteric viruses. Int. J. Environ. Health Res. *23*, 484–506.

Lindgreen, S., Adair, K.L., and Gardner, P.P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. Sci. Rep. *6*, 19233.

Litou, Z.I., Bagos, P.G., Tsirigos, K.D., Liakopoulos, T.D., and Hamodrakas, S.J. (2008). Prediction of cell wall sorting signals in gram-positive bacteria with a hidden markov model: application to complete genomes. J. Bioinform. Comput. Biol. *6*, 387–401.

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics *12*, S4.

Liu, J., Wang, H., Yang, H., Zhang, Y., Wang, J., Zhao, F., and Qi, J. (2013). Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. Nucleic Acids Res. *41*, e3.

Liu, J., Fu, B., Yang, H., Zhao, M., He, B., and Zhang, X.-H. (2015). Phylogenetic shifts of bacterioplankton community composition along the Pearl Estuary: the potential impact of hypoxia and nutrients. Front. Microbiol. *6*, 64.

Luo, H., Benner, R., Long, R.A., and Hu, J. (2009). Subcellular localization of marine bacterial alkaline phosphatases. Proc. Natl. Acad. Sci. U. S. A. *106*, 21219–21223.

MacDonald, N.J., Parks, D.H., and Beiko, R.G. (2012). Rapid identification of high-confidence taxonomic assignments for metagenomic data. Nucleic Acids Res. *40*, e111.

Mägli, A., Wendt, M., and Leisinger, T. (1996). Isolation and characterization of Dehalobacterium formicoaceticum gen. nov. sp. nov., a strictly anaerobic bacterium utilizing dichloromethane as source of carbon and energy. Arch. Microbiol. *166*, 101–108.

Magnus, M., Pawlowski, M., and Bujnicki, J.M. (2012). MetaLocGramN: A meta-predictor of protein subcellular localization for Gram-negative bacteria. Biochim. Biophys. Acta *1824*, 1425–1433.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

Maione, D., Margarit, I., Rinaudo, C.D., Masignani, V., Mora, M., Scarselli, M., Tettelin, H., Brettoni, C., Iacobini, E.T., Rosini, R., et al. (2005). Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science *309*, 148–150.

Mande, S.S., Mohammed, M.H., and Ghosh, T.S. (2012). Classification of metagenomic sequences: methods and challenges. Brief. Bioinform.

Marchand, C.H., Salmeron, C., Bou Raad, R., Méniche, X., Chami, M., Masi, M., Blanot, D., Daffé, M., Tropis, M., Huc, E., et al. (2012). Biochemical disclosure of the mycolate outer membrane of Corynebacterium glutamicum. J. Bacteriol. *194*, 587–597.

Marston, B.J., Lipman, H.B., and Breiman, R.F. (1994). Surveillance for Legionnaires' disease. Risk factors for morbidity and mortality. Arch. Intern. Med. *154*, 2417–2422.

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

McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., et al. (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience *1*, 7.

McLellan, S.L., Newton, R.J., Vandewalle, J.L., Shanks, O.C., Huse, S.M., Eren, A.M., and Sogin, M.L. (2013). Sewage reflects the distribution of human faecal Lachnospiraceae. Environ. Microbiol. *15*, 2213–2227.

Méndez-García, C., Mesa, V., Sprenger, R.R., Richter, M., Diez, M.S., Solano, J., Bargiela, R., Golyshina, O.V., Manteca, Á., Ramos, J.L., et al. (2014). Microbial stratification in low pH oxic and suboxic macroscopic growths along an acid mine drainage. ISME J. *8*, 1259–1274.

Meniche, X., Labarre, C., de Sousa-d'Auria, C., Huc, E., Laval, F., Tropis, M., Bayan, N., Portevin, D., Guilhot, C., Daffé, M., et al. (2009). Identification of a stress-induced factor of Corynebacterineae that is involved in the regulation of the outer membrane lipid composition. J. Bacteriol. *191*, 7323–7332.

Mercante, J.W., and Winchell, J.M. (2015). Current and Emerging Legionella Diagnostics for Laboratory and Outbreak Investigations. Clin. Microbiol. Rev. *28*, 95–133.

von Mering, C., Hugenholtz, P., Raes, J., Tringe, S.G., Doerks, T., Jensen, L.J., Ward, N., and Bork, P. (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. Science *315*, 1126–1130.

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

Mitra, S., Klar, B., and Huson, D.H. (2009). Visual and statistical comparison of metagenomes. Bioinforma. Oxf. Engl. *25*, 1849–1855.

Mitra, S., Rupek, P., Richter, D.C., Urich, T., Gilbert, J.A., Meyer, F., Wilke, A., and Huson, D.H. (2011). Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. BMC Bioinformatics *12 Suppl 1*, S21.

Miyata, M., and Ogaki, H. (2006). Cytoskeleton of mollicutes. J. Mol. Microbiol. Biotechnol. *11*, 256–264.

Mohammed, M.H., Ghosh, T.S., Singh, N.K., and Mande, S.S. (2011a). SPHINX--an algorithm for taxonomic binning of metagenomic sequences. Bioinforma. Oxf. Engl. *27*, 22–30.

Mohammed, M.H., Ghosh, T.S., Reddy, R.M., Reddy, C.V.S.K., Singh, N.K., and Mande, S.S. (2011b). INDUS - a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. BMC Genomics *12 Suppl 3*, S4.

Molohon, K.J., Melby, J.O., Lee, J., Evans, B.S., Dunbar, K.L., Bumpus, S.B., Kelleher, N.L., and Mitchell, D.A. (2011). Structure Determination and Interception of Biosynthetic Intermediates for the Plantazolicin Class of Highly Discriminating Antibiotics. ACS Chem. Biol. *6*, 1307–1313.

Monzoorul Haque, M., Ghosh, T.S., Komanduri, D., and Mande, S.S. (2009). SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinforma. Oxf. Engl. *25*, 1722–1730.

Morris, G.K., Patton, C.M., Feeley, J.C., Johnson, S.E., Gorman, G., Martin, W.T., Skaliy, P., Mallison, G.F., Politi, B.D., and Mackel, D.C. (1979). Isolation of the Legionnaires' disease bacterium from environmental samples. Ann. Intern. Med. *90*, 664–666.

Muder, R.R., and Yu, V.L. (2002). Infection due to Legionella species other than L. pneumophila. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. *35*, 990–998.

Muyzer, G., De Waal, E.C., and Uitterlinden, A.G. (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.

Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. *24*, 34–36.

Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. Proteins *11*, 95–110.

Nalbantoglu, O.U., Way, S.F., Hinrichs, S.H., and Sayood, K. (2011). RAIphy: Phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics *12*, 41.

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

Neufeld, J.D., and Mohn, W.W. (2005). Unexpectedly High Bacterial Diversity in Arctic Tundra Relative to Boreal Forest Soils, Revealed by Serial Analysis of Ribosomal Sequence Tags. Appl. Environ. Microbiol. *71*, 5710–5718.

Neufeld, J.D., Engel, K., Cheng, J., Moreno-Hagelsieb, G., Rose, D.R., and Charles, T.C. (2011). Open resource metagenomics: a model for sharing metagenomic libraries. Stand. Genomic Sci. *5*, 203.

Newton, R.J., Vandewalle, J.L., Borchardt, M.A., Gorelick, M.H., and McLellan, S.L. (2011). Lachnospiraceae and Bacteroidales alternative fecal indicators reveal chronic human sewage contamination in an urban harbor. Appl. Environ. Microbiol. *77*, 6972–6981.

Newton, R.J., Bootsma, M.J., Morrison, H.G., Sogin, M.L., and McLellan, S.L. (2013). A microbial signature approach to identify fecal pollution in the waters off an urbanized coast of Lake Michigan. Microb. Ecol. *65*, 1011–1023.

Nguyen, K.H., and Grove, A. (2012). Metal binding at the Deinococcus radiodurans Dps-1 N-terminal metal site controls dodecameric assembly and DNA binding. Biochemistry (Mosc.) *51*, 6679–6689.

Nguyen, Y., and Sperandio, V. (2012). Enterohemorrhagic E. coli (EHEC) pathogenesis. Front. Cell. Infect. Microbiol. *2*.

Niederweis, M., Danilchanka, O., Huff, J., Hoffmann, C., and Engelhardt, H. (2010). Mycobacterial outer membranes: in search of proteins. Trends Microbiol. *18*, 109–116.

NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J.A., Bonazzi, V., McEwen, J.E., Wetterstrand, K.A., et al. (2009). The NIH Human Microbiome Project. Genome Res. *19*, 2317–2323.

Nishida, H., Beppu, T., and Ueda, K. (2011). Whole-genome comparison clarifies close phylogenetic relationships between the phyla Dictyoglomi and Thermotogae. Genomics *98*, 370–375.

Niu, B., Jin, Y.-H., Feng, K.-Y., Lu, W.-C., Cai, Y.-D., and Li, G.-Z. (2008). Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol. Divers. *12*, 41–45.

Niu, B., Zhu, Z., Fu, L., Wu, S., and Li, W. (2011). FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. Bioinformatics *27*, 1704–1705.

Oh, S., Caro-Quintero, A., Tsementzi, D., DeLeon-Rodriguez, N., Luo, C., Poretsky, R., and Konstantinidis, K.T. (2011). Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. Appl. Environ. Microbiol. *77*, 6000–6011.

Okabe, S., Okayama, N., Savichtcheva, O., and Ito, T. (2007). Quantification of host-specific Bacteroides-Prevotella 16S rRNA genetic markers for assessment of fecal pollution in freshwater. Appl. Microbiol. Biotechnol. *74*, 890–901.

Oksanen, J., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., and Wagner, H. (2015). vegan: Community Ecology Package.

Økstad, O.A., and Kolstø, A.-B. (2011). Genomics of Bacillus Species. In Genomics of Foodborne Bacterial Pathogens, M. Wiedmann, and W. Zhang, eds. (Springer New York), pp. 29–53.

Ortiz-Roque, C.M., and Hazen, T.C. (1987). Abundance and distribution of Legionellaceae in Puerto Rican waters. Appl. Environ. Microbiol. *53*, 2231–2236.

Ounit, R., Wanamaker, S., Close, T.J., and Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics *16*, 236.

Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., et al. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. *33*, 5691–5702.

Pagé, A., Juniper, S.K., Olagnon, M., Alain, K., Desrosiers, G., Quérellou, J., and Cambon-Bonavita, M.-A. (2004). Microbial diversity associated with a Paralvinella sulfincola tube and the adjacent substratum on an active deep-sea vent chimney. Geobiology *2*, 225–238.

Paramasivam, N., and Linke, D. (2011). ClubSub-P: Cluster-Based Subcellular Localization Prediction for Gram-Negative Bacteria and Archaea. Front. Microbiol. *2*.

Parker, S.K., Barkley, R.M., Rino, J.G., and Vasil, M.L. (2009). Mycobacterium tuberculosis Rv3802c encodes a phospholipase/thioesterase and is inhibited by the antimycobacterial agent tetrahydrolipstatin. PloS One *4*, e4281.

Parks, D.H., MacDonald, N.J., and Beiko, R.G. (2011). Classifying short genomic fragments from novel lineages using composition and homology. BMC Bioinformatics *12*, 328.

Parthuisot, N., West, N.J., Lebaron, P., and Baudart, J. (2010). High diversity and abundance of Legionella spp. in a pristine river and impact of seasonal and anthropogenic effects. Appl. Environ. Microbiol. *76*, 8201–8210.

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

Patil, K.R., Haider, P., Pope, P.B., Turnbaugh, P.J., Morrison, M., Scheffer, T., and McHardy, A.C. (2011). Taxonomic metagenome sequence assignment with structured output models. Nat. Methods *8*, 191–192.

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

Peabody, M.A., Rossum, T.V., Lo, R., and Brinkman, F.S. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC Bioinformatics *16*, 363.

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

Petersen, T.N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods *8*, 785–786.

Pible, O., Hartmann, E.M., Imbert, G., and Armengaud, J. (2014). The importance of recognizing and reporting sequence database contamination for proteomics. EuPA Open Proteomics *3*, 246–249.

Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R.D.E., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., et al. (2006). Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science *311*, 392–394.

Poretsky, R., Rodriguez-R, L.M., Luo, C., Tsementzi, D., and Konstantinidis, K.T. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PloS One *9*, e93827.

Porter, M.S., and Beiko, R.G. (2013). SPANNER: taxonomic assignment of sequences using pyramid matching of similarity profiles. Bioinformatics *29*, 1858–1864.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35*, D61-65.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature *490*, 55–60.

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. (2013). A large-scale evaluation of computational protein function prediction. Nat. Methods *10*, 221–227.

Ramulu, H.G., Swathi, A., and Guruprasad, L. (2006). The Rv3799-Rv3807 gene cluster in Mycobacterium tuberculosis genome corresponds to the “Ancient Conserved Region” in CMN mycolyltransferases. Evol. Bioinforma. Online *2*, 377–385.

Rappé, M.S., and Giovannoni, S.J. (2003). The uncultured microbial majority. Annu. Rev. Microbiol. *57*, 369–394.

Rappé, M.S., Connon, S.A., Vergin, K.L., and Giovannoni, S.J. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. Nature *418*, 630–633.

Rasheed, Z., and Rangwala, H. (2012). Metagenomic taxonomic classification using extreme learning machines. J. Bioinform. Comput. Biol. *10*, 1250015.

Rashid, M., Saha, S., and Raghava, G.P. (2007). Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs. BMC Bioinformatics *8*, 337.

Ratcliff, R.M., Lanser, J.A., Manning, P.A., and Heuzenroeder, M.W. (1998). Sequence-based classification scheme for the genus Legionella targeting the mip gene. J. Clin. Microbiol. *36*, 1560–1567.

Read, D.S., Gweon, H.S., Bowes, M.J., Newbold, L.K., Field, D., Bailey, M.J., and Griffiths, R.I. (2015). Catchment-scale biogeography of riverine bacterioplankton. ISME J. *9*, 516–526.

Reddy, R.M., Mohammed, M.H., and Mande, S.S. (2012). TWARIT: an extremely rapid and efficient approach for phylogenetic classification of metagenomic sequences. Gene *505*, 259–265.

Reingold, A.L., Thomason, B.M., Brake, B.J., Thacker, L., Wilkinson, H.W., and Kuritsky, J.N. (1984). Legionella pneumonia in the United States: the distribution of serogroups and species causing human illness. J. Infect. Dis. *149*, 819.

Reinhardt, A., and Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res. *26*, 2230–2236.

Renier, S., Micheau, P., Talon, R., Hébraud, M., and Desvaux, M. (2012). Subcellular localization of extracytoplasmic proteins in monoderm bacteria: rational secretomics-based strategy for genomic and proteomic analyses. PloS One *7*, e42982.

Rey, S., Acab, M., Gardy, J.L., Laird, M.R., deFays, K., Lambert, C., and Brinkman, F.S.L. (2005a). PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res. *33*, D164–D168.

Rey, S., Gardy, J.L., and Brinkman, F.S. (2005b). Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. BMC Genomics *6*, 162.

Richter, D.C., Ott, F., Auch, A.F., Schmid, R., and Huson, D.H. (2008). MetaSim—A Sequencing Simulator for Genomics and Metagenomics. PLoS ONE *3*, e3373.

Riederer, M., and Schönherr, J. (1988). Development of plant cuticles: fine structure and cutin composition of Clivia miniata Reg. leaves. Planta *174*, 127–138.

Robert Huber, T.A.L. (1986). Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90 °C. Arch Microbiol. Arch. Microbiol. *144*, 324–333.

Rodríguez-Ortega, M.J., Norais, N., Bensi, G., Liberatori, S., Capo, S., Mora, M., Scarselli, M., Doro, F., Ferrari, G., Garaguso, I., et al. (2006). Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. Nat. Biotechnol. *24*, 191–197.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Yin, J., and Koonin, E.V. (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. *30*, 4264–4271.

Romero-Saavedra, F., Laverde, D., Wobser, D., Michaux, C., Budin-Verneuil, A., Bernay, B., Benachour, A., Hartke, A., and Huebner, J. (2014). Identification of peptidoglycan-associated proteins as vaccine candidates for enterococcal infections. PloS One *9*, e111880.

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., and Sokhansanj, B. (2008). Metagenome fragment classification using N-mer frequency profiles. Adv. Bioinforma. *2008*, 205969.

Rosen, G.L., Reichenberger, E.R., and Rosenfeld, A.M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinforma. Oxf. Engl. *27*, 127–129.

Rowbotham, T.J. (1980). Preliminary report on the pathogenicity of Legionella pneumophila for freshwater and soil amoebae. J. Clin. Pathol. *33*, 1179–1183.

Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., et al. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. *5*, e77.

Salvadori, M.I., Sontrop, J.M., Garg, A.X., Moist, L.M., Suri, R.S., and Clark, W.F. (2009). Factors that led to the Walkerton tragedy. Kidney Int. Suppl. S33-34.

Sánchez-Busó, L., Comas, I., Jorques, G., and González-Candelas, F. (2014). Recombination drives genome evolution in outbreak-related Legionella pneumophila isolates. Nat. Genet. *46*, 1205–1211.

Saravanan, P., Avinash, H., Dubey, V.K., and Patra, S. (2012). Targeting essential cell wall lipase Rv3802c for potential therapeutics against tuberculosis. J. Mol. Graph. Model. *38*, 235–242.

Sassetti, C.M., Boyd, D.H., and Rubin, E.J. (2003). Genes required for mycobacterial growth defined by high density mutagenesis. Mol. Microbiol. *48*, 77–84.

Sassoubre, L.M., Yamahara, K.M., and Boehm, A.B. (2015). Temporal stability of the microbial community in sewage-polluted seawater exposed to natural sunlight cycles and marine microbiota. Appl. Environ. Microbiol. AEM.03950-14.

Schallmey, M., Singh, A., and Ward, O.P. (2004). Developments in the use of Bacillus species for industrial production. Can. J. Microbiol. *50*, 1–17.

Scher, J.U., Sczesnak, A., Longman, R.S., Segata, N., Ubeda, C., Bielski, C., Rostron, T., Cerundolo, V., Pamer, E.G., Abramson, S.B., et al. (2013). Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. eLife *2*, e01202.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R. a, Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Schmidt, T.M., DeLong, E.F., and Pace, N.R. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. *173*, 4371–4378.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical Assessment of Metagenome Interpretation − a benchmark of computational metagenomics software. bioRxiv 99127.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. Genome Biol. *12*, R60.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods *9*, 811–814.

Sender, R., Fuchs, S., and Milo, R. (2016). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell *164*, 337–340.

Seshadri, R., Kravitz, S.A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: a community resource for metagenomics. PLoS Biol. *5*, e75.

Severin, A., Nickbarg, E., Wooters, J., Quazi, S.A., Matsuka, Y.V., Murphy, E., Moutsatsos, I.K., Zagursky, R.J., and Olmsted, S.B. (2007). Proteomic Analysis and Identification of Streptococcus pyogenes Surface-Associated Proteins. J. Bacteriol. *189*, 1514–1522.

Shakya, M., Quince, C., Campbell, J.H., Yang, Z.K., Schadt, C.W., and Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. Environ. Microbiol. *15*, 1882–1899.

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 379–423.

Sharma, V.K., Kumar, N., Prakash, T., and Taylor, T.D. (2012). Fast and accurate taxonomic assignments of metagenomic sequences using MetaBin. PloS One *7*, e34030.

Shen, H.-B., and Chou, K.-C. (2009). Gpos-mPLoc: a top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins. Protein Pept. Lett. *16*, 1478–1484.

Shen, H.-B., and Chou, K.-C. (2010). Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. J. Theor. Biol. *264*, 326–333.

Shi, J., Pagliaccia, D., Morgan, R., Qiao, Y., Pan, S., Vidalakis, G., and Ma, W. (2014). Novel diagnosis for citrus stubborn disease by detection of a spiroplasma citri-secreted protein. Phytopathology *104*, 188–195.

Simonen, M., and Palva, I. (1993). Protein secretion in Bacillus species. Microbiol. Rev. *57*, 109–137.

Sinclair, L., Osman, O.A., Bertilsson, S., and Eiler, A. (2015). Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform. PloS One *10*, e0116955.

Smith, I. (2003). Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence. Clin. Microbiol. Rev. *16*, 463–496.

Smith, R.J., Jeffries, T.C., Roudnew, B., Fitch, A.J., Seymour, J.R., Delpin, M.W., Newton, K., Brown, M.H., and Mitchell, J.G. (2012). Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems. Environ. Microbiol. *14*, 240–253.

Song, H., Sandie, R., Wang, Y., Andrade-Navarro, M.A., and Niederweis, M. (2008). Identification of outer membrane proteins of Mycobacterium tuberculosis. Tuberc. Edinb. Scotl. *88*, 526–544.

Spencer, R.C. (2003). Bacillus anthracis. J. Clin. Pathol. *56*, 182–187.

Staley, C., Gould, T.J., Wang, P., Phillips, J., Cotner, J.B., and Sadowsky, M.J. (2014). Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. Aquat. Microbiol. *5*, 524.

Stark, M., Berger, S.A., Stamatakis, A., and von Mering, C. (2010). MLTreeMap--accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. BMC Genomics *11*, 461.

Stelma, G.N., and Wymer, L.J. (2012). Research considerations for more effective groundwater monitoring. J. Water Health *10*, 511–521.

Su, E.C.-Y., Chiu, H.-S., Lo, A., Hwang, J.-K., Sung, T.-Y., and Hsu, W.-L. (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics *8*, 330.

Sun, C., Zhao, X.-M., Tang, W., and Chen, L. (2010). FGsub: Fusarium graminearum protein subcellular localizations predicted from primary structures. BMC Syst. Biol. *4 Suppl 2*, S12.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E.E., Ellisman, M., Grethe, J., et al. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic Acids Res. *39*, D546-551.

Sunagawa, S., Mende, D.R., Zeller, G., Izquierdo-Carrasco, F., Berger, S.A., Kultima, J.R., Coelho, L.P., Arumugam, M., Tap, J., Nielsen, H.B., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. Nat. Methods *10*, 1196–1199.

Sutcliffe, I.C. (2010). A phylum level perspective on bacterial cell envelope architecture. Trends Microbiol. *18*, 464–470.

Swanson, M.S., and Hammer, B.K. (2000). Legionella pneumophila pathogesesis: a fateful journey from amoebae to macrophages. Annu. Rev. Microbiol. *54*, 567–613.

Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Poulin, B., Eisner, R., Lu, Z., Anvik, J., Macdonell, C., Fyshe, A., et al. (2004). Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. Nucleic Acids Res. *32*, W365-371.

Takayama, K., Wang, C., and Besra, G.S. (2005). Pathway to Synthesis and Processing of Mycolic Acids in Mycobacterium tuberculosis. Clin. Microbiol. Rev. *18*, 81–101.

Taylor, M., Ross, K., and Bentham, R. (2009). Legionella, protozoa, and biofilms: interactions within complex microbial systems. Microb. Ecol. *58*, 538–547.

Thompson, B.G., and Murray, R.G. (1981). Isolation and characterization of the plasma membrane and the outer membrane of Deinococcus radiodurans strain Sark. Can. J. Microbiol. *27*, 729–734.

Tison, D.L., Baross, J.A., and Seidler, R.J. (1983). Legionella in aquatic habitats in the Mount Saint Helens blast zone. Curr. Microbiol. *9*, 345–348.

Tremlett, H., Bauer, K.C., Appel-Cresswell, S., Finlay, B.B., and Waubant, E. (2017). The gut microbiome in human neurological disease: A review. Ann. Neurol. *81*, 369–382.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature *449*, 804–810.

Turnbaugh, P.J., Hamady, M., Yatsunenko, T., Cantarel, B.L., Duncan, A., Ley, R.E., Sogin, M.L., Jones, W.J., Roe, B.A., Affourtit, J.P., et al. (2009). A core gut microbiome in obese and lean twins. Nature *457*, 480–484.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature *428*, 37–43.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. *43*, D204-212.

Uthandi, S., Saad, B., Humbard, M.A., and Maupin-Furlow, J.A. (2010). LccA, an Archaeal Laccase Secreted as a Highly Stable Glycoprotein into the Extracellular Medium by Haloferax volcanii. Appl. Environ. Microbiol. *76*, 733–743.

Uyaguari-Diaz, M.I., Slobodan, J.R., Nesbitt, M.J., Croxen, M.A., Isaac-Renton, J., Prystajecky, N.A., and Tang, P. (2015). Automated Gel Size Selection to Improve the Quality of Next-generation Sequencing Libraries Prepared from Environmental Water Samples. J. Vis. Exp. JoVE.

Uyaguari-Diaz, M.I., Chan, M., Chaban, B.L., Croxen, M.A., Finke, J.F., Hill, J.E., Peabody, M.A., Van Rossum, T., Suttle, C.A., Brinkman, F.S.L., et al. (2016). A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. Microbiome *4*, 20.

Van Rossum, T., Peabody, M.A., Uyaguari-Diaz, M.I., Cronin, K.I., Chan, M., Slobodan, J.R., Nesbitt, M.J., Suttle, C.A., Hsiao, W.W.L., Tang, P.K.C., et al. (2015). Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. Front. Microbiol. *6*, 1405.

Vasantha, N., Thompson, L.D., Rhodes, C., Banner, C., Nagle, J., and Filpula, D. (1984). Genes for alkaline protease and neutral protease from Bacillus amyloliquefaciens contain a large open reading frame between the regions coding for signal sequence and mature protein. J. Bacteriol. *159*, 811–819.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science *304*, 66–74.

Větrovský, T., and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PloS One *8*, e57923.

Vogel, U., and Claus, H. (2011). Vaccine development against Neisseria meningitidis. Microb. Biotechnol. *4*, 20–31.

Vogel, T.M., Simonet, P., Jansson, J.K., Hirsch, P.R., Tiedje, J.M., van Elsas, J.D., Bailey, M.J., Nalin, R., and Philippot, L. (2009). TerraGenome: a consortium for the sequencing of a soil metagenome. Nat. Rev. Microbiol. *7*, 252–252.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., et al. (2010). Global threats to human water security and river biodiversity. Nature *467*, 555–561.

Voulhoux, R., Bos, M.P., Geurtsen, J., Mols, M., and Tommassen, J. (2003). Role of a Highly Conserved Bacterial Protein in Outer Membrane Protein Assembly. Science *299*, 262–265.

Waites, K.B., and Talkington, D.F. (2004). Mycoplasma pneumoniae and Its Role as a Human Pathogen. Clin. Microbiol. Rev. *17*, 697–728.

Walters, W.A., Caporaso, J.G., Lauber, C.L., Berg-Lyons, D., Fierer, N., and Knight, R. (2011). PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. Bioinformatics *27*, 1159–1161.

Wang, X., Zhang, J., and Li, G.-Z. (2015). Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC Bioinformatics *16 Suppl 12*, S1.

Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., Dugar, B., Feldstein, A.E., Britt, E.B., Fu, X., Chung, Y.-M., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature *472*, 57–63.

West, N.P., Chow, F.M.E., Randall, E.J., Wu, J., Chen, J., Ribeiro, J.M.C., and Britton, W.J. (2009). Cutinase-like proteins of Mycobacterium tuberculosis: characterization of their variable enzymatic functions and active site identification. FASEB J. *23*, 1694–1704.

West, N.P., Cergol, K.M., Xue, M., Randall, E.J., Britton, W.J., and Payne, R.J. (2011). Inhibitors of an essential mycobacterial cell wall lipase (Rv3802c) as tuberculosis drug leads. Chem. Commun. Camb. Engl. *47*, 5166–5168.

Wheeler Alm, E., Burke, J., and Spain, A. (2003). Fecal indicator bacteria are abundant in wet sand at freshwater beaches. Water Res. *37*, 3978–3982.

Whiteside, M.D., Winsor, G.L., Laird, M.R., and Brinkman, F.S.L. (2013). OrthologeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. Nucleic Acids Res. *41*, D366-376.

Whitman, W., Goodfellow, M., Kämpfer, P., Busse, H.-J., Trujillo, M., Ludwig, W., and Suzuki, K. (2012). Bergey's Manual of Systematic Bacteriology: Volume 5: The Actinobacteria (Springer Science & Business Media).

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998). Prokaryotes: The unseen majority. Proc. Natl. Acad. Sci. *95*, 6578–6583.

Wilkes, G., Edge, T., Gannon, V., Jokinen, C., Lyautey, E., Medeiros, D., Neumann, N., Ruecker, N., Topp, E., and Lapen, D.R. (2009). Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia cysts, and hydrological indices for surface waters within an agricultural landscape. Water Res. *43*, 2209–2223.

Winsor, G.L., Griffiths, E.J., Lo, R., Dhillon, B.K., Shay, J.A., and Brinkman, F.S.L. (2016). Enhanced annotations and features for comparing thousands of Pseudomonas genomes in the Pseudomonas genome database. Nucleic Acids Res. *44*, D646-653.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. *74*, 5088–5090.

Wolfe, B.E., and Dutton, R.J. (2015). Fermented foods as experimentally tractable microbial ecosystems. Cell *161*, 49–55.

Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. *15*, R46.

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

Wu, M., and Eisen, J.A. (2008). A simple, fast, and accurate method of phylogenomic inference. Genome Biol. *9*, R151.

Wu, M., and Scott, A.J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinforma. Oxf. Engl. *28*, 1033–1034.

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics *12*, 444.

Xia, X. (2013). DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol. Biol. Evol. *30*, 1720–1728.

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

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

Young, K.T., Davis, L.M., and Dirita, V.J. (2007). Campylobacter jejuni: molecular biology and pathogenesis. Nat. Rev. Microbiol. *5*, 665–679.

Yu, C.-S., Lin, C.-J., and Hwang, J.-K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. Publ. Protein Soc. *13*, 1402–1406.

Yu, C.-S., Chen, Y.-C., Lu, C.-H., and Hwang, J.-K. (2006). Prediction of protein subcellular localization. Proteins *64*, 643–651.

Yu, F., Sun, Y., Liu, L., and Farmerie, W. (2010a). GSTaxClassifier: a genomic signature based taxonomic classifier for metagenomic data analysis. Bioinformation *4*, 46–49.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., et al. (2010b). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics *26*, 1608–1615.

Yu, N.Y., Laird, M.R., Spencer, C., and Brinkman, F.S.L. (2011). PSORTdb--an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. Nucleic Acids Res. *39*, D241-244.

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

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.

Zhang, S., Xia, X., Shen, J., Zhou, Y., and Sun, Z. (2008). DBMLoc: a Database of proteins with multiple subcellular localizations. BMC Bioinformatics *9*, 127.

Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. Bioinformatics *28*, 125–126.

Zhou, M., Boekhorst, J., Francke, C., and Siezen, R.J. (2008). LocateP: genome-scale subcellular-location predictor for bacterial proteins. BMC Bioinformatics *9*, 173.

Appendix A.

qPCR results for primer and probe sets

The following figures are the qPCR results for the positive control and clusters not shown in Figure 3.13.

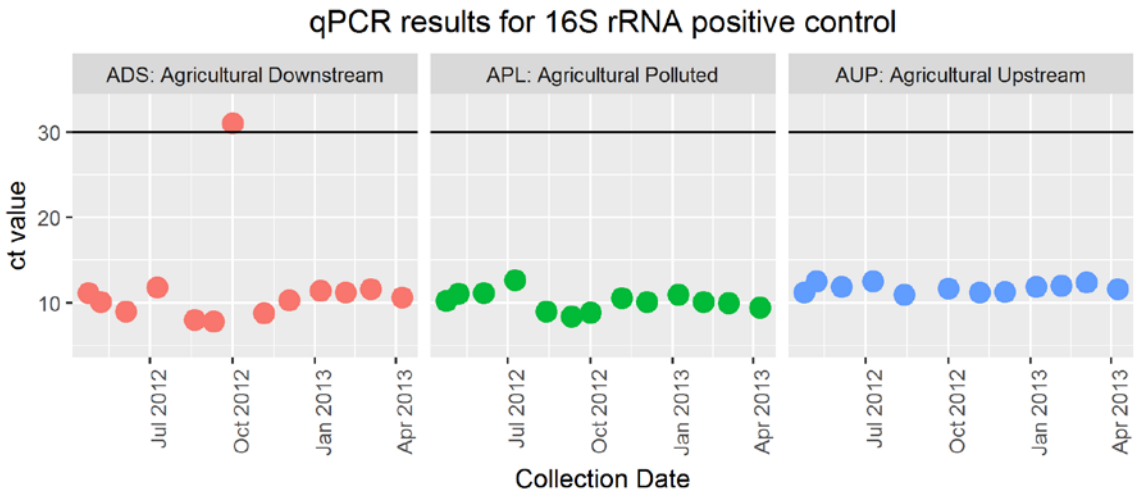

Figure A1 qPCR results for the 16S rRNA positive control.

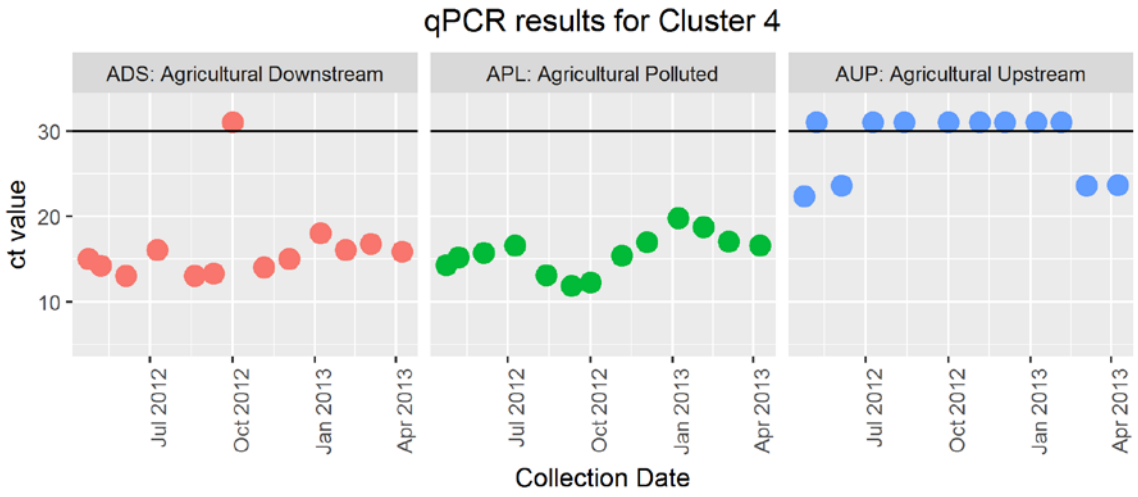

Figure A2 qPCR results for Cluster 4.

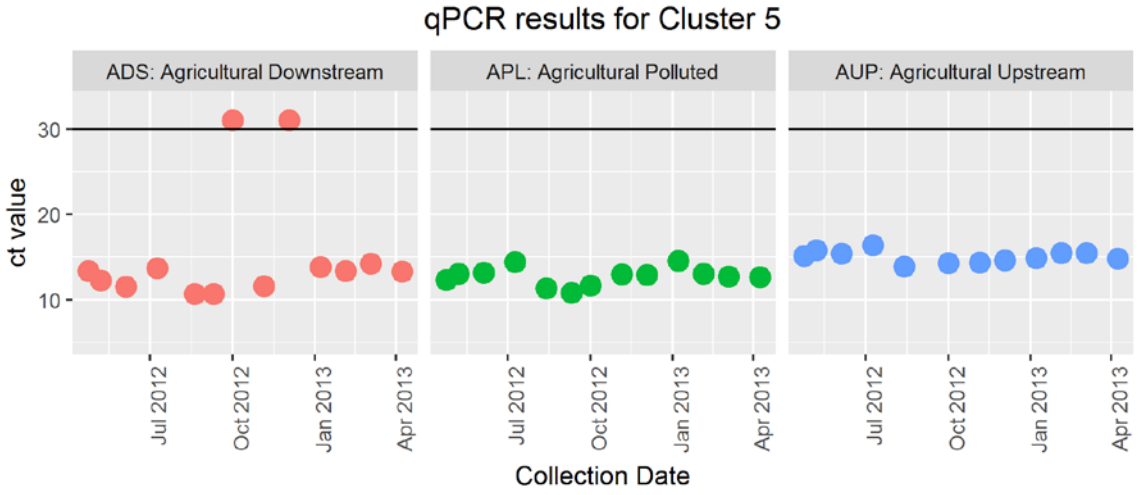

Figure A3 qPCR results for Cluster 5.

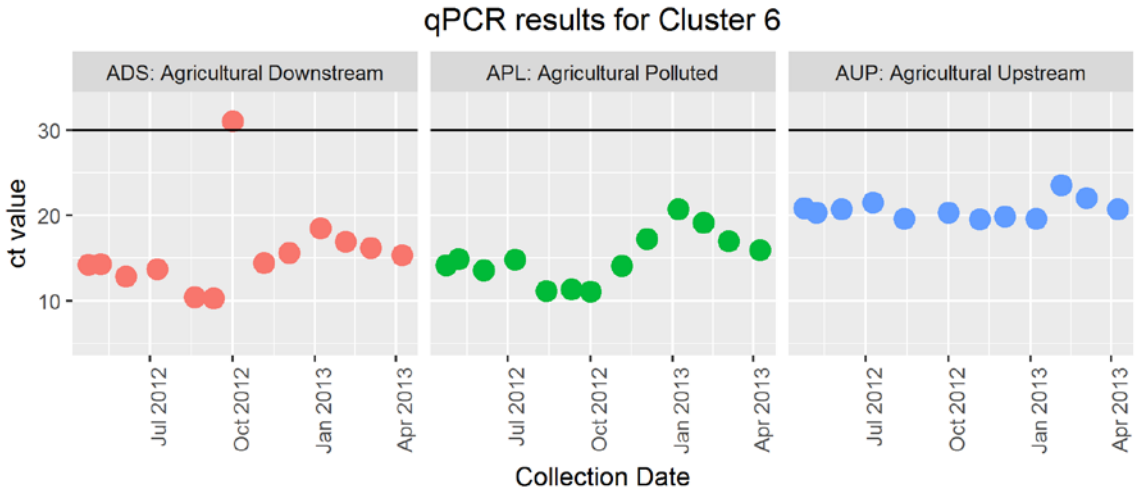

Figure A4 qPCR results for Cluster 6.

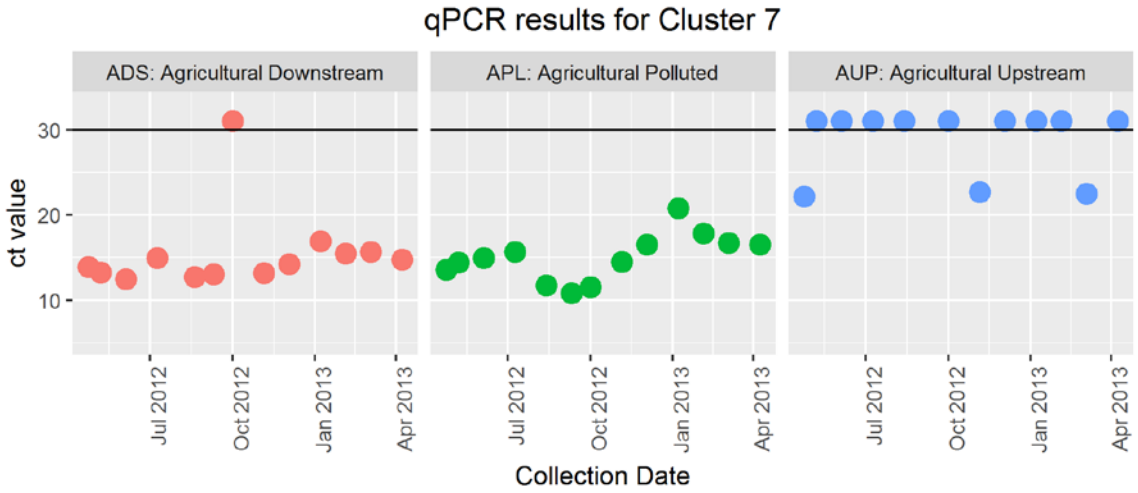

Figure A5 qPCR results for Cluster 7.

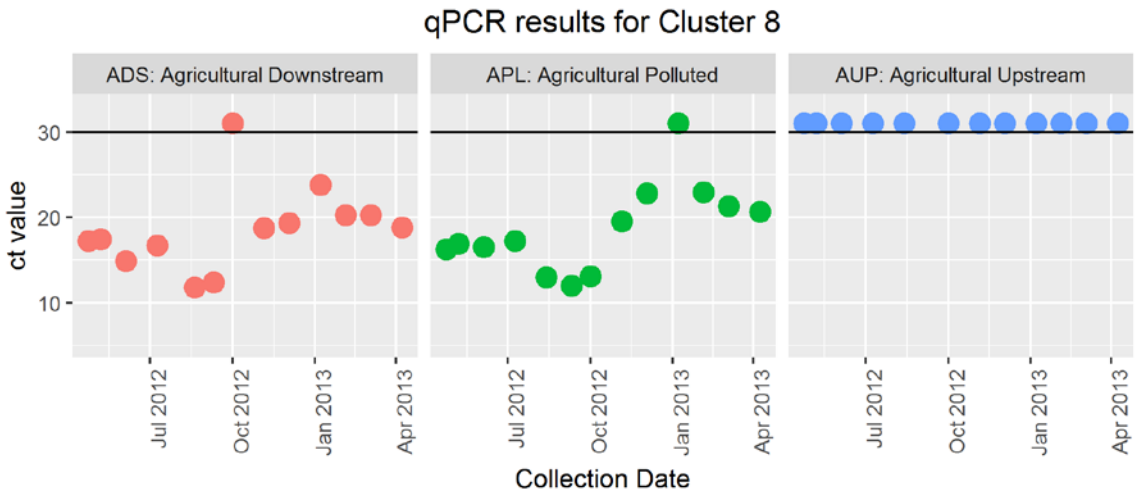

Figure A6 qPCR results for Cluster 8.

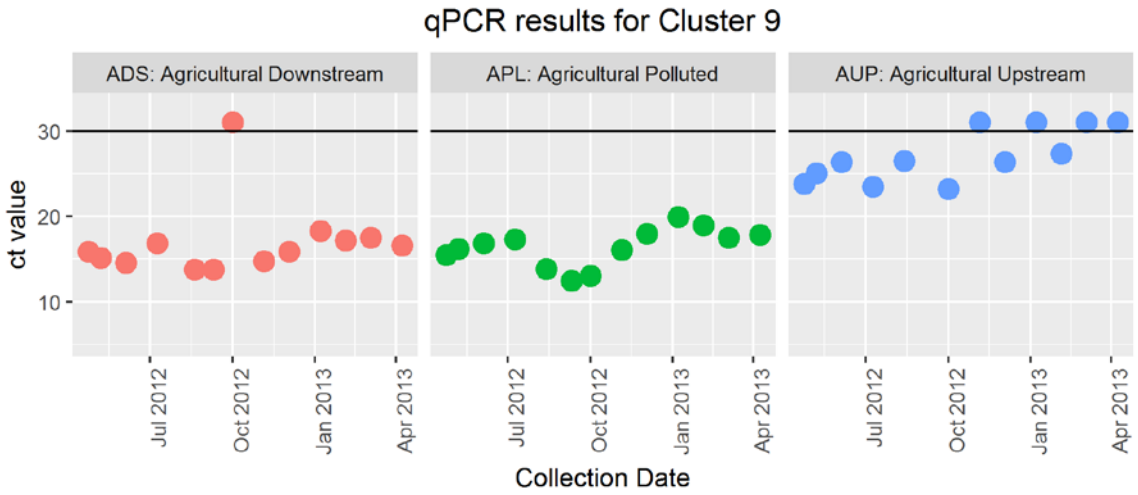

Figure A7 qPCR results for Cluster 9.

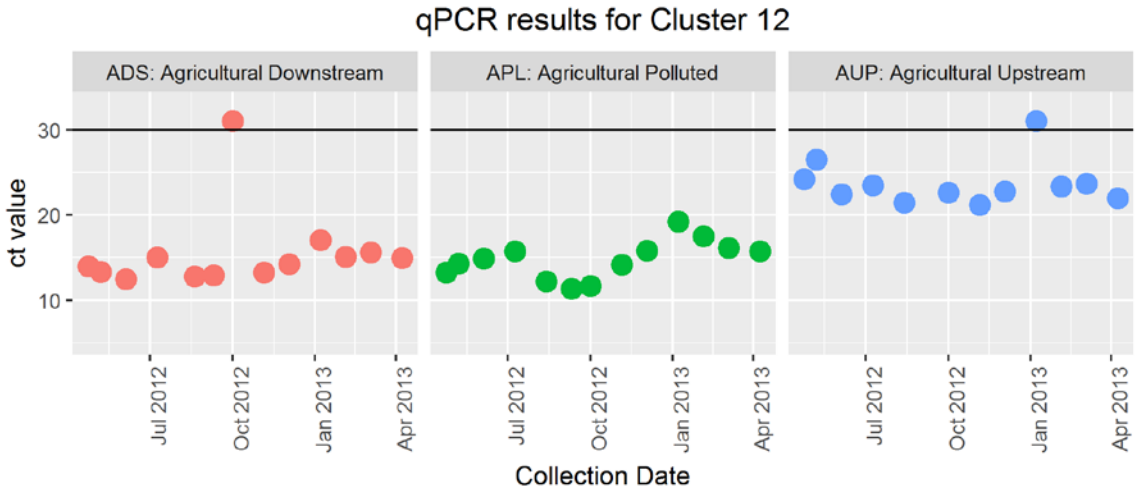

Figure A8 qPCR results for Cluster 12.

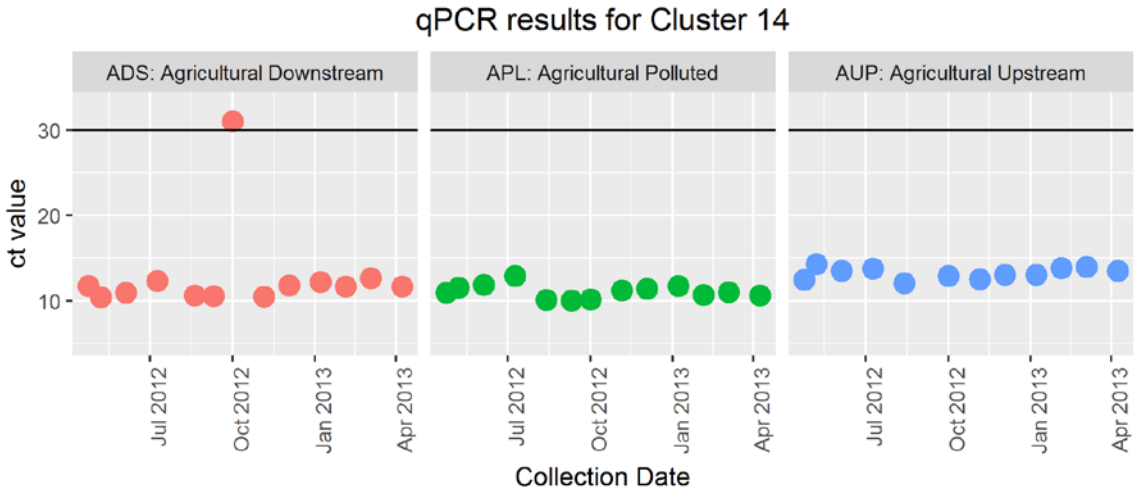

Figure A9 qPCR results for Cluster 14.

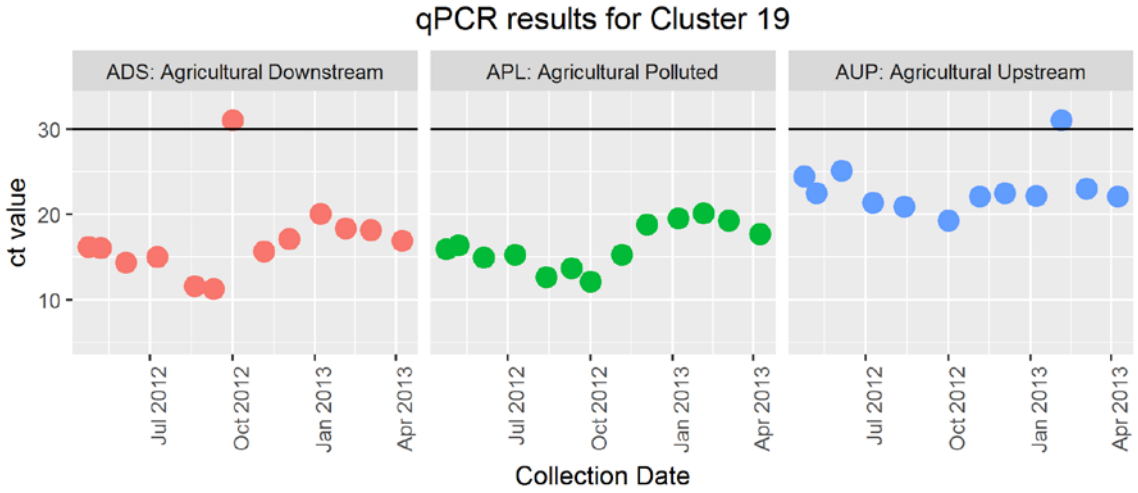

Figure A10 qPCR results for Cluster 19.

Appendix B.

Non-*Legionella* freshwater bacterial pathogens and mip gene analysis

The following figures are of the non-*Legionella* pathogens found in the watershed samples (16S dataset), and a table of metagenomic sequencing reads aligning to the mip gene.

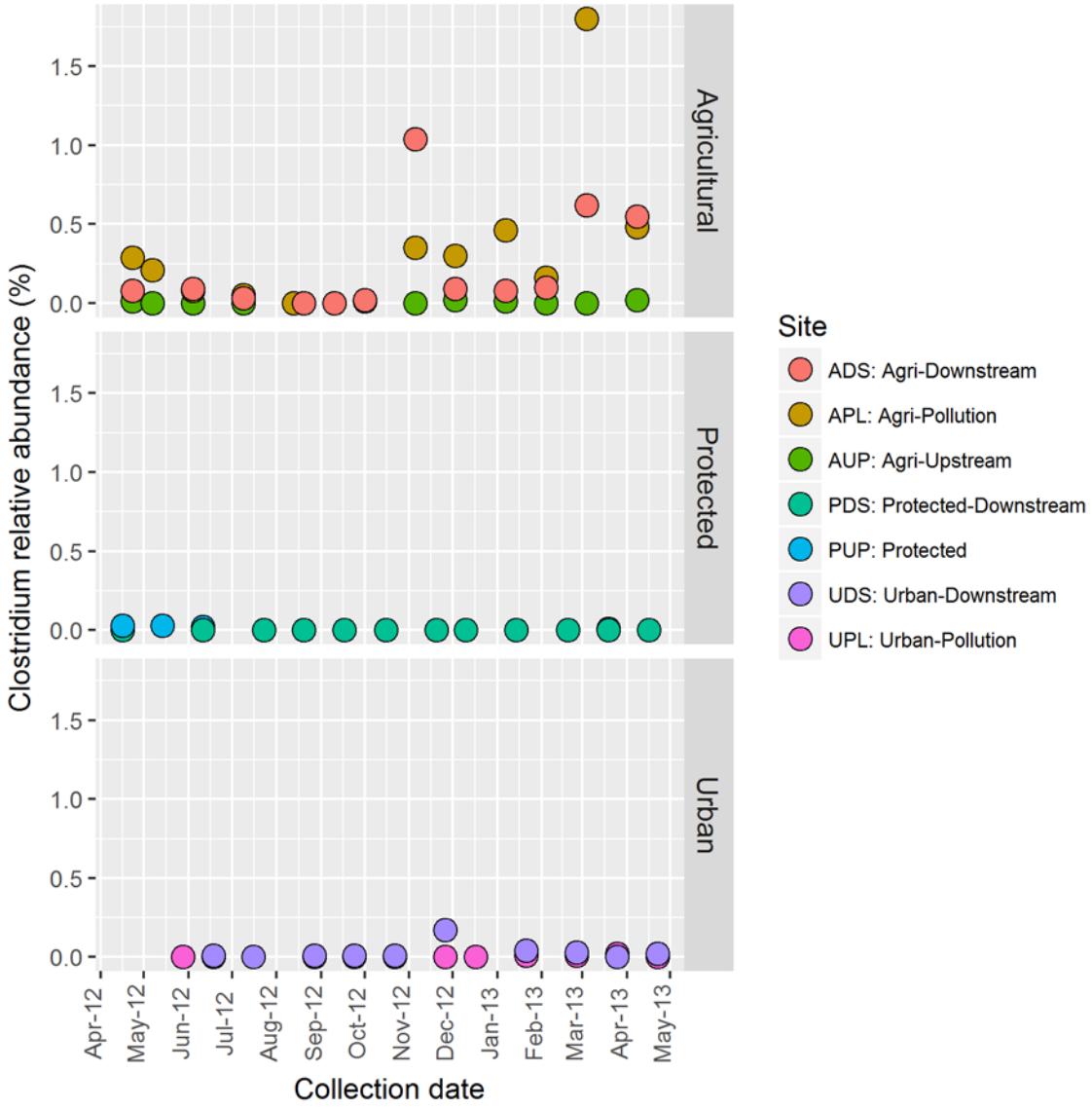

Figure B1. Relative abundance of *Clostridium* over time and sampling sites.

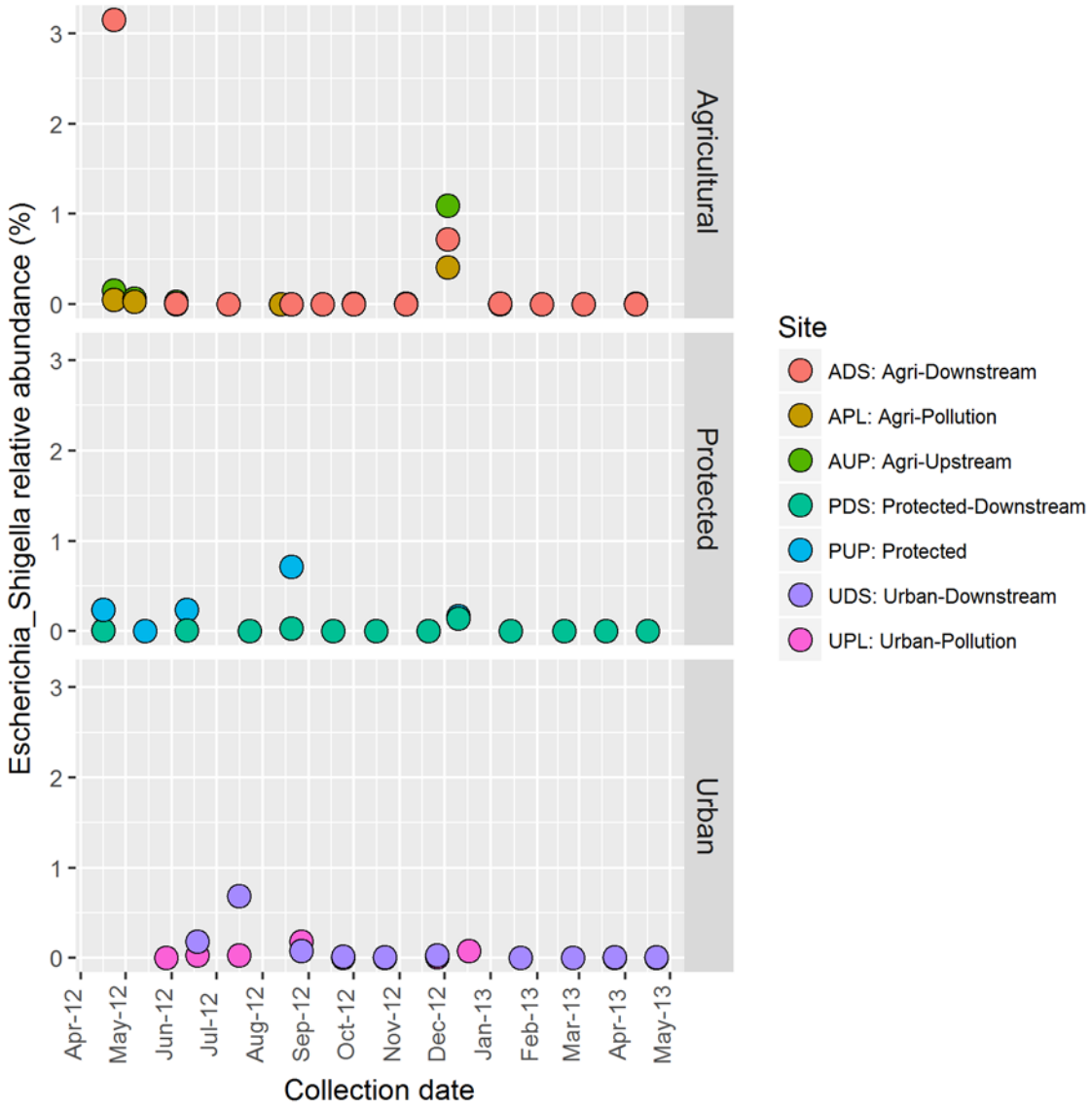

Figure B2. Relative abundance of *Escherichia/Shigella* over time and sampling sites.

Escherichia and *Shigella* cannot be differentiated so they are grouped together.

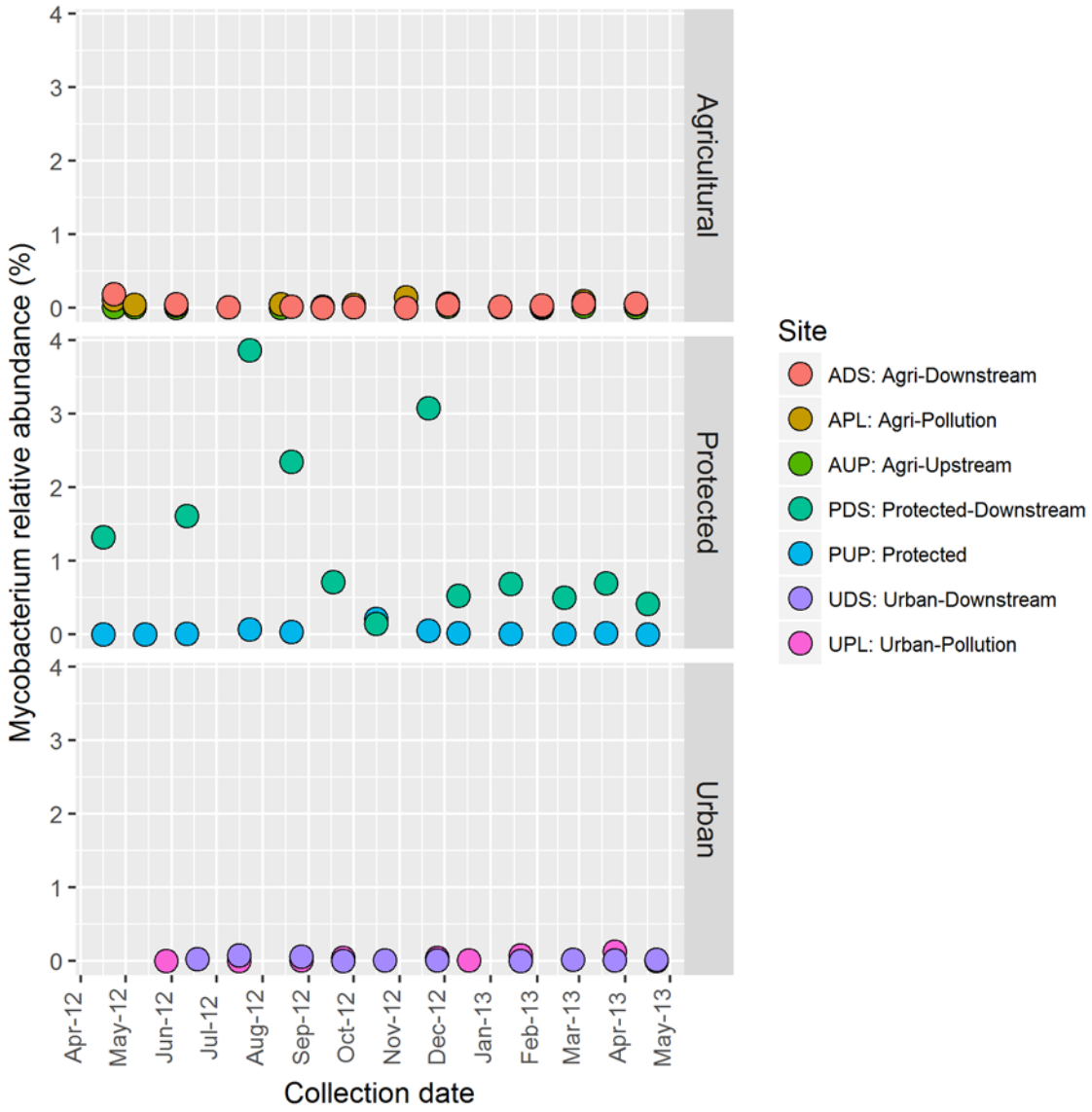

Figure B3. Relative abundance of *Mycobacterium* over time and sampling sites.

Table B1 Detection of shotgun metagenomic sequencing reads aligning to the mip gene.

Species	# Samples	Mean reads/sample	Range % identity with reference mip
Agricultural upstream (AUP)			
L. anisa	1	1	69.37
L. dumoffii	1	1	84.52
L. fallonii	2	1	80.65-91.15

L. hackeliae	1	1	81.55
L. jamestowniensis	1	1	77.97
L. maceachernii	1	1	73.33
L. massiliensis	1	1	71.71
L. micdadei	1	1	72.15
L. oakridgensis	1	1	72.22
L. pneumophila	2	1.5	74.82-77.20
L. quateirensis	2	1	80.00-82.35
L. quinlivanii	1	1	82.39
L. rubrilucens	1	1	80.74
L. sp.	8	1.75	71.10-85.71
L. tunisiensis	2	1	74.10-83.26
Agricultural downstream (ADS)			
L. anisa	1	1	74.18
L. bozemanii	1	1	71.43
L. brunensis	1	1	73.16
L. clemsonensis	1	1	73.20
L. fallonii	1	1	71.77
L. hackeliae	1	1	76.85
L. jamestowniensis	2	1	77.29-81.94
L. lansingensis	1	1	74.31
L. longbeachae	1	1	78.40
L. nagasakiensis	1	1	78.53
L. sp.	8	1.125	72.47-82.16
L. wadsworthii	1	1	68.85
Agricultural polluted (APL)			
L. anisa	1	1	91.08
L. brunensis	1	1	76.19
L. drozanskii	1	1	80.91
L. gormanii	1	1	73.16
L. impletisoli	1	1	69.63
L. longbeachae	1	1	83.93
L. quateirensis	2	1	87.01-87.32
L. sp.	5	1.4	70.41-86.45

Urban downstream (UDS)			
L. anisa	2	1	75.78-88.08
L. brunensis	2	1	76.87-82.58
L. clemsonensis	1	1	81.48
L. fallonii	3	1	72.73-78.67
L. impletisoli	1	1	78.61
L. israelensis	1	1	70.62
L.jamestowniensis	3	1	77.78-80.00
L. longbeachae	1	1	76.15
L. maceachernii	1	1	79.55
L. pneumophila	2	1	73.58-81.03
L. quateirensis	1	1	84.11
L. sainthelensi	2	1	75.16-80.99
L. shakespearei	1	1	89.67
L. sp.	6	1.5	69.52-82.83
L. worsleiensis	1	1	70.21
Urban polluted (UPL)			
L. adelaidensis	2	1	75.56-80.44
L. anisa	1	1	77.27
L. drancourtii	1	1	74.73
L. fallonii	3	1	80.00-91.60
L. gratiana	1	1	70.43
L. jamestowniensis	2	1	77.96-78.72
L. jordanis	1	1	82.50
L. micdadei	1	1	81.25
L. parisiensis	1	1	73.05
L. quateirensis	2	1	82.95-88.46
L. quinlivanii	1	1	76.92
L. shakespearei	1	1	77.53
L. sp.	6	2.5	70.16-88.21
L. tucsonensis	1	1	72.34
L. wadsworthii	1	1	84.66
Protected upstream (PUP)			
L. anisa	3	1	74.18-87.25

L. clemsonensis	1	1	83.80
L. drancourtii	1	2	70.15-85.26
L. dumoffii	1	1	75.50
L. fallonii	3	1.3	76.60-84.86
L. feeleii	1	1	83.21
L. gormanii	1	1	77.01
L. hackeliae	1	1	72.84
L. israelensis	1	1	78.69
L. jamestowniensis	3	1.3	76.15-83.94
L. londiniensis	1	1	74.17
L. massiliensis	1	1	81.38
L. moravica	2	1	82.63-83.76
L. pneumophila	3	1	73.80-77.42
L. quateirensis	2	1	79.29-87.10
L. quinlivanii	1	1	79.13
L. sainthelensi	1	1	72.39
L. santicrucis	1	1	78.83
L. sp.	9	1.7	68.98-93.89
L. spiritensis	1	1	77.24
L. steelei	1	1	86.40
L. tunisiensis	2	1	81.63-90.40
L. waltersii	1	1	74.71
Protected downstream (PDS)			
L. adelaidensis	1	1	70.71
L. brunensis	1	1	80.44
L. clemsonensis	1	1	74.79
L. fallonii	2	1	75.00-89.81
L. jamestowniensis	1	1	78.87
L. micdadei	2	1	73.66-75.29
L. santicrucis	2	1	78.24-79.12
L. sp.	2	2.5	70.34-81.70
L. tucsonensis	1	1	70.49

Appendix C.

Supplementary data for Chapter 5

List C1 Complete Corynebacteriales genomes used in the identification of markers for Corynebacteriales

Amycolicicoccus subflavus DQS3-9A1
Corynebacterium aurimucosum ATCC 700975
Corynebacterium diphtheriae 241
Corynebacterium diphtheriae 31A
Corynebacterium diphtheriae BH8
Corynebacterium diphtheriae C7 (beta)
Corynebacterium diphtheriae CDCE 8392
Corynebacterium diphtheriae HC01
Corynebacterium diphtheriae HC02
Corynebacterium diphtheriae HC03
Corynebacterium diphtheriae HC04
Corynebacterium diphtheriae INCA 402
Corynebacterium diphtheriae NCTC 13129
Corynebacterium diphtheriae PW8
Corynebacterium diphtheriae VA01
Corynebacterium efficiens YS-314
Corynebacterium glutamicum ATCC 13032
Corynebacterium glutamicum ATCC 13032
Corynebacterium glutamicum R
Corynebacterium jeikeium K411
Corynebacterium kroppenstedtii DSM 44385
Corynebacterium pseudotuberculosis 1/06-A
Corynebacterium pseudotuberculosis 1002
Corynebacterium pseudotuberculosis 258
Corynebacterium pseudotuberculosis 267
Corynebacterium pseudotuberculosis 3/99-5
Corynebacterium pseudotuberculosis 31
Corynebacterium pseudotuberculosis 316
Corynebacterium pseudotuberculosis 42/02-A
Corynebacterium pseudotuberculosis C231
Corynebacterium pseudotuberculosis CIP 52.97
Corynebacterium pseudotuberculosis Cp162
Corynebacterium pseudotuberculosis FRC41
Corynebacterium pseudotuberculosis I19
Corynebacterium pseudotuberculosis P54B96
Corynebacterium pseudotuberculosis PAT10
Corynebacterium resistens DSM 45100
Corynebacterium ulcerans 0102
Corynebacterium ulcerans 809
Corynebacterium ulcerans BR-AD22
Corynebacterium urealyticum DSM 7109
Corynebacterium variabile DSM 44702

Gordonia bronchialis DSM 43247
Gordonia polyisoprenivorans VH2
Gordonia sp. KTR9
Mycobacterium abscessus ATCC 19977
Mycobacterium africanum GM041182
Mycobacterium avium 104
Mycobacterium avium subsp. paratuberculosis K-10
Mycobacterium bovis AF2122/97
Mycobacterium bovis BCG str. Mexico
Mycobacterium bovis BCG str. Pasteur 1173P2
Mycobacterium bovis BCG str. Tokyo 172
Mycobacterium canettii CIPT 140010059
Mycobacterium chubuense NBB4
Mycobacterium gilvum PYR-GCK
Mycobacterium intracellulare ATCC 13950
Mycobacterium intracellulare MOTT-02
Mycobacterium intracellulare MOTT-64
Mycobacterium leprae Br4923
Mycobacterium leprae TN
Mycobacterium marinum M
Mycobacterium massiliense str. GO 06
Mycobacterium rhodesiae NBB3
Mycobacterium smegmatis str. MC2 155
Mycobacterium smegmatis str. MC2 155
Mycobacterium sp. JDM601
Mycobacterium sp. JLS
Mycobacterium sp. KMS
Mycobacterium sp. MCS
Mycobacterium sp. MOTT36Y
Mycobacterium sp. Spyr1
Mycobacterium tuberculosis CCDC5079
Mycobacterium tuberculosis CCDC5180
Mycobacterium tuberculosis CDC1551
Mycobacterium tuberculosis CTRI-2
Mycobacterium tuberculosis F11
Mycobacterium tuberculosis H37Ra
Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis H37Rv
Mycobacterium tuberculosis KZN 1435
Mycobacterium tuberculosis KZN 4207
Mycobacterium tuberculosis KZN 605
Mycobacterium tuberculosis RGTB327
Mycobacterium tuberculosis RGTB423
Mycobacterium tuberculosis UT205
Mycobacterium ulcerans Agy99
Mycobacterium vanbaalenii PYR-1
Nocardia cyriacigeorgica GUH-2
Nocardia farcinica IFM 10152
Rhodococcus equi 103S
Rhodococcus erythropolis PR4
Rhodococcus jostii RHA1

Rhodococcus opacus B4
Segniliparus rotundus DSM 44985
Tsukamurella paurometabola DSM 20162

List C2 Complete Thermotogae genomes used in the identification of markers for Thermotogae

Fervidobacterium nodosum Rt17-B1
Fervidobacterium pennivorans DSM 9078
Kosmotoga olearia TBF 19.5.1
Marinitoga piezophila KA3
Mesotoga prima MesG1.Ag.4.2
Petrotoga mobilis SJ95
Thermosipho africanus TCF52B
Thermosipho melanesiensis BI429
Thermotoga elfii
Thermotoga hypogea
Thermotoga lettingae TMO
Thermotoga maritima MSB8
Thermotoga naphthophila RKU-10
Thermotoga neapolitana DSM 4359
Thermotoga petrophila RKU-1
Thermotoga sp. RQ2
Thermotoga thermarum DSM 5069

List C3 Complete Deinococci genomes used in the identification of markers for Deinococci

Deinococcus deserti VCD115
Deinococcus geothermalis DSM 11300
Deinococcus gobiensis I-0
Deinococcus maricopensis DSM 21211
Deinococcus proteolyticus MRP
Deinococcus radiodurans R1
Marinithermus hydrothermalis DSM 14884
Meiothermus ruber DSM 1279
Meiothermus silvanus DSM 9946
Oceanithermus profundus DSM 14977
Thermus scotoductus SA-01
Thermus sp. CCB_US3_UF1
Thermus thermophilus HB27
Thermus thermophilus HB8
Thermus thermophilus JL-18
Thermus thermophilus SG0.5JP17-16
Truepera radiovictrix DSM 17093

Table C1 Corynebacteriales incomplete genomes and which ones contain an ortholog of Rv3802c

Species	Rv3802c
Mycobacterium colombiense	✓
Mycobacterium fortuitum	✓
Mycobacterium hassiacum	✓
Mycobacterium parascrofulaceum	✓
Mycobacterium phlei	✓
Mycobacterium thermoresistible	✓
Mycobacterium tusciae	✓
Mycobacterium vaccae	✓
Mycobacterium xenopi	✓
Corynebacterium accolens	✓
Corynebacterium ammoniagenes	✓
Corynebacterium amycolatum	✓
Corynebacterium bovis	✓
Corynebacterium capitovis	✓
Corynebacterium casei	✓
Corynebacterium caspium	✓
Corynebacterium circoniae	✓
Corynebacterium doosanense	✓
Corynebacterium durum	✓
Corynebacterium genitalium	✓
Corynebacterium glucuronolyticum	✓
Corynebacterium lipophiloflavum	✓
Corynebacterium lubricantis	✓
Corynebacterium mastitidis	✓
Corynebacterium matruchotii	✓
Corynebacterium nuruki	✓
Corynebacterium pilosum	✓
Corynebacterium propinquum	✓
Corynebacterium pseudogenitalium	✓
Corynebacterium striatum	✓
Corynebacterium timonense	-

Corynebacterium tuberculostearicum	✓
Corynebacterium ulceribovis	✓
Turicella otitidis	-
Gordonia aichiensis	✓
Gordonia alkanivorans	✓
Gordonia amarae	✓
Gordonia amicalis	✓
Gordonia araii	✓
Gordonia effusa	✓
Gordonia hirsute	✓
Gordonia malaquae	✓
Gordonia namibiensis	✓
Gordonia neofelicfaecis	✓
Gordonia otitidis	✓
Gordonia paraffinivorans	✓
Gordonia rhizosphera	✓
Gordonia rubripertineta	✓
Gordonia sihwensis	✓
Gordonia soli	✓
Gordonia sputi	✓
Gordonia terrae	✓
Nocardia abscessus	-
Nocardia aobensis	-
Nocardia araoensis	-
Nocardia asiatica	-
Nocardia brevicatena	-
Nocardia carnea	-
Nocardia cerradoensis	-
Nocardia concave	-
Nocardia exalbida	-
Nocardia higoensis	-
Nocardia jiangxiensis	-
Nocardia niigatensis	-
Nocardia otitidiscaviarium	✓
Nocardia paucivorans	-

Nocardia pneumoniae	-
Nocardia takedensis	-
Nocardia tenerifensis	-
Nocardia terpenica	-
Nocardia testacea	-
Nocardia thailandica	-
Nocardia transvalensis	-
Nocardia veterana	-
Nocardia vinacea	-
Rhodococcus imtechensis	✓
Rhodococcus qingshengii	✓
Rhodococcus rhodochrous	✓
Rhodococcus ruber	✓
Rhodococcus triatomae	✓
Rhodococcus wratislaviensis	✓
Smaragdicoccus niigatensis	✓
Dietzia cinnamea	✓
Dietzia alimentaria	✓

Table C2 Characteristics of prospective *Deinococcus-Thermus* protein markers

Characteristic	DR_0042	DR_1021	DR_1474	DR_2007	DR_2136	DR_2156	DR_2318
Accession ID	NP_293768.1	NP_294745.1	NP_295197.1	NP_295730.1	NP_295859.1	NP_295879.1	NP_296039.1
Length (a.a.)	257	167	228	79	206	128	210
Function	Metallophospho-esterase[a]	S-layer-like[b]	unknown	unknown	Putative lipoprotein[a]	unknown	unknown
Domains[a]	PF00149: Metallophos	PF14326: DUF4384	none	PF11609: DUF3248[c]	PF11306: DUF3108	PF11482: DUF3208	none
pSORTb prediction	Cytoplasmic (8.96)	unknown	unknown	unknown	unknown	unknown	unknown
Phylogenetic distribution[d]	16/16 complete 17/19 draft	16/16 complete 11/19 draft D. pimensis, M. cerebereus, M.chliarophilus, M. rufus, M. taiwanensis, M. timidus excepted	16/16 complete 17/19 draft	16/16 complete 16/19 draft *D. ficus excepted*	16/16 complete 16/19 draft *D. ficus excepted*	16/16 complete 19/19 draft[d]	16/16 complete 17/19 draft

[a] Identified using NCBI Conserved Domain Search. Pfam database accession codes given (PFxxxxx).
[b]Function identified using COMBREX protein cluster search tool (Website description: "Search results are grouped into clusters of highly similar, and putatively isofunctional, genes. We use NCBI Protein Clusters as our clustering model.")
[c] DUF3248 comprises ~80% of DR_2007; consensus domain most well conserved of all markers – detects orthologs with highest accuracy (98-100% query coverage; E-value ≤ 1.00e-11)
[d] 16 complete genomes and 19 draft genome assemblies exist for Deinoccocus-Thermus speices (8 and 10, respectively, for Deinococcales only); all proteins except DR_2156 are absent from *D. wulumuqiensis* and *D. xibeiensis* draft genomes.

Table C3 Characteristics of prospective *Deinococcales* protein markers

Characteristic	DR_0638	DR_0889	DR_2001	DR_2271
Accession ID	NP_294361.1	NP_294613.1	NP_295724.1	NP_295992.1
Length (a.a.)	142	116	103	695
Function	unknown	unknown	unknown	precursor; putative membrane protein
Domains[a]	none	PF08899: DUF1844	none	PF13424: TPR_12
pSORTb prediction	unknown	unknown	unknown	Cytoplasmic membrane (9.80)
Phylogenetic distribution[d]	8/8 complete 8/10 draft	8/8 complete 6/10 draft *D. aquatilis & D. pimensis excepted*	8/8 complete 7/10 draft *D. pimensis excepted*	8/8 complete 7/10 draft *D. pimensis excepted*

[a] Identified using NCBI Conserved Domain Search. Pfam database accession codes given (PFxxxxx).
[b]Function identified using COMBREX protein cluster search tool (Website description: "Search results are grouped into clusters of highly similar, and putatively isofunctional, genes. We use NCBI Protein Clusters as our clustering model.")
[c] DUF3248 comprises ~80% of DR_2007; consensus domain most well conserved of all markers – detects orthologs with highest accuracy (98-100% query coverage; E-value ≤ 1.00e-11)
[d] 16 complete genomes and 19 draft genome assemblies exist for Deinoccocus-Thermus speices (8 and 10, respectively, for Deinococcales only); all proteins except DR_2156 are absent from *D. wulumuqiensis* and *D. xibeiensis* draft genomes.

Table C4 Characteristics of prospective *Thermotogae* protein markers

Characteristic	Tlet_0008	Tlet_0231	Tlet_1043
Accession ID	YP_001469645.1	YP_001469865.1	YP_001470672.1
Length (a.a.)	357	211	622
Function[a]	unknown	Organic solvent tolerance protein (Imp/OstA)	unknown
Domains[b]	PF13414: TPR_11	PF06835: LptC	none
pSORTb prediction	Cytoplasmic (8.96)	Unknown (may have multiple localization sites)	Cytoplasmic membrane (9.80)
Phylogenetic distribution	19/19 complete	19/19 complete	19/19 complete

[a]Function identified using COMBREX protein cluster search tool (Website description: "Search results are grouped into clusters of highly similar, and putatively isofunctional, genes. We use NCBI Protein Clusters as our clustering model.") [10]
[b]Identified using NCBI Conserved Domain Search [11]. Pfam database accession codes given (PFxxxxx).

Table C5 Optimal query sequences for all markers

Protein	Sequences	Domains	Avg. % identity
DR_0042	MRKVIAIGDLHADFPALWRALRAAGCADADGLPTEPLRSGLYRVVLLGDLVHPKTPRDYARLTGLEPF DPSDPDHLRLAARAQIRELERLKAFQEAAPGHVHILLGNHDDAVLTGEPVLHNSHGLKHLEFHPEHGG VPLPEHLRAWMAGFPRELRLGGVHFAHVGPVPWLQEYDDLFYADREAKTWWFDTPDYVERMGYRF GVYGHTQMKGGILLKERHGFALIDALDLRNEYLELI	PF00149: Metallophos	67.06
DR_1021	LITRFEPDRGEGATYRVGEEVRFRLTLRRPGYVTLVALDPDGRAYELDRNVYLPAGTPHVLPRPQDGV RYNVAPPPRGLQRVRAIYTDVPPTTDLVLRGVYDNGDWNARTRAYLEASGARDRDVAETYFYIR	PF14326: DUF4384	57.68
DR_1474	FYPAETGLEWSYLPEGEALSSPPYRLRVLGPTVFEGQEALRFRLFGRGADRTYYRQVGAFGVRLLGF EKPGVRVRLTPPWLEYPPEAALAVGLRWGGQTEVRVTLLTPDGGKVQAQGETLRYRYEVLERREVR VPAGTLDVWRITRQIRDDDVGGLFPPATQEVQFAPFVGEVRTPEGLLLTGRNF	N/A	58.40
DR_2007	LEALGGHLVWRIGKAEEEDVLVVRVGLASATPRFAHLPRLRNAPDAELQRLAQEGRVRVEWVD	PF11609: DUF3248	70.79

DR_2136	GVVTVRQGKEEAAAPLLTDYHDPLSLLLALRGRLELEPGEVARFPMPGGRVYVERLPDLEVEGRPARVYRLRPGLALVYVEQ	PF11306: DUF3108 *(incomplete)*	61.79
DR_2156	AVRLLQGYLWHPRDLDVDLEALLPRELDGDAHVLWDEVPPPFAFFEDGTPTATQRFYQFTLLRLTEERPEALHPLAEAASQALGPLLEATPPGVGWQLLEDLRPL	PF11482: DUF3208	66.72
DR_2318	VRLQGVELTLYPERDPGAVWRFRAAEVVEDPGSGETRLTGLLSGERYVEGRLDLRLFAPEVTIDPQDNLRTPYARVEILKGCYRLELSGPGEGPVLIEQGEGFSAPWVRIEAP	N/A	57.14
DR_0638	MNYAATLAVLVVLAFCFPITVRLGAALGIPEALGAAVLGAVLVFGITAYLVRWQVARHSLRLSRLEEARAQVAADPDDSRSYFLEGEHLGQMLLALGRRREAAEVIDRYSRLGGARESEIVALREALSQAERRQRRAQREQ	N/A	74.86
DR_0889	MPHPEFVGLVNSLQATAEAALGDLNAATASAASRDGLGLAEGRARQTAERSLKLLTMLAEKTRGNNLDFTEAELLTDAIGSLRARL	PF08899: DUF1844	83.67
DR_2001	DALFDLAVNRAAAALRGLGLLGRAPADPAAALAAWHARTRFARRVPLAAVRAALARHPAAGEGEWHWAAGGPAGGWQPGKAPFP	N/A	67.64
DR_2271	MKRTLLLLALALGGTASAEVRVDGQTVRVLSGEQVTWQWTFAPALGQVSRPVRLDNQNYVAVGPVVYALSDEGRVLGRADLPGPVTSLDSSGGAVRVTVGGAGRSERLTLGDPTGDALPVEERVVFTPDPAVTGWLARFADALPAQELTRAGAQFPANPFIALREAEAAKTRGDDYAALSAVRRALGVSVPFPAGTQLAARLDRAGFPAAANLALDRAKRDAAARGLDPAVPVSRAALFAYGNPSGYVGTLLAQNRLPRAEVWMSFLRELYPRFEGGDALYRRYADVLDAQGRAGEAEEWRQFARSLREGTLYNLGPNDTRVLRDVARLATFALLLALLAAVLTLSARAWPTQRQDLAALGGRWRSWRRPPQRLRRTALAYASFSERLLLVTLAAALLTALGGWQWANLTGRALQAPALGTGTYSASLALPSTGDADPQPVAATYLLSGLGAQLDGDLAAARTLYGRAGDDACALNNLGAIAQSRDDLPQARTLYQQALALQPDLSAPAYNLGRNPATPETAFQRTYRPGQPRLCYPERRDLARSVSGDLGSVLRQTVTDPLGFLTERRPATRLNWALLAALPLIGALLVLLLIPRAASAGRLGRPAAFRLLAVLFPGAGLLGNAWGGVLLVAWAGVVVALLARWLPFPALPLQTAAVRSALWLALGALYVLNVLALLLIEAAHARRRRLRELEA	PF13424: TPR_12	56.76
Tlet_0008	DITTSTPFVVALRDLLYEGDWELLKEDMKKKENIMKEVQTCEKLEKNVVHLDETVYEPIIFPEFQEWLREENITPKDFKHVSLKGLYDLALEYADRNLYDVAHDIIKFMLDI	PF13414: TPR_11	74.86
Tlet_0231	MTSLRKLLTLAICVSFWGLSFAKTVHIFSEFVKPEKEQAYYEGNVKVEIEEDNLNLYCDSMKVSKYQNEWRIVEANNAEVFFDDGTATSTRLRYDIKLKTGTMSGDVEAEIFDKESTDTILINCDSMEIDLENDIFQGNSTNQVQIYKGKIEASAKSFFYDRKSGKIKLEGDVTLIDHDKNMKMWATSVEITTIDDKMTATNARVELIVEE	PF06835: LptC	44.02
Tlet_1043	LEKSLEIFQKLVEDYESGSPQDPFLSYVKENIPQLQLYRYYRRQLAGSVEKTEFAKTIGDYLFRHIVLYDSKETDWERKLANALFLSYLQSKLSGSKFSESVLKNSPAFNSFFNEYRMFVRSNARNLFRWIIAYYLGLTDTPPGGNLPPVEPPVEFNLGIRNYGFSPNVNHDVHPDIIKLLPSDLEVKLKEAIEEIASSKNFSKDQAEYERNINRQASLLWREIEKNISALQNEVADIFENTTPKEINFWWIRFLVYGVLLVFLRLKYKLWRLTILQIIIAEEILLIWFFDILSLNTSDNMIYGIVAVFAFIFFNIILLRRAFGRRKRRYLYILLSLLFVLLLFIPSYISKREQPELLMDNDSEFENSPYYDQLKNDVFEDPDSKVSTIIRRLNTIALSSKEETKQIVNTLGSLPENLLKIEALKEIEVTKNGIFLFNPDRSKFFSIANFEDRLNLFGKLEGIKDLSDYLSDEKSRYRKYERALKSLDKFVKRITSYSSEKFRQDFERELENLFERYPLIEGVSFSYSTEKSYLSLKPYRTPNGKIFYGITALLIGIFTFFLLFFSAVLGGRYLLFPAAATLFTSILSMIKWKHLEVFVEGIPPLIIETSSTHTFHIEVFLIFVSLFLLYKLFRKGADVL	N/A	44.16
Rv3802c	GLTMTGPRPGGFGALDGRTNEICAQGDLICAAPAQAFSPANLPTTLNTLAGGAGQPVHAMYATPEFWNSDGEPATEWTLNWAHQLIENA	PF01083: Cutinase	60.26

Table C6 List of proteins that are found only in the two available Dictyoglomi genomes

GI	Locus Tag	Assigned name
206901766	DICTH_0033	hypothetical protein
206900134	DICTH_0035	hypothetical protein
206901537	DICTH_0052	hypothetical protein
206901243	DICTH_0364	hypothetical protein
206900777	DICTH_0377	hypothetical protein
206900399	DICTH_0390	hypothetical protein
206901830	DICTH_0391	hypothetical protein
206901345	DICTH_0457	hypothetical protein
206900154	DICTH_0469	hypothetical protein
206900133	DICTH_0481	lipoprotein
206901880	DICTH_0507	hypothetical protein
206900198	DICTH_0580	hypothetical protein
206901771	DICTH_0604	hypothetical protein
206900695	DICTH_0685	MlpD
206901066	DICTH_0747	hypothetical protein
206900350	DICTH_0751	hypothetical protein
206901764	DICTH_0874	hypothetical protein
206901390	DICTH_0886	hypothetical protein
206901410	DICTH_0888	hypothetical protein
206900778	DICTH_0987	lipoprotein
206900494	DICTH_1016	hypothetical protein
206901951	DICTH_1045	hypothetical protein
206901090	DICTH_1046	prepilin-type N-terminal cleavage/methylation domain protein
206901781	DICTH_1048	hypothetical protein
206901592	DICTH_1050	hypothetical protein
206901154	DICTH_1051	prepilin-type N-terminal cleavage/methylation domain protein
206900280	DICTH_1052	hypothetical protein
206901093	DICTH_1053	hypothetical protein

206900300	DICTH_1054	hypothetical protein
206901653	DICTH_1079	hypothetical protein
206900722	DICTH_1112	binding-protein-dependent transport systems inner membrane component
206901643	DICTH_1124	hypothetical protein
206901446	DICTH_1129	hypothetical protein
206901941	DICTH_1144	hypothetical protein
206900744	DICTH_1271	hypothetical protein
206901954	DICTH_1296	hypothetical protein
206901321	DICTH_1327	hypothetical protein
206900213	DICTH_1349	hypothetical protein
206900370	DICTH_1397	chromosome partition protein
206900838	DICTH_1413	DNA double-strand break repair Rad50 ATPase
206900261	DICTH_1436	hypothetical protein
206900846	DICTH_1669	hypothetical protein
206900809	DICTH_1725	hypothetical protein
206901979	DICTH_1734	hypothetical protein
206900887	DICTH_1740	hypothetical protein
206901286	DICTH_1749	hypothetical protein
206901293	DICTH_1770	hypothetical protein
206900960	DICTH_1771	hypothetical protein
206900460	DICTH_1810	hypothetical protein
206901824	DICTH_1864	hypothetical protein
206900723	DICTH_1880	lipoprotein
206900230	DICTH_1904	hypothetical protein
206901563	DICTH_1919	hypothetical protein
206901785	DICTH_1950	hypothetical protein
206901852	DICTH_1962	hypothetical protein