ASSESSING THE DIVERSITY OF FRESHWATER BACTERIA AND VIRUSES

EMPLOYING METAGENOMICS TO CAPTURE THE DYNAMICS AND THE DIVERSITY OF FRESHWATER BACTERIA AND VIRUSES

By MOHAMMAD MOHIUDDIN, M.Sc., Hons. B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2019) McMaster University

Department of Biology Hamilton, Ontario, Canada

TITLE: Employing Metagenomics to Capture the Dynamics and the Diversity of Freshwater Bacteria and Viruses

AUTHOR: Mohammad Mohiuddin, M.Sc., Hons. B.Sc. (University of Dhaka, Bangladesh)

SUPERVISOR: Professor Herb E. Schellhorn

NUMBER OF PAGES: xx, 241

Lay Abstract

Bacteria and viruses are abundant in freshwaters, yet our understanding of bacterial and viral communities in freshwaters remains inadequate. In addition, the changing patterns of waterborne diseases in recent years have warranted strong monitoring of recreational freshwaters across the globe. The goals of this thesis were to assess the dynamics and the diversity of freshwater bacterial and viral communities of the lower Great Lakes region and identify pathogenic bacterial and viral species. Employing a metagenomic approach, I have investigated the spatial and temporal distribution of bacterial and viral communities in recreational freshwaters. Using computational approaches, I have identified pathogenic bacteria and viruses and investigated the link between connected habitats. My research advances our understanding of bacterial and viral communities of the lower Great Lakes region and provides useful information to water quality decision makers for the optimization of municipal sampling programs currently employed in the Great Lakes area.

Abstract

Microbial water communities are a complex consortium of bacteria, viruses and protozoa. These complex microbial and viral communities may contain pathogens that cannot be detected by conventional methods. Next-generation sequencing (NGS) offers the potential to exhaustively characterize all microbial and viral components of a given water sample and facilitates the identification and quantification of pathogens of interest. The goals of this thesis were to assess the dynamics and the diversity of freshwater bacterial and viral communities of the lower Great Lakes region and identify pathogenic bacterial and viral species. We first assessed the diversity of viral communities in six different beaches of Lake Ontario and Lake Erie, two of the largest freshwater reservoirs in North America. We employed a robust and routinely applicable approach that can provide a comprehensive analysis of bacterial and viral community composition. Our analysis suggests that the viral communities of the lower Great Lakes region are dominated by bacteriophages but also contained viruses of plants and animals. Exhaustive characterization of bacterial communities indicates that the bacterial community composition is highly diverse, and the diversity differs between recreational waters and beach sands. In addition, we identified sequences of pathogens that are not currently included in traditional water monitoring schemes in both recreational water and beach sand. To investigate the impact of spatiotemporal and environmental factors on the distribution of bacterial species, we employed a computational approach and our analysis suggests that dissolved oxygen (DO) level is strongly associated with bacterial

community diversity. Using a computational approach, we have also identified habitat-specific bacterial species and a possible link between inter-connected habitats. The findings of this thesis aid in our understanding of bacterial and viral community diversity in recreational waters and provide useful information to water quality decision makers.

Acknowledgements

First and foremost, I would like to thank Dr. Herb Schellhorn for giving me the opportunity to work with him and for being an outstanding advisor and mentor. His patience and guidance have helped me grow intellectually, both as a researcher and as a person. I am truly grateful for my time learning from him. I am also thankful for the freedom and resources I have received from him to complete my graduate studies.

I would also like to thank my supervisory committee members, Dr. Radhey Gupta and Dr. Brian Golding, for their guidance and support throughout my graduate studies. Thank you for meeting with me and for thoughtful discussions. Thanks are due to the collaborators, Dr. Tom Edge, Joshua Diamond, Glen Hudgin, for providing samples and resources throughout my graduate studies. I am grateful for the generous financial support I received from Norgen Biotek, Natural Sciences and Engineering Research Council (NSERC) of Canada, Niagara Peninsula Conservation Authority (NPCA), WaterSmart Niagara, MacWater and Niagara Region Public Health, for my research.

Throughout my time in the Schellhorn lab, I have enjoyed the company of dynamic lab members. Thank you all for your support, for the creative ideas and for all the fun I had in and outside the lab. Special thanks go to Dr. Athanasios Paschos, Steven Botts and Yasser Salama for their input in my research.

Finally, thank you to all my family, friends and loved ones, for their support and encouragement throughout my graduate studies.

Table of Contents

Lay Abstract ..... **iii**

Abstract ..... **iv**

Acknowledgements ..... **vi**

Table of Contents ..... **vii**

List of Figures ..... **xii**

List of Tables ..... **xviii**

Declaration of Academic Achievement ..... **xix**

Chapter 1: Assessing the diversity of freshwater bacteria and viruses in the metagenomic era **1**

Abstract 2

1. Introduction 3

2. Freshwater bacterial communities 5

2.1 Bacterial community structure 5

2.2 Diversity of bacterial communities 13

3. Freshwater viruses 15

3.1 Viral community structure 15

3.2 Diversity of viral communities 20

4. Metagenomics- advantages and limitations 21

4.1 Factors that impact metagenomic studies 22

5. Conclusion 37

References 40

Chapter 2: Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analyses 59

Abstract 60

1. Introduction 61

2. Materials and methods 64

2.1 Sampling sites and sample collection 64

2.2 Sample fractionation 64

2.3 Extraction of DNA from bacterial and VLP fraction 65

2.4 Preparation of DNA samples for sequencing 66
2.5 Bioinformatic analyses 66
3. Results 68
3.1 Efficiency of the fractionation scheme 68
3.2 DNA in the environmental water samples 68
3.3 Abundance of bacterial and viral genome copies in freshwater 69
3.4 Overview of the viromes 70
3.5 Taxonomic composition of the viromes 71
3.6 Comparison of the viromes 71
3.7 Phylogenetic analysis of the viromes 73
4. Discussion 74
5. Acknowledgements 79
References 80
Chapter 3: Shotgun metagenomic sequencing reveals freshwater beach sand as reservoirs of bacterial pathogens 101
Abstract 102
1. Introduction 103
2. Materials and methods 106
2.1 Sampling and DNA extraction 106
2.2 Library preparation and sequencing 107
2.3 Bioinformatic analysis 108
2.4 Statistical analysis 109
3. Results 110
3.1 Community complexity and diversity 110
3.2 Taxonomic composition 111
3.3 Differences between sand and water microbial communities 112
3.4 Potential human pathogens and fecal indicators in freshwater beaches 115
4. Discussion 116
5. Conclusions 121
6. Acknowledgement 123
7. Author contributions 123

References 124
Chapter 4: Temporal and spatial changes in bacterial diversity in mixed use watersheds of the Great Lakes region 138
Abstract 139
1. Introduction 140
2. Materials and methods 142
2.1 Sample collection and processing 142
2.2 Library preparation for 16S rRNA gene amplicons and sequencing 144
2.3 Bioinformatic analyses 145
2.4 Statistical analyses 145
2.5 Nucleotide sequence accession numbers 147
3. Results 147
3.1 Physicochemical characteristics of water 147
3.2 Bacterial community complexity and diversity 149
3.3 Taxonomic composition of bacterial communities 151
3.4 Microbial community connectivity between habitats 153
3.5 Habitat-specific bacterial species in watersheds 154
3.6 Fecal indicators and potential human pathogens 154
3.7 Temporal changes of bacterial diversity in watersheds 155
4. Discussion 156
5. Conclusions 164
6. Acknowledgement 165
7. Author contributions 165
References 166
Chapter 5: Conclusion and Future Research 182
Chapter 6 / Appendix A: Application of high-throughput 16S rRNA sequencing to identify fecal contamination sources and to complement the detection of fecal indicator bacteria in rural groundwater 185
Abstract 190
1. Introduction 191
2. Methods 191

2.1 Study site description 193
2.2 Groundwater tap sample collection 193
2.3 Septic tank effluent (STE) and manure sample collection 194
2.4 FIB detection assay 194
2.5 DNA extraction and library generation for sequencing 194
2.6 Processing and analysis of 16S rRNA gene sequences 195
2.7 Preparation of DNA standards for quantification of human fecal pollution 197
2.8 Quantification of human *Bacteroidales* marker in sewage and groundwater samples 198
3. Results 199
3.1 Detection of FIB in groundwater wells 200
3.2 Basic sequencing data 200
3.3 Identification of potential STE and animal manure contamination markers 200
3.4 Abundance of STE and manure-based OTUs in groundwater samples 201
3.5 Identification of human fecal contamination 202
4. Discussion 202
4.1 Characterization of fecal microbiota and identification of potential fecal markers 203
4.2 Fecal source identification in residential groundwater sites using 16S rRNA sequencing 203
4.3 Quantification of human-based fecal contamination 205
5. Conclusion 207
Acknowledgement 207
References 210
Chapter 7 / Appendix B: Epidemiology and ecology of emerging viruses in two freshwater lakes of the Northern Hemisphere 219
Abstract 220
1. Introduction 221
2. Viral community structure 223
2.1 Viral pathogens 224
2.2 Cyanophages 227
2.3 Phycodnaviruses 227

2.4 Viral Hemorrhagic Septicemia virus 228
2.5 Giant viruses 228
2.6 Other viruses 229
3. Sources of pathogenic viruses 229
4. Current approaches for the identification of pathogens, their limitations and scope 231
5. Summary and Conclusion 234
References 235

List of Figures

Chapter 2

Figure 2.1 Schematic diagram showing water sample fractionation steps and the efficiency of the fractionation scheme. Water samples were separated into three distinct fractions: bacterial, VLP (Virus like particles) and eDNA fraction. eDNA is defined as the DNA released in the environment after lysis (natural decay or lysis) of bacteria, viruses and higher organisms. The efficiency of the fractionation scheme was determined as percent recovery of bacteriophage (MHS 16) mixed with the water sample prior to the fractionation steps. 91

Figure 2.2 Recovery of DNA from freshwater environment. DNA was extracted from three different fractions of twelve individual 1L samples. Six samples from Lake Ontario (four samples from Lakeside Beach, one sample from Queen's Royal Beach and Fifty Point Beach) and six samples from Lake Erie (four samples from Long Beach and one sample from Long Beach Conservation Area East and Nickel Beach) were used to extract DNA and amount of DNA recovered from all three fractions were normalized to 1L water sample. 92

Figure 2.3 DNA yield from VLP, bacterial and eDNA fraction. DNA recovered from six samples from Lake Ontario (four samples from Lakeside Beach, one sample from Queen's Royal Beach and Fifty Point Beach) and six samples from Lake Erie (four samples from Long Beach and one sample from Long Beach Conservation Area East and Nickel Beach) were used to calculate DNA yield from each fraction. eDNA accounts for most of the DNA (~60%) in freshwater environment while VLP DNA and bacterial DNA are present in almost equal amount. 93

Figure 2.4 Annotation of the virome sequence reads. Sequences from both VLP DNA and eDNA fractions of Lakeside and Long Beach samples (collected in 2012) were compared against the non-redundant M5NR database (E-value $<10^{-3}$ in blastX). About 25% of the sequence reads mapped to the known protein sequences of M5NR database while majority of the sequence reads were categorized as unknown. Among the unknown reads, unknown sequences are sequences for which gene prediction tools could not predict any protein or rRNA sequence and unknown proteins represent predicted proteins with unknown functions. 94

Figure 2.5 Taxonomy of virome domains. Sequence reads from Lakeside Beach and Long Beach VLP DNA and eDNA fractions (of the year 2012) were compared against the M5NR database (E-value $<10^{-3}$ in blastX). Majority of the sequence reads (60% to 80%) mapped to bacterial sequences and 20% - 35% sequence reads mapped to virus sequences of the database. 95

Figure 2.6 Relative abundance of top 5 viral families in the lower Great Lakes water samples. blastX comparison (E-value $< 10^{-3}$) of VLP DNA and eDNA sequences of Lakeside Beach and Long Beach samples (sampled in 2012). Unclassified viruses

and viral families constituting less than 1% of total viral reads are categorized as "Other." 96
Figure 2.7 Neighbour-joining phylogenetic tree of the capsid assembly protein G20 (T4-like phages or cyanophages) (pfam07230) of Lakeside Beach virome. Lakeside -1 virome library was used as a representative metagenome profile for the Lakeside Beach site. Sequence reads with significant similarity (E value < 10^{-3} using blastX) to the G20 marker sequences were obtained, assembled at 98% identity in 35 bp using Cap3 and used to draw the phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Sequences for which the best blast hit did not correspond to the G20 marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines. 97
Figure 2.8 Neighbour-joining tree of the capsid assembly protein G20 (T4-like phages or cyanophages) (pfam07230) of Long Beach virome. Long Beach -1 virome library was used as a representative metagenome profile for the Long Beach site. Sequence homologs (E value < 10^{-3} using blastX) to the G20 marker sequences were obtained from the virome library, assembled at 98% identity in 35 bp using Cap3 and used to draw phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Sequences for which the best blast hit did not correspond to the G20 marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines. 98
Figure 2.9 Neighbour-joining phylogenetic tree of the TerL (Terminase Large subunit – for *Caudovirales* including *Myoviridae, Podoviridae* and *Siphoviridae*) (pfam03237) of Lakeside Beach virome. Lakeside -1 virome library was used as a representative metagenome profile for the Lakeside Beach site. Sequence reads with significant similarity (E value < 10^{-3} using blastX) to the TerL marker sequences were obtained, assembled at 98% identity in 35 bp using Cap3 and used to draw phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Bootstrap values of ≥ 80 are highlighted with black lines. Sequences for which the best blast hit did not correspond to the TerL marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines. 99
Figure 2.10 Neighbour-joining phylogenetic tree of the TerL (Terminase Large subunit – for *Caudovirales* including *Myoviridae, Podoviridae* and *Siphoviridae*) (pfam03237) of Long Beach virome. Long Beach -1 virome library was used as a representative metagenome profile for the Long Beach site. Sequence homologs (E value < 10^{-3} using blastX) to the TerL marker sequences were obtained from the virome library, assembled at 98% identity in 35 bp using Cap3 and used to draw phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Sequences for which the best blast hit did not correspond to the TerL marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines. 100

Chapter 3

Figure 3.1 Rarefaction and richness analysis of sequence data. Rarefaction analysis was performed on all reads with both functional (A) and taxonomic assignments (B). Each subset used for rarefaction was repeated 10 times. Sand and water samples are distinguished by color. Unique features refer to distinct and non-redundant taxonomic or functional features. (A) Functional assignments did not saturate whereas (B) taxonomic assignments saturated for all but one of the samples. (C) Average taxonomic features showing pore samples are more taxonomically rich than water samples. Standard error of the mean is indicated by the error bar. 127
Figure 3.2 Shannon diversity of samples from both Lake Ontario and Lake Erie. (A) Species level Shannon index and (B) phylum level Shannon index. Standard errors of the mean are indicated by the error bars. Shannon diversity is enriched in sand environments. 128
Figure 3.3 Taxonomic composition at the phylum level. Reads assigned to bacterial taxa were aggregated to the phylum level and plotted as proportions. Reads assigned to bacterial taxa whose phylum classification was unknown are aggregated and labelled as unclassified. Sites are distinguished by environment..... 129
Figure 3.4 Differentially abundant phyla in sand and water environments. Numerous phyla are differentially abundant between beach environments. Differential abundance of phyla was determined by DESeq2 using a paired model. log2 fold changes are indicated on the left with the associated adjusted p-values of significance (cutoff of 0.01). The normalized counts for each phylum in each environment are plotted on the right..... 130
Figure 3.5 Mean proportions of bacterial taxa at the family level in beach sand and water. The mean proportion of families in both sand and water are plotted against each other with standard error of the mean proportion indicated for each environment. The marginal histogram describes the frequency of phyla at the given proportions. Families labelled as unclassified were removed prior to determining mean proportions and are not plotted..... 131
Figure 3.6 Functions exhibiting differential abundance between beach environments. Differential abundance of functions was determined by DESeq2 using a paired model. log2 fold changes are indicated, with negative fold changes signifying enrichment in sand and positive fold changes signifying enrichment in water. Functions depicted are in level three of the SEED subsystem hierarchy and are colored by their broader level two categorization and significance was determined at an adjusted *p*-value of 0.01. 132
Figure 3.7 Many pathogens exhibit differential abundance between beach environments. Reads were classified using CLARK as described in the methods. Counts were normalized using DESeq2's size factor estimations and mean normalized counts for each environment were plotted on a $\log_{10}$ scale with the error bars indicating

standard error of the mean. Significance between environments was determined at an adjusted *p*-value of 0.05 using DESeq2's negative binomial model testing and signified by the asterisks. A confidence score determined by CLARK of 95% was used as a threshold for classification. 133

Figure 3S.1 Taxonomic composition at the superkingdom level. The proportion of reads assigned to each superkingdom is indicated. Reads belonging to any node below the superkingdom nodes were counted as belonging to that superkingdom. The proportions were calculated by dividing the number of reads belonging to a superkingdom by all reads assigned to any node in the NCBI reference tree. The "Unclassified" entry refers to sequences which have been entered in the NCBI database but have not been assigned to a superkingdom (e.g. environmental sequences). 134

Figure 3S.2 Taxonomic composition of the proteobacteria at the class rank. Proportions were determined by the number of reads assigned to each proteobacterial class divided by the total number of proteobacterial hits. The "Unclassified" label refers to reads whose assignments occurred to the proteobacterial phylum and have no classification at the class level. 135

Figure 3S.3 Differentially abundant species between beach environments. Differential abundance of species was determined by DESeq2 using a paired model. log2 fold changes are indicated, with negative l2FC signifying enrichment in the sand and positive l2FC signifying enrichment in the water. Species are coloured according to their phylum classification and significance was determined at an adjusted *p*-value of 0.05. 136

Figure 3S.4 Functional capacity is stable between samples and environments. Functional classification to the SEED subsystem hierarchy aggregated to level two is plotted as proportions of the number of reads with a functional assignment for each sample. Samples corresponding to sand and water sites are separated. 137

Chapter 4

Figure 4.1 Shannon index (a) and bacterial richness as measured by abundance-based coverage estimate (ACE index) (b) in all the habitats at the genus level. The whiskers represent minimum and maximum values. Among all the habitats, taxonomic diversity was highest in lakes and lowest in stormwater outfalls. 171

Figure 4.2 Principal coordinates analysis of bacterial communities (Bray-Curtis dissimilarities) from all the habitats. Samples were clustered based on the habitats. 172

Figure 4.3 Distribution of major phyla and classes across all habitats. Because of the higher abundance of Proteobacteria phylum among all samples, taxonomic composition of Proteobacteria phylum is shown at the class level. Only top five

classes are included. “Other Proteobacteria” refer to OTUs from TA18 and Zetaproteobacteria. OTUs from remaining phyla are labeled as other phyla. 173
Figure 4.4 Proportional contribution of source habitats (creek, river, canal and stormwater outfall habitats) in shaping lake microbial community structure. SourceTracker, a Bayesian-based approach was used to quantify the contribution of each source habitat. Here, other refers to sequence proportion that was found only in lake habitat..... 174
Figure 4.5 Relative abundance of potential human pathogen containing genera and fecal indicators in watersheds. OTUs corresponding to pathogens and FIBs were resolved at the genus level. Error bars indicate standard deviations of the mean. 175
Figure 4.6 Temporal changes of microbial communities in creek (a, b), river (c, d), and lake (e, f) habitats based on Shannon diversity and species richness (ACE index). 176

Figure 4S.1 Sampling locations of Niagara Peninsula. Sampling locations are indicated by colored circles. Green, yellow and red colored circles indicate Water Quality Index (fair, marginal and poor, respectively) used by the Canadian Council of Ministers of the Environment (CCME) to summarize data collected from Niagara Peninsula Conservation Authority (NPCA) water quality monitoring stations. Image: NPCA..... 177
Figure 4S.2 Temporal distribution of major bacterial phyla in creek (a), river (b), and lake habitats (c). Relative abundances of only top 10 phyla are included..... 178

Chapter 6

Figure 6.1 Location of sampling sites..... 212
Figure 6. 2 Detection of *E. coli* and *Enterococcus* in well water samples (n = 48) at locations in the boil water advisory zone in 2015. 213
Figure 6.3 *E. coli* and *Enterococcus* levels in wells waters B and K compared with the other well water sites. Error bars represent standard deviation of the mean. 214
Figure 6.4 Relative abundance of potential manure and sewage markers in waste samples collected in Wainfleet. Error bars represent standard deviation of the mean. Three biological replicates for each of the sample was used to calculate mean and standard deviation..... 215
Figure 6.5 Relative abundance of potential STE and animal-specific contamination markers in well water sites within the active boil water advisory zone. Error bars represent standard deviation of the mean..... 216
Figure 6.6 Presence of human *Bacteroidales* marker in selected septic tank and groundwater wells within the boil water advisory zone in Wainfleet. Error bars represent standard deviation of the mean..... 217

Figure 6S.1 Standard curve (A) and melt curve (B) from qPCR of HF183 gene. (A) High R^2 value and robust E value indicate the high quality of the amplification reaction and (B) melt curve shows the specificity of the amplification reaction. 218

List of Tables

Chapter 2

Table 2.1 Relative abundance of bacteria and viruses in freshwater 84
Table 2.2 Temporal change of major viral families in the Lakeside and the Long Beach viromes (VLP fraction) 85

Table 2S.1 Major viral families in the Lakeside Beach and the Long Beach viromes (VLP fraction) 86
Table 2S.2 Major viral families in the Lake Ontario and the Lake Erie viromes (VLP fractions) 87
Table 2S.3 Variation in relative abundance of virus families from Lakeside and Long Beach 88

Chapter 3

Table 3.1 Phyla unique to beach sand environment. 113

Chapter 4

Table 4.1 Physicochemical parameters of habitats examined[a] 148

Table 4S.1 Alpha diversity indices for bacterial communities 179
Table 4S.2 Analysis of similarity (ANOSIM) of bacterial communities 180
Table 4S.3 Association between microbial community diversity with factors of interest 181
Table 4S.4 Relative abundance of top twenty families from all habitats 182
Table 4S.5 Differentially abundant bacterial species in habitats 183

Chapter 7

Table 7.1 List of pathogenic viruses identified in the lower Great Lakes region 241

Declaration of Academic Achievement

This dissertation is organized in "sandwich" thesis format and comprises 5 chapters and 2 appendices. Chapter 1 provides a general introduction and a review of the findings of chapters 2, 3, 4 and other recent, related published literature and is being prepared for submission. Chapters 2, 3 and 4 are published chapters. Chapter 5 provides a general conclusion and future directions of the research performed. Appendix A is a co-authored article that employed the methodology used in chapter 4. Appendix B is a book chapter that is derived from the research described in chapter 2 and other relevant literature and is accepted for publication. The individual research articles, book chapters and the contribution of each author in each of these publications is included below.

Chapter 1:	Mohiuddin, M.M. & H.E. Schellhorn, Assessing the diversity of freshwater bacteria and viruses in the metagenomic era. (In preparation for submission).
Comments:	M.M.M and H.E.S. conceived the manuscript; M.M.M wrote the manuscript, H.E.S provided conceptual insights and edits to the manuscript
Chapter 2:	Mohiuddin, M. & H.E. Schellhorn, (2015) Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* 6: 960.
Comments:	M.M.M. and H.E.S. designed the study; M.M.M collected samples, performed laboratory experiments, analyzed data and wrote the manuscript; H.E.S supervised the study and provided edits to the manuscript

Chapter 3:	Mohiuddin, M.M., Y. Salama, H.E. Schellhorn & G.B. Golding, (2017) Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Res* 115: 360-369.
Comments:	M.M.M. and H.E.S. designed the study; M.M.M. collected samples and performed laboratory experiments; M.M.M. and Y.S. performed bioinformatic analyses; M.M.M. and Y.S. wrote the manuscript; H.E.S. and G.B.G. supervised the study and provided edits to the manuscript.
Chapter 4:	Mohiuddin, M.M., S. R. Botts, A. Paschos & H.E. Schellhorn, (2019) Temporal and spatial changes in bacterial diversity in mixed-use watersheds of the Great Lakes region. *J Great Lakes Res* 45(1): 109-118
Comments:	M.M.M. and H.E.S. designed the study; M.M.M. and A.P. collected samples and performed laboratory experiments; S.R.B. provided the QIIME script for bioinformatic analysis; M.M.M. performed bioinformatic and statistical analyses, analyzed data and wrote the manuscript; All authors provided edits to the manuscript
Appendix A:	Naphtali, P., M.M. Mohiuddin, A. Paschos, H.E. Schellhorn, (2018) Application of high-throughput 16S rRNA sequencing to identify fecal contamination sources and to complement the detection of fecal indicator bacteria in rural groundwater. (In press, J Water health)
Comments:	P.N. and H.E.S. designed the study; P.N. and A.P. collected samples; P.N. performed laboratory experiments; M.M.M. provided experimental protocols for amplicon sequencing; P.N. and M.M.M. analyzed data; P.N. and M.M.M. wrote the manuscript; H.E.S. supervised the study and provided edits to the manuscript.
Appendix B:	Mohiuddin, M.M., and H.E. Schellhorn. 2019. Ecology and epidemiology of emerging viruses in two freshwater lakes of the Northern Hemisphere. *In* Emerging and Re-emerging Viral Pathogens (ed. M. M. Ennaji). (In press, *Elsevier*)
Comments:	M.M.M. and H.E.S. conceived the book chapter; M.M.M. wrote the chapter; H.E.S. provided conceptual insights and edits to the chapter.

Chapter 1: Assessing the diversity of freshwater bacteria and viruses in the metagenomic era

Mahi M. Mohiuddin, Herb E. Schellhorn
Department of Biology, McMaster University, Hamilton, ON, Canada

***Correspondence:** Herb E. Schellhorn, LSB 433, Department of Biology, McMaster University,
1280 Main St W, Hamilton, ON, L8S 4K1, Canada.
E-mail: schell@mcmaster.ca

Keywords: Freshwater; Metagenomics; Bacterial Diversity; Viral Diversity; Next-Generation Sequencing

Presented in manuscript format and in under review.

Abstract

The emergence and continuous improvement of next-generation sequencing technologies have expanded our interest in bacteria and viruses from complex environments. While bacteria and viruses are ubiquitous in freshwaters, our understanding of the extent of the diversity remains incomplete. In addition, the rapid improvement of sequencing technologies has led to the development of many methodologies for metagenomic analysis of microbial and viral communities, each with the potential to introduce bias in downstream analyses. In this review, we examine the diversity of freshwater bacteria and viruses, with a focus on the major taxonomic groups as well as providing an overview of methodologies used pertaining to each processing step and general recommendations which will allow researchers obtaining more insight embarking on metagenomic analyses of bacterial and viral communities.

1. Introduction

Freshwater environments are an essential component of many ecosystems which may also supply drinking water, water for agricultural, domestic, industrial and recreational activities and may be important transport routes. In addition, freshwater ecosystems play essential roles in nutrient processing through cycling of terrestrial organic matters (Battin et al., 2009;Tranvik et al., 2018). Viruses and bacteria represent a significant biomass of aquatic ecosystems (Whitman et al., 1998;Farnell-Jackson and Ward, 2003;Suttle, 2005) and play central roles in the transformation of elemental nutrients and, therefore, are considered major players in global biogeochemical and ecological cycles (Fuhrman, 1999;Farnell-Jackson and Ward, 2003;Suttle, 2007;Madsen, 2011).

While bacteria and viruses play important roles in the cycling of nutrients in both freshwater and marine environments, marine environments, compared to freshwater habitats, have received more attention in the past decade (Brum et al., 2015;Sunagawa et al., 2015;Coutinho et al., 2017). This is primarily due to the efforts initiated by marine microbiologists and virologists who investigated the diverse aspects of bacteria and viruses in marine environments in the late twentieth century (microbial ecology), and early twenty-first century (viral ecology) (Giovannoni et al., 1990;Stein et al., 1996;Breitbart et al., 2002;Angly et al., 2006). The relatively small fraction of the earth's surface (<1%) occupied by freshwater and the extensive anthropogenic usage and the diverse nature of inhabiting bacterial and viral populations in marine environments have

also contributed to the marine habitats being the primary focus (Shiklomanov, 1993;Brum et al., 2015;Sunagawa et al., 2015).

Not yet fully explored, freshwater bacterial and viral communities exhibit similar diversity profile as their marine counterparts (Tamames et al., 2010;Roux et al., 2012;Ruiz-Gonzalez et al., 2015a). Freshwater ecosystems often receive inputs from point and non-point sources, which include wastewater, agricultural farm runoffs, stormwater runoffs, and anthropogenic usage. Unlike marine environments, freshwater ecosystems are limited by land boundaries. Because of the lack of flow, the received nutrient inputs from point and non-point sources often lead to a rapid eutrophication of freshwater environments (Smith et al., 1999) which may result in diverse populations of bacteria and viruses in freshwater ecosystems (Cloutier et al., 2015;Korajkic et al., 2015;Chopyk et al., 2018). While salinity is a major determinant in shaping microbial community structure (Lozupone and Knight, 2007), comparison of microbial community diversity between freshwater and marine environments remains incomplete. Comparisons between different aquatic environments have indicated that freshwater environments can harbour more diverse and unique microbial populations than that of marine environments (Rusch et al., 2007;Wang et al., 2012).

During the past decade, freshwater microbial and viral ecology has moved away from using culture-based and targeted PCR-based detection of individual bacteria or viruses towards the identification of total bacterial and viral populations using metagenomic approaches. The metagenomics approach is culture-independent and involves the extraction of DNA from whole bacterial and viral communities and

subsequent sequencing. The metagenomic approach, in combination with a dramatic reduction in the cost of sequencing and the development of standard analysis pipelines in recent years has revolutionized the study of viruses and bacteria in aquatic environments (Muir et al., 2016). In this review, we examine the diversity of freshwater bacteria and viral communities – largely focusing on the major bacterial and viral groups identified in different freshwater environments, host-virus interactions, and key advantages and limitations of metagenomic approaches in capturing freshwater bacterial and viral diversity. We also describe the key technical challenges in producing viral and bacterial metagenomes and the standard analysis pipelines available for analyzing these metagenomes.

2. Freshwater bacterial communities

2.1 Bacterial community structure

Freshwater bacterial communities are primarily enriched in bacteria belonging to the phyla Proteobacteria (mainly of the classes Alpha, Beta and Gammaproteobacteria), Actinobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia and Firmicutes as revealed in several time-scale studies (Tamames et al., 2010;Fortunato et al., 2013;Read et al., 2015;Lee et al., 2016). However, the abundance of each phylum may differ between freshwater habitats. For instance, the Betaproteobacteria class and the Bacteroidetes phylum are enriched in surface water and groundwater (Read et al., 2015;Braun et al., 2016), while the Firmicutes phylum is more abundant primarily in wastewater (Tamames et al., 2010). In addition, chemical inputs associated with different land cover influence the abundance of some bacterial groups, as evident by the

enrichment of Verrucomicrobia phylum in potassium-rich habitats and Acidobacteria phylum in organic carbon contaminated habitats (Staley et al., 2014). Geographic location, water flow, and seasonal changes can also influence the abundance of major bacterial groups in freshwater environments (Roguet et al., 2015;Staley and Sadowsky, 2016). In addition to the major bacterial groups, unclassified phyla often constitute a large proportion (approx. 5-20% of the total bacterial population) of total bacterial communities in freshwater environments (Staley and Sadowsky, 2016;Mohiuddin et al., 2017). These unclassified phyla represent bacterial groups for which full taxonomic classification is not yet available. While the majority of the freshwater bacterial communities present are indigenous to their specific environment (Staley and Sadowsky, 2016), bacterial populations (and pathogens) can also be introduced into aquatic environments through external sources including point source (wastewater) (Vijayavel et al., 2010) and non-point-source pollution (Piggot et al., 2012). These non-endogenous bacteria adapt to local ecosystems and become part of the naturalized community (Russell et al., 2012;Whitman et al., 2014). Therefore, understanding the microbial communities of aquatic environments may provide important insight into their ecology.

2.1.1 Proteobacteria

Among the main freshwater bacterial phyla, Proteobacteria is the most abundant phylum in most freshwater environments. Proteobacteria consists of seven major classes including Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, Deltaproteobacteria, Epsilonproteobacteria, Acidithiobacillia and Zetaproteobaceria. Among these seven classes, Alpha-, Beta-, and Gammaproteobacteria are three major

classes of Proteobacteria commonly detected in most freshwaters (Ghai et al., 2011;Kolmakova et al., 2014;Read et al., 2015;Staley and Sadowsky, 2016;Mohiuddin et al., 2017). Compared to other bacterial groups, members of the Betaproteobacteria class are by far the most abundant bacteria in aquatic environments. Their high abundance and their amenability to culturing have contributed to the detailed understanding of the roles of this particular bacterial group in many aquatic environments (Newton et al., 2011). While mostly free-living and planktonic, Betaproteobacteria also harbor groups such as Polynucelobacter spp. that form a symbiotic relationship with protists (Croue et al., 2013;Vannini et al., 2017) and these symbiotic Betaproteobacteria are perhaps the best-understood group in aquatic environments. Free-living Betaproteobacteria such as *Limnohabitans* spp. and Alcaligenaceae family are abundant in freshwater environments and changes in the abundance of *Limnohabitans* spp. and Alcaligenaceae can also indicate recent environmental disturbances in freshwater lakes and rivers (Balmonte et al., 2016;Sharuddin et al., 2017). Alphaproteobacteria are ubiquitous in both marine and freshwater environments. The SAR11 clade of Alphaproteobacteria can constitute up to 25-50% of total bacterial populations in marine water (Morris et al., 2002). Compared to their marine cousins, freshwater SAR11 clade (LD12) are genetically distinct (Logares et al., 2010) and their abundance on the global scale is still unclear. Although the LD12 clade was found to be as ubiquitous as marine SAR11 clade in some freshwater environments (Salcher et al., 2011;Eiler et al., 2014), the abundance of this group may not be as high in other environments as indicated by the higher abundance of Beta-, and Gammaproteobacteria than Alphaproteobacteria (Logares et al., 2010;Ghai et al.,

2011;Wang et al., 2012;Uyaguari-Diaz et al., 2016;White et al., 2016). Other notable members of Alphaproteobacteria include *Novosphingobium* spp. and *Sphingopyxis* spp. that are commonly detected in humic lakes (Salcher et al., 2013;Chen et al., 2015). Members of Alphaproteobacteria are often found in higher abundance in low nutrient environments and are associated with degrading organic carbon compounds (Salcher et al., 2013;Salka et al., 2014;Henson et al., 2018). Gammaproteobacteria, another class of Proteobacteria, are also abundant in freshwater environments, however, members of this group are more abundant in marine water (Zwart et al., 2002;Sunagawa et al., 2015). The high abundance of Gammaproteobacteria in oceans or salt water is mostly due to the high abundance of SAR86 clade (Sunagawa et al., 2015). Members of Gammaproteobacteria that include *Escherichia coli*, *Shigella s*pp., *Salmonella* spp., *Yersinia* spp., *Pseudomonas* spp., *Legionella* spp., *Enterobacter, Klebsiella* spp. are often associated with waterborne diseases and are, therefore, used as water quality indicators across the globe (Harwood et al., 2014;Whitman et al., 2014). Deltaproteobacteria and Epsilonproteobacteria are also enriched in some freshwater ecosystems. Deltaproteobacteria which comprises both aerobic and anaerobic bacteria are often present in high abundance in nearshore region of lakes (Staley and Sadowsky, 2016), in pore water (interstitial water between sand particles) (Keshri et al., 2018), and in the upper layer of sediments (Dai et al., 2016). Deltaproteobacteria are often enriched in sand and sediments. Beach sand are washed with wave action, and this wave action subsequently washes off the Deltaproteobacteria from sand into the water column and subsequently leads to an increase in abundance of Deltaproteobacteria. Epsilonproteobacteria, another class of Proteobacteria, are

commonly detected in both water and sediments (Briee et al., 2007;Liu et al., 2014;Noguerola et al., 2015). However, Epsilonproteobacteria are more abundant in water than in sediments (Briee *et al.*, 2007). Epsilonproteobacteria are also present in higher abundance in runoffs from agricultural farms and wastewater treatment plants (Engberg et al., 2000;Hutchison et al., 2005). These runoffs often drain into the nearby freshwater reservoirs (such as lakes, rivers and canals) and results in higher abundance of Epsilonproteobacteria in these environments (Mohiuddin et al., 2019).

2.1.2 Actinobacteria

Actinobacteria, another major group is also highly abundant in many freshwater environments. While being the second most dominant phylum in freshwater environments, Actinobacteria can contribute to more than 50% of the total bacterial population in some freshwater environments (Glockner et al., 2000). The high abundance of Actinobacteria in freshwater ecosystems is interesting since historically, soil is considered to be the primary habitat of Actinobacteria (Goodfellow and Williams, 1983). However, since the first detection of Actinobacteria in the water column of Adirondack mountain lakes (Hiorns et al., 1997;Glockner et al., 2000), presence of this phylum has been reported in many freshwater environments across the globe indicating that Actinobacteria are ubiquitous in aquatic environments (for a comprehensive review on freshwater bacteria (Newton *et al.*, 2011)). The high abundance of Actinobacteria in freshwater is primarily due to the enrichment of two orders of Actinobacteria, Actinomycetales and Acidimicrobiales (Ghai et al., 2014). Members of these two orders are planktonic and photoheterotrophic and, thereby explains the high abundance of

Actinobacteria in aquatic environments (Ghai et al., 2014). Genomic features of freshwater Actinobacteria also differ from their terrestrial counterparts and 16S rRNA gene-based analysis have indicated that freshwater Actinobacteria often cluster in distinct lineages (Warnecke et al., 2004). Compared to the terrestrial Actinobacteria for which the G+C content is high (~51-70%) (Ventura et al., 2007), the G+C content of freshwater Actinobacteria is low (~42%) (Ghai et al., 2012;Neuenschwander et al., 2018).

2.1.3 Bacteroidetes

Members of this phylum are commonly known as environmental bacteria and are found in high abundance in freshwater lakes, rivers, streams, and sediments. However, the most abundant members of this group are the *Flavobacteriaceae* family. Members of this group are mostly free-living and cultivable (Riemann and Winding, 2001;Zeder et al., 2009), however, some can form a symbiotic relationship (Frias-Lopez et al., 2002;Faust et al., 2012). Another major group of Bacteroidetes are the *Bacteriodales* and members of this group are often used for microbial source tracking (MST), an approach commonly used to trace fecal pollution of aquatic environments.

2.1.4 Cyanobacteria

Toxin-producing Cyanobacteria are responsible for harmful algal blooms (HABs) which pose a threat to water quality. The worldwide increase in HABs over the past decade poses a risk to both drinking and recreational water quality (Cheung et al., 2013). Lake ecosystems are subject to many anthropogenic pressures including wetland loss, shoreline development, nutrient over-enrichment (eutrophication), and contaminant inputs. While larger and deeper freshwater environments are less prone to eutrophication,

smaller and shallower environments are susceptible to increased nutrient loading. Factors promoting CHABs are relatively well-understood and include nutrient over-enrichment, primarily with nitrogen and phosphorous (Davis et al., 2009;Rigosi et al., 2014;Müller and Mitrovic, 2015) and human activities that lead to the eutrophication of freshwater environments (Paerl, 2014). Several potent toxins are produced by Cyanobacteria, with microcystin being the most potent. Microcystin is produced by a wide range of cyanobacteria including *Microcystis, Planktothrix, Anabaena* (currently called *Dolichospermum*) (Wacklin et al., 2009) and *Nostoc* species. Cyanobacteria are common in recreational waters and sand interstitial waters (Rinta-Kanto and Wilhelm, 2006;Hotto et al., 2007;Davis et al., 2015;Mohiuddin et al., 2017) Among the Cyanobacteria present in freshwater environments, *Microcystis aeruginosa* is dominant, followed by *Planktothrix* spp., *Nostoc* spp., and *Synechococcus* spp. (Rinta-Kanto and Wilhelm, 2006;Hotto et al., 2007;Davis et al., 2015;Lee et al., 2015).

2.1.5 Firmicutes

Firmicutes, another major bacterial group which constitute the normal microbial flora of human and mice gut, are abundant in freshwater environments. While Firmicutes are common in freshwaters, sewage-contaminated lakes, rivers, or streams exhibit a higher abundance of Firmicutes than in pristine freshwaters (Saarenheimo et al., 2017;Chu et al., 2018). Member of this group such as *Clostridiaceae* and *Lachnospiraceae* families are present at high concentration in untreated sewage (McLellan et al., 2010;McLellan et al., 2013;Shanks et al., 2013). These groups, although

not currently included in traditional water quality monitoring scheme, have the potential to serve as biological indicators of fecal contamination in aquatic environments.

2.1.6 Verrucomicrobia

Verrucomicrobia, a member of the PVC superphylum, are ubiquitous in freshwater environments, however, their abundance is higher in freshwater lakes and rivers (Zwart *et al.*, 2002). Verrucomicrobia are present in most freshwater lakes (up to 90%) and their abundance can range between 1% and 41% of total bacterial communities in freshwater lakes (Chiang et al., 2018;Tran et al., 2018) and up to 20% of total bacterial populations in soil (Bergmann et al., 2011). Verrucomicrobia are also present at a higher abundance (approx. 19% of total bacterial populations) in humic lakes (Arnds et al., 2010). Verrucomicrobia degrade polysaccharide and fix nitrogen in freshwaters (Cabello-Yeves et al., 2017;He et al., 2017), and therefore, play a direct role in the cycling of both carbon and nitrogen in aquatic habitats.

2.1.7 Other bacterial phyla

In addition to the major bacterial phyla commonly detected in most freshwater environments, freshwater bacterial communities also harbor minor bacterial groups that include Acidobacteria, Armatimonadetes, Chlamydiae, Cladithrix, Chlorobi, Chloroflexi, Crenarchaeota, Deinococcus-Thermus, Elusimicrobia, Fibrobacteres, Firmicutes, Fusobacteria, Gemmatimonadetes, Lentisphaerae, Nitrospirae, Planctomycetes, Spirochaetes, Tenericutes, and members of candidate phyla (Ji et al., 2018;Proctor et al., 2018;Wang et al., 2018). Candidate phyla consist of bacterial groups that do not have cultured representatives and this particular group consists of bacteria from phyla that

include Latescibacteria (WS3), Omnitrophica (OP3), Microgenomates (OP11), Aminicenantes (OP8), Katanobacteria (WWE3), Parcubacteria (OD1), Berkelbacteria (ACD58), Saccharibacteria (TM7), WS6, Peregrinibacteria (PER), and Kazan phyla (Kumar and Saravanan, 2010;Farag et al., 2014;Brown et al., 2015;Farag et al., 2017). However, because of their low abundance, many of these minor phyla are often referred as "other phyla" in most metagenomic studies (Staley and Sadowsky, 2016;Wang et al., 2016) and, hence, the abundance of these minor bacterial groups and their potential role in freshwater environments is still unclear. Some of these candidate phyla such as Latescibacteria, Aminicenantes, and Microgenomates are often detected in hydrocarbon impacted and in anoxic environments (Kumar and Saravanan, 2010;Farag et al., 2014;Youssef et al., 2015;Hu et al., 2016) suggesting their potential role in the carbon and hydrogen cycle. Among these minor phyla, bacterial species belonging to the phyla Acidobacteria, Chlorobi, Chlamydiae, Chloroflexi, Nitrospirae, and Planctomycetes are commonly detected in freshwaters (Wang et al., 2012;Wang et al., 2016;Mohiuddin et al., 2019).

2.2 Diversity of bacterial communities

The top five bacterial phyla represent 60-80% of total bacterial sequences in most freshwater environments, and, therefore, the diversity of bacterial communities is difficult to capture at a broad taxonomic level (such as the phylum level). Resolving taxonomy at a higher resolution (lower taxonomic levels such as at the family or genus level) provide more information about bacterial diversity as well as niche specific information. In-depth analyses of bacterial communities have indicated that bacterial community diversity may

differ significantly between different regions (from regional to continental-scale) (Nemergut et al., 2011;Roguet et al., 2015;Staley and Sadowsky, 2016;Mohiuddin et al., 2019). The diversity of bacterial communities may also differ on a temporal scale (Liu et al., 2018;Mohiuddin et al., 2019).

The differences in diversity are primarily due to the deterministic and neutral processes that influence the shape of microbial community structure (Lee et al., 2013;Dini-Andreote et al., 2015;Pagaling et al., 2017). Deterministic process, also known as species sorting, predicts environmental conditions (interactions between biotic and abiotic factors) and spatial factors that shape microbial community structure (Leibold, 1995) whereas neutral processes assume that the microbial community assembly within a trophic level is determined by dispersal (migration) and speciation through natural means (McKane et al., 2004). Freshwater environments receive input from many different sources and these sources often differ in nutrient and microbial loadings across regions which ultimately impact the diversity of bacterial populations in freshwater ecosystems. Among the physicochemical factors, pH is the primary determinant in shaping bacterial community diversity (Lindstrom et al., 2005;Lliros et al., 2014;Liu et al., 2015b;Ren et al., 2015). Other abiotic factors that also influence bacterial diversity include temperature (Lindstrom et al., 2005;Roguet et al., 2015), dissolved oxygen (DO) concentration (Spietz et al., 2015;Mohiuddin et al., 2019), organic carbon content (Lliros et al., 2014;Ruiz-Gonzalez et al., 2015), total phosphorous (TP) (Jones and Lennon, 2010;Eiler et al., 2014), total nitrogen (TN) (Wilhelm et al., 2011;Logue et al., 2012), and conductivity (Szekely et al., 2013;Eiler et al., 2014). Agricultural practices, runoffs from agricultural

farms and wastewater treatment plants can also influence microbial community composition in freshwaters (Cloutier et al., 2015;Mohiuddin et al., 2019). Bacterial competition for resources and space with other microorganisms and predation of bacteria by protozoan grazers and viruses can also impact the diversity of aquatic microbial communities in aquatic environments (Berdjeb et al., 2011;Andersson et al., 2018).

3. Freshwater viruses

3.1 Viral community structure

Viruses are the most abundant and the most diverse microorganism on earth (Guemes et al., 2016). Viruses infect all domains of life including archaea, bacteria and eukaryotes (Rohwer and Thurber, 2009). The dynamics and the diversity of such large viral communities, their role in aquatic environments and the interactions between viruses and their diverse hosts have been the focus of viral ecology research in past three decades. It is now well-established that viruses influence processes ranging from global biogeochemical and ecological cycles (Suttle, 2007;Roux et al., 2016) to host virulence and pathogenesis (Brussow et al., 2004). Viruses modulate their hosts through infection (Evans and Brussaard, 2012), mortality (Payet and Suttle, 2013), horizontal gene transfer (Brouwer et al., 2013;Winstel et al., 2013;Moon et al., 2016), expression of accessory host proteins (Frye et al., 2005;Hargreaves et al., 2014), and niche adaptation (Chibani-Chennoufi et al., 2004a;Sullivan et al., 2006). However, similar to marine microbial communities, marine viruses and their ecological aspects have received more attention

than freshwater viruses. Freshwater viral communities, although less complex and less abundant than marine viral communities, harbour a diverse range of viruses.

Viruses outnumber their prokaryotic hosts by a ratio of fourteen to one in freshwater environments (Guemes *et al.*, 2016). The majority of the aquatic viruses identified are bacteriophages (viruses that infect bacteria) (Paez-Espino et al., 2016). In addition to bacteriophages, aquatic environments harbour a wide range of diverse archaeal viruses (Krupovic et al., 2018). Human and animal viruses and to a lesser extent, plant viruses are also present in aquatic environments (Fancello et al., 2013;Tseng et al., 2013;Mohiuddin and Schellhorn, 2015). Exploring the diversity of such highly abundant and complex viral communities may, therefore, shed light into the ecology and the role of viruses in aquatic environments.

3.1.1 DNA Viruses

3.1.1.1 Tailed phages

With a proportion of 60-93% of total viral populations, tailed phages are by far the most abundant DNA viruses in freshwater environments. Tailed phages are diverse in their morphotypes and there are three major families of tailed phages- Myoviridae (long contractile tails), Siphoviridae (long, noncontractile tails), and Podoviridae (short tails). These families belong to the order Caudovirales and among these three families, Siphoviridae is the most abundant (60%) followed by Myoviridae and Podoviridae (25% and 15% respectively) (Clokie et al., 2011). However, the abundance of theses tailed viral families vary between freshwater environments suggesting that geographic location doesn't seem to influence the abundance of these viral families. In most freshwater

environments Siphoviridae (Roux et al., 2012;Gong et al., 2018) and Podoviridae (Cai et al., 2016;Skvortsov et al., 2016) predominate whereas Myoviridae predominates in others (Matteson et al., 2011;Mohiuddin and Schellhorn, 2015).

Among the Caudovirales, cyanophages dominate the tailed phage community in both freshwater and marine environments (Wilhelm et al., 2006;Roux et al., 2012;Mohiuddin and Schellhorn, 2015;Skvortsov et al., 2016). Cyanophages are abundant in the environment and the concentration of cyanophages can be as high as 8.8 $\times 10^7$ pfu ml^{-1} in freshwater ecosystems (Matteson *et al.*, 2011). Such a high concentration of cyanophage in some environments require in-depth exploration since the concentration of viruses in most aquatic environments range between 10^6 to 10^8 PFU per ml of water. Cyanophages of diverse morphotypes are present in aquatic environments, with phages belonging to the Myoviridae, Siphoviridae and Podoviridae family. Non-tailed cyanophages, although not very common as tailed cyanophages, are also present in aquatic ecosystems (Gao et al., 2012). However, the majority of the cyanophages identified to date belong to the Myoviridae family (Roux et al., 2012;Ma et al., 2014;Mohiuddin and Schellhorn, 2015).

3.1.1.2 Non-tailed DNA viruses

Non-tailed viruses are viruses with icosahedral or quasi-icosahedral, rod-shaped bodies. The genomic content of these viruses can be either DNA or RNA and their nucleic acid can be either single-stranded or double-stranded (Ackermann, 2007). Host range of non-tailed viruses include bacteria, plants, algae, birds, humans and other mammals. Major families of polyhedral viruses include Microviridae, Circoviridae,

Naonoviridae, Inoviridae, Geminiviridae, Phycodnaviridae, and Tectiviridae and these viruses are often found in freshwater viral metagenomic datasets at a higher abundance (Roux et al., 2012). Although tailed phages predominate most freshwater viromes, non-tailed viruses were found to be as high as 70% of total viral populations in Lake Bourget and Lake Pavin (Roux et al., 2012). However, non-tailed single-stranded DNA viruses are often present at a higher abundance in polar freshwater lakes and ponds (both the Arctic and the Antarctic) (Lopez-Bueno et al., 2009;Zawar-Reza et al., 2014;de Carcer et al., 2015;Gong et al., 2018).

Some of the non-tailed viruses commonly found in freshwater metagenomes can also be grouped as giant viruses (also known as nucleocytoplasmic large DNA viruses or NCLDVs). Compared to other viruses, giant viruses possess extremely large genomes and often carry genes that are common in bacteria and eukaryotes (Arslan et al., 2011;Aherfi et al., 2016). These giant viruses are often visible under the optical microscope and include members of the family Phycodnaviridae, Mimiviridae, Marseilleviridae, Pandoraviridae, Iridoviridae and Poxviridae. While the giant viruses that primarily infect amoeba are well studied (Philippe et al., 2013;Aherfi et al., 2016;Yoshikawa et al., 2019), many giant viruses that infect insects, algae, humans and animals are also present in aquatic environments (Popgeorgiev et al., 2013;Yolken et al., 2014;Maruyama and Ueki, 2016). Although giant viruses are common in viral metagenomes, the proportion of giant viruses do not exceed 15% of total viral populations (Roux et al., 2012;Mohiuddin and Schellhorn, 2015;Cai et al., 2016).

3.1.1.3 Virophages

Virophages are small, icosahedral, dsDNA viruses that infect giant viruses. These viruses are considered obligatory parasites of giant viruses because they co-infect the unicellular hosts of giant viruses and are dependent on giant viruses for replication. Virophage infection leads to the deformation and abortion of giant viruses and increased survival of the eukaryotic hosts. These parasitic viruses belong to the Lavidaviridae family which contains two genera – Sputnikvirus and Mavirus (Krupovic et al., 2016). Virophages attenuate Phycodnaviruses that infect algae and in turn, may control harmful algal blooms (Yau et al., 2011). Virophages, although identified in metagenomic datasets from the soil, sediments, bioreactors, marine and freshwater environments, the abundance of virophages are highest in freshwaters (Zhou et al., 2013). While still new to viral ecologists and new viruses are consistently being identified, continuous research has indicated that these parasitic viruses share high sequence similarity to their host giant viruses (Phycodnaviridae and Mimiviridae) suggesting horizontal gene transfer between these viruses and their hosts (La Scola et al., 2008;Yau et al., 2011;Zhou et al., 2013;Gong et al., 2016).

3.1.2 RNA Viruses

While metagenomic studies of viruses are mainly focused on the identification of DNA viruses, only a few studies have investigated the RNA viruses in freshwater environments – one in Lake Needwood (Maryland, USA) and another in Lake Limnopolar (Antarctica) (Djikeng et al., 2009;Lopez-Bueno et al., 2015). Among the RNA viruses, members of the Picornaviridae and Dicistroviridae families dominate RNA

viromes in aquatic environments (Djikeng et al., 2009;Lopez-Bueno et al., 2015). Members of the Nodaviridae, Herpeviridae, and Leviviridae are also present to some extent in RNA viral metagenomes. RNA viral metagenomes show a clear distinction from DNA viral metagenomes – bacteriophages predominate DNA viral communities whereas viruses of insects, mammals and other animals, and plants predominate RNA viral metagenomes (Djikeng et al., 2009;Lopez-Bueno et al., 2015).

3.2 Diversity of viral communities

Diversity of viruses are often expressed as the abundance profile of major viral families. However, the majority of the viruses identified in viral metagenomes are dsDNA viruses (or tailed phages) belonging to the Myoviridae, Siphoviridae and the Podoviridae family (Mohiuddin and Schellhorn, 2015;Cai et al., 2016;Skvortsov et al., 2016). These viruses comprise >60% of total viromes in most freshwater environments. Differences in the abundance of major viral families were observed in some freshwater environments– particularly in Lake Bourget (France) and Lake Pavin (France) and in polar freshwaters (both the Arctic and the Antarctica) (Lopez-Bueno et al., 2009;Roux et al., 2012;de Carcer et al., 2015). These environments share a similar abundance profile of viruses with ssDNA and non-tailed dsDNA phages being present at a higher relative abundance than most other freshwater habitats.

Freshwater environments often exhibit temporal changes in viral diversity. However, these differences in diversity are mostly limited to the changes in the abundance profile of major viral families in a given environment (Tseng et al., 2013;Mohiuddin and Schellhorn, 2015). An exception to this phenomenon is the

Antarctic viromes (Lopez-Bueno *et al.*, 2009). While ssDNA phages predominate the Antarctic freshwater viromes in the spring months, tailed phages (dsDNA viruses) are present at a higher abundance in the viromes collected in the summer months. Differences in viral diversity can be linked to the changes in host-virus dynamics (Arkhipova et al., 2018). Phages (or viruses) require hosts for replication. Since most of the viruses present in viromes are bacteriophages, changes in microbial communities directly impact the diversity of viral populations in a given environment. Therefore, factors (such as salinity, temperature, pH, dissolved oxygen content, nutrient availability, and climactic disturbances) that impact the diversity of bacterial populations can also impact the diversity of viruses (Almeida et al., 2015;Cabral et al., 2017;Hanson et al., 2017). In addition, factors that influence the lysogenic conversion of lytic phages may also play a role in the diverse profile of viruses in aquatic environments (Palesse et al., 2014;Knowles et al., 2016). The diversity of viruses that infect eukaryotes (algae, for instance) also changes depending on seasonal host dynamics (Lopez-Bueno et al., 2009;Wang et al., 2015;Long and Short, 2016).

4. Metagenomics- advantages and limitations

With the advances in sequencing technologies, a sharp reduction in costs of sequencing, coupled with additional technological advances, high-throughput metagenomic analysis of complex bacterial and viral communities has revolutionized the study of bacteria and viruses in the past decade. Bacterial and viral communities can be extremely diverse, can inhabit every possible environment and the culture methods for the majority of the bacteria and viruses are still unavailable (Paez-Espino et al.,

2016;Thompson et al., 2017). Metagenomic approaches are culture independent and, therefore, overcome many limitations of culture-based methods and offer unrestricted access to the total microbial and viral populations in any given environment. While metagenomic approaches are comprehensive and effective in determining microbial and viral diversity, challenges remain in the application of metagenomic approaches. These challenges primarily arise from the goal of a study and the resources available to address the specific needs. Common challenges that are often encountered include the costs of sequencing, low DNA yield from environmental samples, expertise required to analyze sequence data, and the lack of a standard metagenomics workflow.

4.1 Factors that impact metagenomic studies

The widespread use of metagenomic approaches has led to the development of many methodologies that introduce systematic biases in metagenomic studies. Such biases can introduce differences in the observed community diversity and impact the subsequent interpretation of metagenomic datasets. From study design to sample collection, sample processing (fresh vs frozen samples), DNA extraction methods, amplification, purification and library preparation methods, choice of sequencing technology and/or platform to be used and the final bioinformatic analyses pipelines to be used, there are many methodological choices to be made (discussed in details below). Use of different methodology often limits the comparability and/or reproducibility of datasets that are generated from the same or related environments. Here we review current literature that include key information about experimental design for metagenomic studies.

4.1.1 Sequencing strategy

There are two primary approaches available for metagenomic studies and these include shotgun metagenomic sequencing and amplicon sequencing. Amplicon sequencing uses marker genes such as 16S rRNA, 5S rRNA, 23S rRNA and cpn60 as target regions to identify microbial community diversity (Pace, 1997;Hunt et al., 2006;Pei et al., 2012;Johnson et al., 2015). Among these marker genes, the 16S rRNA gene is ubiquitous in bacteria, and since its first use in 1997 on environmental samples in capturing bacterial diversity, the 16S rDNA gene has been the most commonly used marker gene to identify bacterial communities (Pace, 1997). Amplicon sequencing is fast and inexpensive and is useful in assessing taxonomic diversity at a broader level (Tessler et al., 2017). However, amplicon sequencing does not provide detailed functional characteristics of microbial communities and is not applicable to viruses because of the lack of a universal signature marker gene in viruses. Shotgun metagenomic sequencing, on the other hand, provides detailed information about functional characteristics of microbial communities and shotgun metagenomic sequencing is the only method that can provide taxonomic profile as well as functional characteristics of viral communities.

While shotgun metagenomic sequencing can also be used to assess the biodiversity, unlike amplicon sequencing, shotgun metagenomic sequencing is expensive. In addition, obtaining similar community diversity as obtained from amplicon sequencing, shotgun metagenomic sequencing have to a be performed at a higher depth which can be even more expensive and may not be economically feasible for many labs (Clooney et al., 2016). Signature sequence genes used in amplicon sequencing, however,

often fall short of resolving taxonomy at a fine level (such as species level) whereas a mid-depth shotgun metagenomic sequencing can identify microbial communities at the species level (Clooney *et al.*, 2016). Therefore, both these approaches have advantages and limitations and the choice of sequencing technology to be used depends primarily on the study requirement.

4.1.2 Sequencing platform

The decision as to which sequencing platform to be used is the first consideration in a microbial or viral metagenomics project. This decision is often influenced by factors such as the length of amplicons, desired sequence length (for shotgun), data output, cost per sequencing run, and error rate of each platform (sequence quality). Depending on the resource availability and the requirement, several options are available- the Pacific Biosciences RSII, the Illumina HiSeq (most popular for shotgun metagenomic sequencing), NovaSeq, NextSeq, MiSeq (most popular platform for amplicon sequencing) and the Thermo Fisher Scientific Ion PGM (Personal Genome Machine) and SOLiD. Each of these platforms has its strengths and weaknesses (for a review on sequencing technologies, see (Goodwin et al., 2016)). The use of different sequencing platforms (D'Amore et al., 2016) and the use of different library preparation methods within the same platform (Jones et al., 2015) were shown to introduce biases in genomic and functional profile of microorganisms. Differences were seen in species counts between Illumina MiSeq and the Ion PGM platforms when evaluated for 16S amplicon sequencing approach (Clooney et al., 2016;Fouhy et al., 2016). This suggests that a degree of caution should be applied when comparing datasets generated by different sequencing platforms

as well as by different library preparation methods within the same platform. Therefore, the choice of sequencing platforms and the standardization of experimental design remains a critical consideration before embarking on any metagenomic study.

4.1.3 Sample collection

The first step in any metagenomic study is the sample collection. While metagenomic approaches are commonly used to assess bacterial and viral diversity if freshwaters, there is no standard guideline regarding the volume of samples to be collected. This is primarily due to the difference in the abundance of bacteria and viruses in freshwaters and the amount of DNA extracted from water samples. The volume of samples to be processed depends whether DNA extracted from the sample is enough for metagenomic sequencing as well as represent most of the bacteria and viruses present in the given environment. For bacteria, collection of 300 mL- 2.0 L water samples is a common practice whereas, for viruses, 300 mL to 20 L water samples are commonly processed. Sample processing also adds another variable in metagenomic studies. Another variable that is often encountered in metagenomic studies is the difference in microbial diversity between fresh and flash-frozen samples.

4.1.4 Sample processing

The next step to sample collection is the filtration of water samples. Filtration of water samples is usually achieved using 0.22 µm pore-size polycarbonate membrane filters. If filters get clogged, sequential filtration of water samples is recommended (Staley et al., 2013). During sequential filtration, water samples are filtered sequentially through 5 µm, 0.8 µm and 0.2 µm membrane filters to ensure large particles are removed

from water samples before bacterial fraction is filtered. The filter paper that contains bacterial fraction is then used to extract DNA using commercially available kits.

Sample processing for viruses, however, is trickier than for bacteria and requires additional steps. The first step after sample collection is the removal of any cellular fraction and this is usually achieved by filtering the water samples with 0.10-0.45 µm pore-size filters. Intermediate filtration steps may be required in some case to prevent filter clogging and this is done by using 1.0-5.0 µm pore-size filters (Pesant et al., 2015). Using 0.10-0.20 µm pore-size filters is the most common approach, however, use of such filters also introduce some bias, excluding or under-representing large viruses (Philippe *et al.*, 2013). Filtration through 0.45 µm pore-size filters allows some cellular fractions to pass. Therefore, generating a complete profile of viruses requires a combination of filters. Careful consideration is required when removing cellular fractions from water samples since high pressure can damage the capsid of virus particles and result in the under-representation of viruses. The next step after filtration is the concentration of virus-like particles (VLPs) from filtered water samples. The most common methodologies used for this purpose include the tangential flow filtration (TFF) (Cai et al., 2015) and iron-chloride ($FeCl_3$) flocculation (John et al., 2011). Use of these approaches reduces the volume of the water samples to 10-500 mL and often requires additional concentration and this additional concentration is achieved through ultracentrifugation, PEG (polyethylene glycol) precipitation or CsCl centrifugation.

The next step is the extraction of nucleic acid from concentrated virus particles. Depending on the project goal, the nucleic acid could be either DNA or RNA. Before the extraction of nucleic acid, contaminating cellular or free DNA or RNA should be removed to avoid difficulties in downstream bioinformatic analyses (Thurber et al., 2009;Roux et al., 2013). The nucleic acid can be extracted using commercially available kits. Although viruses are abundant in environmental waters, their small genome size often limits the amount of nucleic acid extracted. While the advancement of library preparation methods and sequencing technologies have reduced the requirement of starting nucleic acid amount, the low yield of nucleic acid could still pose a problem and amplification of nucleic acids may become necessary. There are several methods available for amplification of nucleic acids and these include multiple displacement amplification (MDA) (Dean et al., 2002), linear amplification for deep-sequencing (LADS) (Hoeijmakers et al., 2011), linker amplified library construction (LA) (Duhaime et al., 2012), and sequence-independent single-primer amplification (SISPA) (Karlsson et al., 2013)- each having some advantages and limitations. The amplified nucleic acid is then used for library preparation which is described below.

4.1.5 Library preparation

The choice of library preparation method adds another variable in a metagenomic study. The first step of library preparation methods (for amplicon sequencing) is the choice of 16S rDNA hypervariable region. The 16S rRNA gene consists of 9 hypervariable regions (V1-V9). Sequencing of the full-length 16S rRNA region is not possible with most second-generation short-read sequencers. In addition, currently there

is no consensus regarding a 16S rRNA region that can capture the complete bacterial profile and each region shows bias towards specific bacterial groups. For example, Deltaproteobacteria (Phylum Proteobacteria) are overrepresented in the V4 and V7-V8 region whereas the members of the same class are underrepresented in the V6-V8 region (Tremblay et al., 2015). Verrucomicrobia, another bacterial phylum commonly detected in freshwaters and Sphingobacteria were shown to be overrepresented in the V4 region as compared to the V6-V8 and V7-V8 region. In addition, underrepresentation of Gammaproteobacteria and/or overrepresentation of Firmicutes were observed with the V7-V8 region (Tremblay et al., 2015). Because of the bias of 16S rRNA hypervariable regions toward specific bacterial groups, different combinations of these hypervariable regions were evaluated in separate experiments (Tremblay et al., 2015;Fouhy et al., 2016) and primer pairs that target the V3-V4, V4 and V4-V5 regions were shown to produce comparable results than other regions (Klindworth et al., 2013;D'Amore et al., 2016;Teng et al., 2018). Recently, sequencing of full-length 16S rRNA gene amplified through SMRT (single-molecule real-time amplification) or using approaches that combine sequences generated from the amplification of multiple regions of the 16S rRNA gene and/or amplifying multiple hypervariable regions of the 16S rRNA gene were shown to be more effective in capturing more complete profile of microorganisms (Singer et al., 2016;Fuks et al., 2018;Schriefer et al., 2018).

Choice of the primer pair is the most critical decision in an amplicon sequencing workflow. It is now well documented in literature that primer specificity can introduce bias in a metagenomic study. Primer design that includes the length of the primer,

sequence, position, degenerations and combinations can impact the richness and the diversity estimates of microbial communities (Klindworth et al., 2013;Tremblay et al., 2015;D'Amore et al., 2016). Based on trials, several primer pairs are recommended for the analysis of microbial communities from complex environmental samples and these include the primer pairs used for the Earth Microbiome Project, 515F (Parada)-806R (Apprill) and 515F (Parada)-926R (Quince) (modified from the original primer pair 515F-806R as used in (Caporaso et al., 2011)) (Apprill et al., 2015;Parada et al., 2016), Human Microbiome Project (357F-926R) (Muyzer et al., 1993) and *in silico* and experimentally evaluated primer pair (S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21) (Klindworth et al., 2013). However, primer pairs amplify the corresponding 16S rRNA target region and, therefore, using a primer pair that amplify a single 16S rRNA region will have similar bias as described above. Another consideration in choosing a primer pair is the amplicon length. Choice of sequencing platforms can limit the length of the amplicon to be sequenced. Therefore, to avoid biases associated with primer pairs and/or 16S rRNA hypervariable regions, one should choose primer pairs that amplify different hypervariable regions (preferable V3, V4 and V5 region) and the length of the amplicon remains within the capacity of the sequencing platform. This can ensure avoiding biases associated with primer pairs as well as can provide a better estimate of bacterial community diversity. Another consideration in choosing a primer pair is the amplicon length and, the choice of sequencing platforms can also limit the choice of primer pair.

In addition to the primer pairs, factors that impact amplification reactions such as presence of PCR inhibitors (organic and inorganic matters) (Schrader et al., 2012),

number of PCR cycles (Wu et al., 2010;Ahn et al., 2012), use of high fidelity polymerases (Gohl et al., 2016) and PCR method (one-step vs two-step) (Miller et al., 2013;Sinclair et al., 2015) can lead to skewed interpretation of datasets. However, most commercial kits available can minimize the presence of PCR inhibitors during DNA extraction steps. If amplification is impossible to achieve, an alternative method such as metabarcoding, which does not rely on amplification of DNA can be utilized (Taberlet et al., 2012). While the number of PCR cycles can lead to an increased species abundance, PCR cycle number do not have an impact on the overall community structure (Wu et al., 2010;Ahn et al., 2012). Although higher cycle number does not make any change to bacterial community structure, higher cycle numbers often lead to increased chimera formation and this can be avoided by lowering PCR cycle number (Kanagawa, 2003). Careful consideration is required when comparing datasets generated by one-step and two-step PCR methods. While both approaches have some advantages and limitations, two-step PCR methods were shown to introduce differences in alpha and beta diversity among replicates (Sinclair et al., 2015). Use of high-fidelity polymerases (polymerases with proof-reading activity) were shown to minimize substitution errors and, thereby, improve the accuracy of amplification by reducing the number of spurious OTUs (operational taxonomic units) (Gohl et al., 2016). In addition, high-fidelity polymerases can identify additional taxa that are not detected by conventional polymerases (Gohl et al., 2016).

Shotgun metagenomic sequencing, like amplicon sequencing, is also impacted by the choice of library preparation methods. Using four library preparation methods that included Illumina Nextera XT and Illumina TruSeq DNA PCR-free kits, and the KAPA

Biosystems Hyper Prep PCR and PCR-free systems for Illumina HiSeq 2500 (V4), and a mock microbial community, differences were shown in taxonomic abundance, error rate, and functional predictions for all library preparation methods (Jones et al., 2015). Comparison between PCR-based and PCR-free-based methods indicates that PCR-free-based methods reduce PCR bias in abundance estimates and improve assemblies for accurate taxonomic assignment (Jones et al., 2015).

4.1.6 Bioinformatic analyses

Continuing advances in sequencing technologies and reduction in costs have led to the exponential use of metagenomic approaches and consequently, to the development of many bioinformatic tools for the analysis of datasets generated. The scientific community continuously evaluates and validates these tools and have identified standard bioinformatic tools which can accurately analyze sequence data generated from both marker gene amplification and shotgun metagenomic sequencing. It is, therefore, recommended to use well-validated bioinformatic pipelines for an unbiased and accurate analysis of sequence data.

4.1.6.1 Amplicon sequencing

Analysis of amplicon sequence data involves several steps- quality filtering, clustering of sequences into operational taxonomic units (OTUs), taxonomic assignment, and diversity analysis. There are several open-source bioinformatic platforms available that can accomplish each of these steps without introducing potential biases. These platforms include QIIME (Caporaso et al., 2010), newly developed QIIME 2 (Bolyen et al., 2018), Mothur (Schloss et al., 2009), MG-RAST (Meyer et al., 2008), SILVAngs

(Quast et al., 2013). Among these platforms, QIIME, QIIME 2 and Mothur incorporate many algorithms and thereby, offer the users a variety of choices. Both MG-RAST and SILVAngs are web-based applications with graphical user interfaces (GUI) and offer fully automated pipelines for the analysis of datasets.

The first step in the analysis of a metagenomic dataset is quality filtering. Quality filtering is performed to remove sequences of unexpected length, sequences containing ambiguous bases and low-quality base-pairs, error in barcode sequences, sequences with primer mismatches and chimeric sequences. Tools that are most commonly used for quality filtering includes CUTADAPT (removes adapter sequence) (Martin, 2011), FLASH (merge paired-end reads) (Magoc and Salzberg, 2011), Trimmomatic (removes adapter and trims sequences based on quality score) (Bolger et al., 2014) and FASTX toolkit (trim sequences based on quality score) (http://hannonlab.cshl.edu/fastx_toolkit/). Presence of chimeric sequences can lead to the erroneous detection of microorganisms and therefore, it is essential to remove such sequences (especially in human microbiome studies). Chimeric sequences can be removed using UCHIME (Edgar et al., 2011) and Chimera Slayer (Haas et al., 2011).

The next step to quality filtering is the clustering of sequences into OTUs (termed, OTU picking). Clustering of sequences is achieved using *a priori* defined thresholds (usually with a 97% similarity thresholds) that consolidates similar sequences into a single OTU. There are several clustering algorithms available for OTU clustering based on sequence length or pairwise alignment. Commonly used algorithms include CD-HIT

(Fu et al., 2012), UCLUST (Edgar, 2010), UPARSE (Edgar, 2013), DNACLUST (Ghodsi et al., 2011), and BLAST. By default, QIIME and MG-RAST use the UCLUST algorithm whereas, Mothur uses its own algorithm based on the furthest neighbor, nearest neighbor and the UPGMA (unweighted-pair group method using average linkages) algorithms (Schloss *et al.*, 2009). Currently, the classifier algorithm used by Mothur is also supported by QIIME. Taxonomic assignment of OTUs is a key analysis step of metagenomic datasets. Taxonomy is usually achieved by comparing the sequence to a suitable reference using a machine learning approach such as RDP classifier which uses a naïve Bayesian based approach (Wang et al., 2007). Taxonomic assignment requires the use of reference databases of marker genes and the most commonly utilized databases include Greengenes (DeSantis et al., 2006), RDP (Cole et al., 2014) and SILVA (Quast et al., 2013). Among these three databases, Greengenes has not been updated since 2013 and, therefore, lacks the 16S rRNA gene sequences from recently discovered bacterial species, whereas, RDP and SILVA databases are more up to date. There are two primary strategies of OTU picking: reference-based and *de novo*. Based on the overlap of amplicons, reference-based OTU picking can be divided into two types- closed-reference (used when amplicons do not overlap) and open-reference (amplicons overlap). Both these reference-based OTU picking methods require a reference sequence collection and matches the query sequences to the reference database. Sequences that do not match to the reference are discarded in closed-reference OTU picking method whereas such sequences are clustered de novo in open-reference OTU picking method. *De novo* OTU picking is used

when amplicons overlap but there is no reference database available. In this case, sequences are clustered without using a reference.

Diversity analysis is another key step in metagenomic approaches. However, users should consider the differences in 16S rRNA gene copy number (1-15 per bacterial genome) among bacterial species prior to the diversity analysis (Vetrovsky and Baldrian, 2013;Stoddard et al., 2015). Correcting for 16S rRNA gene copy number is a prerequisite for an accurate diversity estimate analysis. Tools that can be used for correcting 16S rRNA gene copy numbers include rrnDB database (Stoddard *et al.*, 2015), Copyrighter (Angly et al., 2014), "pplacer" and "picante" packages within R (Kembel et al., 2012). Following correction for 16S rRNA gene copy numbers, the overall diversity of microbial communities is assessed by alpha and beta diversity. Alpha diversity estimates the diversity within individual samples and is commonly assessed as Chao 1, ACE (abundance-based coverage estimates) or Shannon index estimates. Beta diversity estimates species diversity between samples and generates a distance matrix. Bray-Curtis dissimilarity measure (estimate phylogenetic distances between groups) and weighted UniFrac (a distance metric) are the most commonly used methods for estimating beta diversity. Following beta diversity estimation, the significance of beta diversity between groups is assessed using PERMANOVA (Anderson, 2001) and ANOSIM (Clarke, 1993). Diversity analysis can be performed using QIIME, QIIME 2, Mothur, STAMP (Parks et al., 2014) and the "vegan" package within R (Dixon, 2003).

4.1.6.2 Shotgun metagenomic sequencing

Similar to 16S rRNA sequence analysis, the first step in analyzing shotgun metagenomic dataset is quality trimming. Quality filtering can be performed using open-source bioinformatic tools such as Cutadapt (Martin, 2011), Trimmomatic (Bolger et al., 2014), BBtools (JGI), and FastQC (Babraham-Bioinformatics). Following quality filtering, taxonomy is assigned and there are two approaches for taxonomy assignment-read-based profiling where sequences are matched against a reference database or assembly of sequence reads into longer sequences (contigs) which can be matched against reference databases. For read-based profiling, taxonomic assignment is achieved by aligning sequences against reference databases (such as GenBank or RefSeq) using BLAST (Basic Local Alignment Search Tool) or BLAST-based programs such as MG-RAST, MetaPhyler (Liu et al., 2011) or CARMA (Gerlach et al., 2009). The output from BLAST or BLAST-based programs can be used in MEGAN (Metagenome Analyzer) for taxonomic and functional profiling. The GUI-based MG-RAST provides pipelines for both taxonomic and functional analysis. For functional analysis, sequences are clustered using UCLUST and compared against reference databases which include GenBank (Benson et al., 2014), SEED (Overbeek et al., 2014), IMG (Markowitz et al., 2012), KEGG (Kanehisa et al., 2012) and eggNOG (Powell et al., 2014). Both MG-RAST and MEGAN can be used to visualize the profiled data. A similar platform to MG-RAST is the CyVerse cyberinfrastructure (formerly iPlant Collaborative) which provide users access to a range of analysis toolkits (Goff et al., 2011). The CyVerse cyberinfrastructure replaces the now retired Community Cyberinfrastructure for Advanced Marine Microbial

Research and Analysis (CAMERA) which also integrated bioinformatic tools for the analysis of metagenomic dataset (Seshadri et al., 2007). However, BLAST-based methods are computationally intensive and time-consuming. To address this, k-mer based approaches which replace alignment, can be used. Most common tools that use k-mers matches to assign taxonomy include Kraken (Wood and Salzberg, 2014) and CLARK (CLAssifier based on Reduced *K*-mers) (Ounit et al., 2015b). Sequence reads generated from shotgun metagenomic sequencing can also be assigned taxonomy using programs that identify universal, single-copy marker elements from genomic datasets. MetaPhlAn 2 (Truong et al., 2015) and TIPP (Nguyen et al., 2014) are two such programs that use marker gene method for taxonomic assignment. For the assembly-based approach, several tools can be used to assemble sequences into contigs. These include MetaVelvet (Namiki et al., 2012), SOAPdenovo (Li et al., 2010), Meta IBDA (Peng et al., 2011), MetaSPAdes (Nurk et al., 2017) and MEGAHIT (Li et al., 2015). Assembled contigs can then be used to match against reference database and assign taxonomy using MG-RAST, MEGAN, CyVerse cyberinfrastructure or CARMA.

Analysis of viral metagenomic datasets is not as straightforward as analyzing bacterial metagenomic datasets. As viruses lack universal marker genes, identification of viruses from sequence reads are limited to the alignment of sequences against reference databases. Similar to the analysis of bacterial datasets, taxonomic assignment of sequences generated from viral metagenomic datasets can be performed using both read-based profiling and contig assembly methods. Tools that are used to analyze bacterial shotgun metagenomic datasets such as MEGAN and MG-RAST can also be used to

analyze viral metagenomic datasets. Specialized bioinformatic pipelines for analyzing viral metagenomic datasets such as the Viral Informatics Resource for Metagenome Exploration (VIROME) (Wommack et al., 2012), the Viral MetaGenome Annotation Pipeline (VMGAP) (Lorenzi et al., 2011), VirusSeeker (Zhao et al., 2017), and the newly developed iVirus community resources (Bolduc et al., 2017) which provides access to a range of viromes analysis tools can also be used. In addition to the reference database-based analysis, viral metagenomic datasets can also be analyzed in a reference-independent manner. Tools that are commonly used for such purposes include PHACCS (Phage Communities from Contig Spectrum) (Wommack et al., 2015), MaxiPhi (Angly et al., 2006) and CrAss (Cross-Assembly) (Dutilh et al., 2012). However, a complete analysis of viral datasets requires additional steps. Prophages, viruses that integrate their genome into bacterial DNA, are often identified as bacterial species due to their similarity to bacterial genomes. Use of specialized tools such as Prophage Finder (Bose and Barber, 2006), PHAST (its successor PHASTER and PHASTEST) (Arndt et al., 2017) can identify prophages from metagenomic datasets. VirSorter (Roux et al., 2015), a tool accessible through the iVirus community, MetaPhinder (Jurtz et al., 2016) and the newly developed VHost-Classifier (Kitson et al., 2019) can be used to investigate the virus-host relationships from metagenomic datasets.

5. Conclusion

In this review, we discuss the diversity of bacteria and viruses in freshwater environments with a focus on dominant groups. In addition, we reviewed current literature to identify the best practices for employing metagenomic approaches, from

study design to sample processing and downstream bioinformatic analyses of metagenomic datasets. The diversity profile of bacteria and viruses in most freshwater environments are similar with a few exceptions (most notably polar freshwaters). However, due to the lack of computational resources and the lack of reliability of metagenomic tools at a higher taxonomic resolution (such as the species level), exploration of diversity is often limited to a lower taxonomic resolution (such as the phylum level) in most studies. As the field is continuously evolving and moving towards a level of precision in providing a true representation of diversity, in-depth and reliable exploration of microbial and viral community diversity through metagenomic approaches do not appear to be farfetched.

While high-throughput sequencing has enabled the investigation of complex microbial and viral communities in discrete environments and expanded our knowledge of bacterial and viral community diversity through the identification of previously unidentified microorganisms, challenges remain many. One such challenge is the lack of comparability between metagenomic datasets that are generated using different methodologies. Continuous advancement of sequencing technologies and reduction in costs have led to an overwhelming number of methodologies which may pose a problem in identifying methodologies that can provide an accurate and unbiased representation of bacterial and viral community diversity. The scientific community continuously strives to identify the best methodologies that can avoid introducing biases and provide an opportunity for comparing datasets generated from similar or related environments. It is, therefore, essential to use well-validated methodologies and be consistent with the

application of such methodologies which will ensure the comparability of metagenomic datasets generated from identical niches.

References

Ackermann, H.W., (2007) 5500 Phages examined in the electron microscope. *Archives of Virology* **152**: 227-243.

Aherfi, S., P. Colson, B. La Scola & D. Raoult, (2016) Giant Viruses of Amoebas: An Update. *Frontiers in Microbiology* **7**.

Ahn, J.H., B.Y. Kim, J. Song & H.Y. Weon, (2012) Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *Journal of Microbiology* **50**: 1071-1074.

Almeida, R.M., F. Roland, S.J. Cardoso, V.F. Farjalla, R.L. Bozelli & N.O. Barros, (2015) Viruses and bacteria in floodplain lakes along a major Amazon tributary respond to distance to the Amazon River. *Frontiers in Microbiology* **6**: 158.

Anderson, M.J., (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**: 32-46.

Andersson, A., J. Ahlinder, P. Mathisen, M. Hagglund, S. Backman, E. Nilsson, *et al.*, (2018) Predators and nutrient availability favor protozoa-resisting bacteria in aquatic systems. *Scientific Reports* **8**.

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

Angly, F.E., B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, *et al.*, (2006) The marine viromes of four oceanic regions. *PLoS Biology* **4**: e368.

Apprill, A., S. McNally, R. Parsons & L. Weber, (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* **75**: 129-137.

Arkhipova, K., T. Skvortsov, J.P. Quinn, J.W. McGrath, C.C. Allen, B.E. Dutilh, *et al.*, (2018) Temporal dynamics of uncultured viruses: a new dimension in viral diversity. *ISME J* **12**: 199-211.

Arnds, J., K. Knittel, U. Buck, M. Winkel & R. Amann, (2010) Development of a 16S rRNA-targeted probe set for Verrucomicrobia and its application for fluorescence in situ hybridization in a humic lake. *Systematic and Applied Microbiology* **33**: 139-148.

Arndt, D., A. Marcu, Y. Liang & D.S. Wishart, (2017) PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Brief Bioinform*.

Arslan, D., M. Legendre, V. Seltzer, C. Abergel & J.M. Claverie, (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 17486-17491.

Babraham-Bioinformatics, FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. In., pp.

Balmonte, J.P., C. Arnosti, S. Underwood, B.A. Mckee & A. Teske, (2016) Riverine Bacterial Communities Reveal Environmental Disturbance Signatures within the Betaproteobacteria and Verrucomicrobia. *Frontiers in Microbiology* **7**.

Battin, T.J., S. Luyssaert, L.A. Kaplan, A.K. Aufdenkampe, A. Richter & L.J. Tranvik, (2009) The boundless carbon cycle. *Nature Geoscience* **2**: 598-600.

Benson, D.A., K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell & E.W. Sayers, (2014) GenBank. *Nucleic Acids Research* **42**: D32-37.

Berdjeb, L., T. Pollet, I. Domaizon & S. Jacquet, (2011) Effect of grazers and viruses on bacterial community structure and production in two contrasting trophic lakes. *BMC Microbiology* **11**.

Bergmann, G.T., S.T. Bates, K.G. Eilers, C.L. Lauber, J.G. Caporaso, W.A. Walters, *et al.*, (2011) The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biology & Biochemistry* **43**: 1450-1455.

Bolduc, B., K. Youens-Clark, S. Roux, B.L. Hurwitz & M.B. Sullivan, (2017) iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *ISME J* **11**: 7-14.

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

Bolyen, E., J.R. Rideout, M.R. Dillon, N.A. Bokulich, C. Abnet, G.A. Al-Ghalith, *et al.*, (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science.

Bose, M. & R.D. Barber, (2006) Prophage Finder: a prophage loci prediction tool for prokaryotic genome sequences. *In Silico Biol* **6**: 223-227.

Braun, B., J. Schröder, H. Knecht & U. Szewzyk, (2016) Unraveling the microbial community of a cold groundwater catchment system. *Water Research* **107**: 113-126.

Breitbart, M., P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, *et al.*, (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

Briee, C., D. Moreira & P. Lopez-Garcia, (2007) Archaeal and bacterial community composition of sediment and plankton from a suboxic freshwater pond. *Research in Microbiology* **158**: 213-227.

Brouwer, M.S.M., A.P. Roberts, H. Hussain, R.J. Williams, E. Allan & P. Mullany, (2013) Horizontal gene transfer converts non-toxigenic Clostridium difficile strains into toxin producers. *Nat Commun* **4**.

Brown, C.T., L.A. Hug, B.C. Thomas, I. Sharon, C.J. Castelle, A. Singh, *et al.*, (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**: 208-U173.

Brum, J.R., J.C. Ignacio-Espinoza, S. Roux, G. Doulcier, S.G. Acinas, A. Alberti, *et al.*, (2015) Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**: 1261498.

Brussow, H., C. Canchaya & W.D. Hardt, (2004) Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* **68**: 560-+.

Cabello-Yeves, P.J., R. Ghai, M. Mehrshad, A. Picazo, A. Camacho & F. Rodriguez-Valera, (2017) Reconstruction of Diverse Verrucomicrobial Genomes from Metagenome Datasets of Freshwater Reservoirs. *Frontiers in Microbiology* **8**.

Cabral, A.S., M.M. Lessa, P.C. Junger, F.L. Thompson & R. Paranhos, (2017) Virioplankton dynamics are related to eutrophication levels in a tropical urbanized bay. *PLoS One* **12**: e0174653.

Cai, L.L., Y.L. Yang, N.Z. Jiao & R. Zhang, (2015) Evaluation of Tangential Flow Filtration for the Concentration and Separation of Bacteria and Viruses in Contrasting Marine Environments. *PLoS One* **10**.

Cai, L.L., R. Zhang, Y. He, X.Y. Feng & N.Z. Jiao, (2016) Metagenomic Analysis of Virioplankton of the Subtropical Jiulong River Estuary, China. *Viruses-Basel* **8**.

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

Caporaso, J.G., C.L. Lauber, W.A. Walters, D. Berg-Lyons, C.A. Lozupone, P.J. Turnbaugh, *et al.*, (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 4516-4522.

Chen, N., X.J. Yu, J.S. Yang, E.T. Wang, B.Z. Li & H.L. Yuan, (2015) Novosphingobium tardum sp nov., isolated from sediment of a freshwater lake. *Antonie Van Leeuwenhoek International Journal of General and Molecular Microbiology* **108**: 51-57.

Cheung, M.Y., S. Liang & J. Lee, (2013) Toxin-producing cyanobacteria in freshwater: A review of the problems, impact on drinking water safety, and efforts for protecting public health. In: Journal of Microbiology. pp. 1-10.

Chiang, E., M.L. Schmidt, M.A. Berry, B.A. Biddanda, A. Burtner, T.H. Johengen, *et al.*, (2018) Verrucomicrobia are prevalent in north-temperate freshwater lakes and display class-level preferences between lake habitats. *PLoS One* **13**.

Chibani-Chennoufi, S., A. Bruttin, M.L. Dillmann & H. Brussow, (2004) Phage-host interaction: an ecological perspective. *Journal of Bacteriology* **186**: 3677-3686.

Chopyk, J., S. Allard, D.J. Nasko, A. Bui, E.F. Mongodin & A.R. Sapkota, (2018) Agricultural Freshwater Pond Supports Diverse and Dynamic Bacterial and Viral Populations. *Frontiers in Microbiology* **9**.

Chu, B.T.T., M.L. Petrovich, A. Chaudhary, D. Wright, B. Murphy, G. Wells, *et al.*, (2018) Metagenomics Reveals the Impact of Wastewater Treatment Plants on the Dispersal of Microorganisms and Genes in Aquatic Sediments. *Applied and Environmental Microbiology* **84**.

Clarke, K.R., (1993) Non-parametric multivariate analyses of changes in community structure. *Austral Ecology* **18**: 117-143.

Clokie, M.R., A.D. Millard, A.V. Letarov & S. Heaphy, (2011) Phages in nature. *Bacteriophage* **1**: 31-45.

Clooney, A.G., F. Fouhy, R.D. Sleator, A.O. Driscoll, C. Stanton, P.D. Cotter, *et al.*, (2016) Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLoS One* **11**.

Cloutier, D.D., E.W. Alm & S.L. McLellan, (2015) Influence of Land Use, Nutrients, and Geography on Microbial Communities and Fecal Indicator Abundance at Lake Michigan Beaches. *Applied and Environmental Microbiology* **81**: 4904-4913.

Cole, J.R., Q. Wang, J.A. Fish, B.L. Chai, D.M. McGarrell, Y.N. Sun, *et al.*, (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Research* **42**: D633-D642.

Coutinho, F.H., C.B. Silveira, G.B. Gregoracci, C.C. Thompson, R.A. Edwards, C.P.D. Brussaard, *et al.*, (2017) Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **8**: 15955.

Croue, J., N.J. West, M.L. Escande, L. Intertaglia, P. Lebaron & M.T. Suzuki, (2013) A single betaproteobacterium dominates the microbial community of the crambescidine-containing sponge Crambe crambe. *Scientific Reports* **3**.

D'Amore, R., U.Z. Ijaz, M. Schirmer, J.G. Kenny, R. Gregory, A.C. Darby, *et al.*, (2016) A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics* **17**.

Dai, Y., Y. Yang, Z. Wu, Q. Feng, S. Xie & Y. Liu, (2016) Spatiotemporal variation of planktonic and sediment bacterial assemblages in two plateau freshwater lakes at different trophic status. *Applied Microbiology and Biotechnology* **100**: 4161-4175.

Davis, T.W., D.L. Berry, G.L. Boyer & C.J. Gobler, (2009) The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of Microcystis during cyanobacteria blooms. *Harmful Algae* **8**: 715-725.

Davis, T.W., G.S. Bullerjahn, T. Tuttle, R.M. McKay & S.B. Watson, (2015) Effects of Increasing Nitrogen and Phosphorus Concentrations on Phytoplankton Community Growth and Toxicity during Planktothrix Blooms in Sandusky Bay, Lake Erie. *Environmental Science and Technology* **49**: 7197-7207.

de Carcer, D.A., A. Lopez-Bueno, D.A. Pearce & A. Alcami, (2015) Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances* **1**.

Dean, F.B., S. Hosono, L.H. Fang, X.H. Wu, A.F. Faruqi, P. Bray-Ward, *et al.*, (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 5261-5266.

DeSantis, T.Z., P. Hugenholtz, N. Larsen, M. Rojas, E.L. Brodie, K. Keller, *et al.*, (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**: 5069-5072.

Dini-Andreote, F., J.C. Stegen, J.D. van Elsas & J.F. Salles, (2015) Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proceedings of the National Academy of Sciences of the United States of America* **112**: E1326-E1332.

Dixon, P., (2003) VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* **14**: 927-930.

Djikeng, A., R. Kuzmickas, N.G. Anderson & D.J. Spiro, (2009) Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**: e7264.

Duhaime, M.B., L. Deng, B.T. Poulos & M.B. Sullivan, (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a

rigorous assessment and optimization of the linker amplification method. *Environmental Microbiology* **14**: 2526-2537.

Dutilh, B.E., R. Schmieder, J. Nulton, B. Felts, P. Salamon, R.A. Edwards, *et al.*, (2012) Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **28**: 3225-3231.

Edgar, R.C., (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.

Edgar, R.C., (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**: 996-+.

Edgar, R.C., B.J. Haas, J.C. Clemente, C. Quince & R. Knight, (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194-2200.

Eiler, A., K. Zaremba-Niedzwiedzka, M. Martinez-Garcia, K.D. McMahon, R. Stepanauskas, S.G. Andersson, *et al.*, (2014) Productivity and salinity structuring of the microplankton revealed by comparative freshwater metagenomics. *Environmental Microbiology* **16**: 2682-2698.

Engberg, J., S.L. On, C.S. Harrington & P. Gerner-Smidt, (2000) Prevalence of Campylobacter, Arcobacter, Helicobacter, and Sutterella spp. in human fecal samples as estimated by a reevaluation of isolation methods for Campylobacters. *Journal of Clinical Microbiology* **38**: 286-291.

Evans, C. & C.P. Brussaard, (2012) Regional variation in lytic and lysogenic viral infection in the Southern Ocean and its contribution to biogeochemical cycling. *Applied and Environmental Microbiology* **78**: 6741-6748.

Fancello, L., S. Trape, C. Robert, M. Boyer, N. Popgeorgiev, D. Raoult, *et al.*, (2013) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *ISME J* **7**: 359-369.

Farag, I.F., J.P. Davis, N.H. Youssef & M.S. Elshahed, (2014) Global Patterns of Abundance, Diversity and Community Structure of the Aminicenantes (Candidate Phylum OP8). *PLoS One* **9**.

Farag, I.F., N.H. Youssef & M.S. Elshahed, (2017) Global Distribution Patterns and Pangenomic Diversity of the Candidate Phylum "Latescibacteria" (WS3). *Applied and Environmental Microbiology* **83**.

Farnell-Jackson, E.A. & A.K. Ward, (2003) Seasonal patterns of viruses, bacteria and dissolved organic carbon in a riverine wetland. *Freshwater Biology* **48**: 841-851.

Faust, K., J.F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, *et al.*, (2012) Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology* **8**.

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

Fouhy, F., A.G. Clooney, C. Stanton, M.J. Claesson & P.D. Cotter, (2016) 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiology* **16**.

Frias-Lopez, J., A.L. Zerkle, G.T. Bonheyo & B.W. Fouke, (2002) Partitioning of bacterial communities between seawater and healthy, black band diseased, and dead coral surfaces. *Applied and Environmental Microbiology* **68**: 2214-2228.

Frye, J.G., S. Porwollik, F. Blackmer, P. Cheng & M. McClelland, (2005) Host gene expression changes and DNA amplification during temperate phage induction. *Journal of Bacteriology* **187**: 1485-1492.

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

Fuhrman, J.A., (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541-548.

Fuks, G., M. Elgart, A. Amir, A. Zeisel, P.J. Turnbaugh, Y. Soen, *et al.*, (2018) Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome* **6**.

Gao, E.B., J.F. Gui & Q.Y. Zhang, (2012) A Novel Cyanophage with a Cyanobacterial Nonbleaching Protein A Gene in the Genome. *Journal of Virology* **86**: 236-245.

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

Ghai, R., K.D. McMahon & F. Rodriguez-Valera, (2012) Breaking a paradigm: cosmopolitan and abundant freshwater actinobacteria are low GC. *Environ Microbiol Rep* **4**: 29-35.

Ghai, R., C.M. Mizuno, A. Picazo, A. Camacho & F. Rodriguez-Valera, (2014) Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molecular Ecology* **23**: 6073-6090.

Ghai, R., F. Rodriguez-Valera, K.D. McMahon, D. Toyama, R. Rinke, T. Cristina Souza de Oliveira, *et al.*, (2011) Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS One* **6**: e23785.

Ghodsi, M., B. Liu & M. Pop, (2011) DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* **12**.

Giovannoni, S.J., T.B. Britschgi, C.L. Moyer & K.G. Field, (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**: 60-63.

Glockner, F.O., E. Zaichikov, N. Belkova, L. Denissova, J. Pernthaler, A. Pernthaler, *et al.*, (2000) Comparative 16S rRNA analysis of lake bacterioplankton reveals globally distributed phylogenetic clusters including an abundant group of actinobacteria. *Applied and Environmental Microbiology* **66**: 5053-5065.

Goff, S.A., M. Vaughn, S. McKay, E. Lyons, A.E. Stapleton, D. Gessler, *et al.*, (2011) The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front Plant Sci* **2**: 34.

Gohl, D.M., P. Vangay, J. Garbe, A. MacLean, A. Hauge, A. Becker, *et al.*, (2016) Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology* **34**: 942-+.

Gong, C.W., W.J. Zhang, X.W. Zhou, H.M. Wang, G.W. Sun, J.Z. Xiao, *et al.*, (2016) Novel Virophages Discovered in a Freshwater Lake in China. *Frontiers in Microbiology* **7**.

Gong, Z., Y.T. Liang, M. Wang, Y. Jiang, Q.W. Yang, J. Xia, *et al.*, (2018) Viral Diversity and Its Relationship With Environmental Factors at the Surface and Deep Sea of Prydz Bay, Antarctica. *Frontiers in Microbiology* **9**.

Goodfellow, M. & S.T. Williams, (1983) Ecology of actinomycetes. *Annual Review of Microbiology* **37**: 189-216.

Goodwin, S., J.D. McPherson & W.R. McCombie, (2016) Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* **17**: 333-351.

Guemes, A.G.C., M. Youle, V.A. Cantu, B. Felts, J. Nulton & F. Rohwer, (2016) Viruses as Winners in the Game of Life. *Annual Review of Virology, Vol 3* **3**: 197-214.

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

Hanson, A.M., J.A. Berges & E.B. Young, (2017) Virus morphological diversity and relationship to bacteria and chlorophyll across a freshwater trophic gradient in the Lake Michigan watershed. *Hydrobiologia* **794**: 93-108.

Hargreaves, K.R., A.M. Kropinski & M.R. Clokie, (2014) Bacteriophage behavioral ecology: How phages alter their bacterial host's habits. *Bacteriophage* **4**: e29866.

Harwood, V.J., C. Staley, B.D. Badgley, K. Borges & A. Korajkic, (2014) Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. *FEMS Microbiology Reviews* **38**: 1-40.

He, S.M., S.L.R. Stevens, L.K. Chan, S. Bertilsson, T.G. del Rio, S.G. Tringe, *et al.*, (2017) Ecophysiology of Freshwater Verrucomicrobia Inferred from Metagenome-Assembled Genomes. *Msphere* **2**.

Henson, M.W., V.C. Lanclos, B.C. Faircloth & J.C. Thrash, (2018) Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *Isme Journal* **12**: 1846-1860.

Hiorns, W.D., B.A. Methe, S.A. NierzwickiBauer & J.P. Zehr, (1997) Bacterial diversity in Adirondack Mountain lakes as revealed by 16S rRNA gene sequences. *Applied and Environmental Microbiology* **63**: 2957-2960.

Hoeijmakers, W.A.M., R. Bartfai, K.J. Francoijs & H.G. Stunnenberg, (2011) Linear amplification for deep sequencing. *Nature Protocols* **6**: 1026-1036.

Hotto, A.M., M.F. Satchwell & G.L. Boyer, (2007) Molecular characterization of potential microcystin-producing cyanobacteria in Lake Ontario embayments and nearshore waters. *Applied and Environmental Microbiology* **73**: 4570-4578.

Hu, P., L. Tom, A. Singh, B.C. Thomas, B.J. Baker, Y.M. Piceno, *et al.*, (2016) Genome-Resolved Metagenomic Analysis Reveals Roles for Candidate Phyla and Other Microbial Community Members in Biogeochemical Transformations in Oil Reservoirs. *Mbio* **7**.

Hunt, D.E., V. Klepac-Ceraj, S.G. Acinas, C. Gautier, S. Bertilsson & M.F. Polz, (2006) Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Applied and Environmental Microbiology* **72**: 2221-2225.

Hutchison, M.L., L.D. Walters, S.M. Avery, F. Munro & A. Moore, (2005) Analyses of livestock production, waste storage, and pathogen levels and prevalences in farm manures. *Applied and Environmental Microbiology* **71**: 1231-1236.

JGI, BBDuk guide. http://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/. In.: Department of Energy, pp.

Ji, B., H. Qin, S.D. Guo, W. Chen, X.C. Zhang & J.C. Liang, (2018) Bacterial communities of four adjacent fresh lakes at different trophic status. *Ecotoxicology and Environmental Safety* **157**: 388-394.

John, S.G., C.B. Mendez, L. Deng, B. Poulos, A.K.M. Kauffman, S. Kern, *et al.*, (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation (vol 3, pg 195, 2011). *Environ Microbiol Rep* **3**: 809-809.

Johnson, L.A., B. Chaban, J.C.S. Harding & J.E. Hill, (2015) Optimizing a PCR protocol for cpn60-based microbiome profiling of samples variously contaminated with host genomic DNA. *BMC Research Notes* **8**.

Jones, M.B., S.K. Highlander, E.L. Anderson, W.Z. Li, M. Dayrit, N. Klitgord, *et al.*, (2015) Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 14024-14029.

Jones, S.E. & J.T. Lennon, (2010) Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 5881-5886.

Jurtz, V.I., J. Villarroel, O. Lund, M. Voldby Larsen & M. Nielsen, (2016) MetaPhinder-Identifying Bacteriophage Sequences in Metagenomic Data Sets. *PLoS One* **11**: e0163111.

Kanagawa, T., (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering* **96**: 317-323.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi & M. Tanabe, (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40**: D109-114.

Karlsson, O.E., S. Belak & F. Granberg, (2013) The Effect of Preprocessing by Sequence-Independent, Single-Primer Amplification (Sispa) on Metagenomic Detection of Viruses. *Biosecurity and Bioterrorism-Biodefense Strategy Practice and Science* **11**: S227-S234.

Kembel, S.W., M. Wu, J.A. Eisen & J.L. Green, (2012) Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Computational Biology* **8**.

Keshri, J., A.S.P. Ram & T. Sime-Ngando, (2018) Distinctive Patterns in the Taxonomical Resolution of Bacterioplankton in the Sediment and Pore Waters of Contrasted Freshwater Lakes. *Microbial Ecology* **75**: 662-673.

Kitson, E., C.A. Suttle & J. Wren, (2019) VHost-Classifier: virus-host classification using natural language processing. *Bioinformatics*.

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

Knowles, B., C.B. Silveira, B.A. Bailey, K. Barott, V.A. Cantu, A.G. Cobian-Guemes, *et al.*, (2016) Lytic to temperate switching of viral communities. *Nature* **531**: 466-470.

Kolmakova, O.V., M.I. Gladyshev, A.S. Rozanov, S.E. Peltek & M.Y. Trusova, (2014) Spatial biodiversity of bacteria along the largest Arctic river determined by next-generation sequencing. *FEMS Microbiology Ecology* **89**: 442-450.

Korajkic, A., L.W. Parfrey, B.R. McMinn, Y.V. Baeza, W. VanTeuren, R. Knight, *et al.*, (2015) Changes in bacterial and eukaryotic communities during sewage decomposition in Mississippi river water. *Water Research* **69**: 30-39.

Krupovic, M., V. Cvirkaite-Krupdvic, J. Iranzo, D. Prangishvili & E.V. Koonin, (2018) Viruses of archaea: Structural, functional, environmental and evolutionary genomics. *Virus Research* **244**: 181-193.

Krupovic, M., J.H. Kuhn & M.G. Fischer, (2016) A classification system for virophages and satellite viruses. *Archives of Virology* **161**: 233-247.

Kumar, M.R. & V.S. Saravanan, (2010) Candidate OP Phyla: Importance, Ecology and Cultivation Prospects. *Indian journal of microbiology* **50**: 474-477.

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

Lee, C., J.W. Marion, M. Cheung, C.S. Lee & J. Lee, (2015) Associations among human-associated fecal contamination, microcystis aeruginosa, and microcystin at lake erie beaches. *International Journal of Environmental Research and Public Health* **12**: 11466-11485.

Lee, C.S., M. Kim, C. Lee, Z. Yu & J. Lee, (2016) The microbiota of recreational freshwaters and the implications for environmental and public health. *Frontiers in Microbiology* **7**.

Lee, J.E., H.L. Buckley, R.S. Etienne & G. Lear, (2013) Both species sorting and neutral processes drive assembly of bacterial communities in aquatic microcosms. *FEMS Microbiology Ecology* **86**: 288-302.

Leibold, M.A., (1995) The Niche Concept Revisited - Mechanistic Models and Community Context. *Ecology* **76**: 1371-1382.

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

Li, R.Q., H.M. Zhu, J. Ruan, W.B. Qian, X.D. Fang, Z.B. Shi, *et al.*, (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20**: 265-272.

Lindstrom, E.S., M.P. Kamst-Van Agterveld & G. Zwart, (2005) Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. *Applied and Environmental Microbiology* **71**: 8201-8206.

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

Liu, S., H.X. Ren, L.D. Shen, L.P. Lou, G.M. Tan, P. Zheng, *et al.*, (2015) pH levels drive bacterial community structure in sediments of the Qiantang River as determined by 454 pyrosequencing. *Frontiers in Microbiology* **6**.

Liu, T., A.N. Zhang, J.W. Wang, S.F. Liu, X.T. Jiang, C.Y. Dang, *et al.*, (2018) Integrated biogeography of planktonic and sedimentary bacterial communities in the Yangtze River. *Microbiome* **6**.

Liu, Y., J.X. Zhang, L. Zhao, X.L. Zhang & S.G. Xie, (2014) Spatial distribution of bacterial communities in high-altitude freshwater wetland sediment. *Limnology* **15**: 249-256.

Lliros, M., O. Inceoglu, T. Garcia-Armisen, A. Anzil, B. Leporcq, L.M. Pigneur, *et al.*, (2014) Bacterial Community Composition in Three Freshwater Reservoirs of Different Alkalinity and Trophic Status. *PLoS One* **9**.

Logares, R., J. Brate, F. Heinrich, K. Shalchian-Tabrizi & S. Bertilsson, (2010) Infrequent Transitions between Saline and Fresh Waters in One of the Most Abundant Microbial Lineages (SAR11). *Molecular Biology and Evolution* **27**: 347-357.

Logue, J.B., S. Langenheder, A.F. Andersson, S. Bertilsson, S. Drakare, A. Lanzen, *et al.*, (2012) Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species-area relationships. *Isme Journal* **6**: 1127-1136.

Long, A.M. & S.M. Short, (2016) Seasonal determinations of algal virus decay rates reveal overwintering in a temperate freshwater pond. *ISME J* **10**: 1602-1612.

Lopez-Bueno, A., A. Rastrojo, R. Peiro, M. Arenas & A. Alcami, (2015) Ecological connectivity shapes quasispecies structure of RNA viruses in an Antarctic lake. *Molecular Ecology* **24**: 4812-4825.

Lopez-Bueno, A., J. Tamames, D. Velazquez, A. Moya, A. Quesada & A. Alcami, (2009) High Diversity of the Viral Community from an Antarctic Lake. *Science* **326**: 858-861.

Lorenzi, H.A., J. Hoover, J. Inman, T. Safford, S. Murphy, L. Kagan, *et al.*, (2011) The Viral MetaGenome Annotation Pipeline (VMGAP): An automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Standards in Genomic Sciences* **4**: 418-429.

Lozupone, C.A. & R. Knight, (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci U S A* **104**: 11436-11440.

Ma, Y.F., L.Z. Allen & B. Palenik, (2014) Diversity and genome dynamics of marine cyanophages using metagenomic analyses. *Environ Microbiol Rep* **6**: 583-594.

Madsen, E.L., (2011) Microorganisms and their roles in fundamental biogeochemical cycles. *Current Opinion in Biotechnology* **22**: 456-464.

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

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

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

Maruyama, F. & S. Ueki, (2016) Evolution and Phylogeny of Large DNA Viruses, Mimiviridae and Phycodnaviridae Including Newly Characterized Heterosigma akashiwo Virus. *Frontiers in Microbiology* **7**.

Matteson, A.R., S.N. Loar, R.A. Bourbonniere & S.W. Wilhelm, (2011) Molecular enumeration of an ecologically important cyanophage in a laurentian Great Lake. *Applied and Environmental Microbiology* **77**: 6772-6779.

McKane, A.J., D. Alonso & R.V. Sole, (2004) Analytic solution of Hubbell's model of local community dynamics. *Theoretical Population Biology* **65**: 67-73.

McLellan, S.L., S.M. Huse, S.R. Mueller-Spitz, E.N. Andreishcheva & M.L. Sogin, (2010) Diversity and population structure of sewage-derived microorganisms in wastewater treatment plant influent. *Environmental Microbiology* **12**: 378-392.

McLellan, S.L., R.J. Newton, J.L. Vandewalle, O.C. Shanks, S.M. Huse, A.M. Eren, *et al.*, (2013) Sewage reflects the distribution of human faecal Lachnospiraceae. *Environmental Microbiology* **15**: 2213-2227.

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

Miller, C.S., K.M. Handley, K.C. Wrighton, K.R. Frischkorn, B.C. Thomas & J.F. Banfield, (2013) Short-Read Assembly of Full-Length 16S Amplicons Reveals Bacterial Diversity in Subsurface Sediments. *PLoS One* **8**.

Mohiuddin, M. & H.E. Schellhorn, (2015) Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Frontiers in Microbiology* **6**.

Mohiuddin, M.M., S.R. Botts, A. Paschos & H.E. Schellhorn, (2019) Temporal and spatial changes in bacterial diversity in mixed use watersheds of the Great Lakes region. *Journal of Great Lakes Research* **45**: 109-118.

Mohiuddin, M.M., Y. Salama, H.E. Schellhorn & G.B. Golding, (2017) Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Research* **115**: 360-369.

Moon, B.Y., J.Y. Park, D.A. Robinson, J.C. Thomas, Y.H. Park, J.A. Thornton, *et al.*, (2016) Mobilization of Genomic Islands of Staphylococcus aureus by Temperate Bacteriophage. *PLoS One* **11**.

Morris, R.M., M.S. Rappe, S.A. Connon, K.L. Vergin, W.A. Siebold, C.A. Carlson, *et al.*, (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806-810.

Muir, P., S.T. Li, S.K. Lou, D.F. Wang, D.J. Spakowicz, L. Salichos, *et al.*, (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology* **17**.

Müller, S. & S.M. Mitrovic, (2015) Phytoplankton co-limitation by nitrogen and phosphorus in a shallow reservoir: progressing from the phosphorus limitation paradigm. *Hydrobiologia* **744**: 255-269.

Muyzer, G., E.C. Dewaal & A.G. Uitterlinden, (1993) Profiling of Complex Microbial-Populations by Denaturing Gradient Gel-Electrophoresis Analysis of Polymerase

Chain Reaction-Amplified Genes-Coding for 16s Ribosomal-Rna. *Applied and Environmental Microbiology* **59**: 695-700.

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

Nemergut, D.R., E.K. Costello, M. Hamady, C. Lozupone, L. Jiang, S.K. Schmidt, *et al.*, (2011) Global patterns in the biogeography of bacterial taxa. *Environmental Microbiology* **13**: 135-144.

Neuenschwander, S.M., R. Ghai, J. Pernthaler & M.M. Salcher, (2018) Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *ISME J* **12**: 185-198.

Newton, R.J., S.E. Jones, A. Eiler, K.D. McMahon & S. Bertilsson, (2011) A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews* **75**: 14-49.

Nguyen, N.P., S. Mirarab, B. Liu, M. Pop & T. Warnow, (2014) TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics* **30**: 3548-3555.

Noguerola, I., A. Picazo, M. Lliros, A. Camacho & C.M. Borrego, (2015) Diversity of freshwater Epsilonproteobacteria and dark inorganic carbon fixation in the sulphidic redoxcline of a meromictic karstic lake. *FEMS Microbiology Ecology* **91**.

Nurk, S., D. Meleshko, A. Korobeynikov & P.A. Pevzner, (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**: 824-834.

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

Overbeek, R., R. Olson, G.D. Pusch, G.J. Olsen, J.J. Davis, T. Disz, *et al.*, (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* **42**: D206-214.

Pace, N.R., (1997) A molecular view of microbial diversity and the biosphere. *Science* **276**: 734-740.

Paerl, H., (2014) Mitigating Harmful Cyanobacterial Blooms in a Human- and Climatically-Impacted World. *Life* **4**: 988-1012.

Paez-Espino, D., E.A. Eloe-Fadrosh, G.A. Pavlopoulos, A.D. Thomas, M. Huntemann, N. Mikhailova, *et al.*, (2016) Uncovering Earth's virome. *Nature* **536**: 425-+.

Pagaling, E., K. Vassileva, C.G. Mills, T. Bush, R.A. Blythe, J. Schwarz-Linek, *et al.*, (2017) Assembly of microbial communities in replicate nutrient-cycling model ecosystems follows divergent trajectories, leading to alternate stable states. *Environmental Microbiology* **19**: 3374-3386.

Palesse, S., J. Colombet, A.S. Pradeep Ram & T. Sime-Ngando, (2014) Linking host prokaryotic physiology to viral lifestyle dynamics in a temperate freshwater lake (Lake Pavin, France). *Microbial Ecology* **68**: 740-750.

Parada, A.E., D.M. Needham & J.A. Fuhrman, (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology* **18**: 1403-1414.

Parks, D.H., G.W. Tyson, P. Hugenholtz & R.G. Beiko, (2014) STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* **30**: 3123-3124.

Payet, J.P. & C.A. Suttle, (2013) To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnology and Oceanography* **58**: 465-474.

Pei, A.N., H.R. Li, W.E. Oberdorf, A.V. Alekseyenko, T. Parsons, L.Y. Yang, *et al.*, (2012) Diversity of 5S rRNA genes within individual prokaryotic genomes. *FEMS Microbiology Letters* **335**: 11-18.

Peng, Y., H.C.M. Leung, S.M. Yiu & F.Y.L. Chin, (2011) Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* **27**: I94-I101.

Pesant, S., F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, *et al.*, (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data* **2**.

Philippe, N., M. Legendre, G. Doutre, Y. Coute, O. Poirot, M. Lescot, *et al.*, (2013) Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341**: 281-286.

Piggot, A.M., J.S. Klaus, S. Johnson, M.C. Phillips & M.S.G. Helena, (2012) Relationship between enterococcal levels and sediment biofilms at recreational beaches in South Florida. *Applied and Environmental Microbiology* **78**: 5973-5982.

Popgeorgiev, N., G. Michel, H. Lepidi, D. Raoult & C. Desnues, (2013) Marseillevirus Adenitis in an 11-Month-Old Child. *Journal of Clinical Microbiology* **51**: 4102-4105.

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, *et al.*, (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research* **42**: D231-239.

Proctor, C.R., M.D. Besmer, T. Langenegger, K. Beck, J.C. Walser, M. Ackermann, *et al.*, (2018) Phylogenetic clustering of small low nucleic acid-content bacteria across diverse freshwater ecosystems. *Isme Journal* **12**: 1344-1359.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, *et al.*, (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**: D590-D596.

Read, D.S., H.S. Gweon, M.J. Bowes, L.K. Newbold, D. Field, M.J. Bailey, *et al.*, (2015) Catchment-scale biogeography of riverine bacterioplankton. *ISME J* **9**: 516-526.

Ren, L., E. Jeppesen, D. He, J. Wang, L. Liboriussen, P. Xing, *et al.*, (2015) pH Influences the Importance of Niche-Related and Neutral Processes in Lacustrine Bacterioplankton Assembly. *Applied and Environmental Microbiology* **81**: 3104-3114.

Riemann, L. & A. Winding, (2001) Community dynamics of free-living and particle-associated bacterial assemblages during a freshwater phytoplankton bloom. *Microbial Ecology* **42**: 274-285.

Rigosi, A., C.C. Carey, B.W. Ibelings & J.D. Brookes, (2014) The interaction between climate warming and eutrophication to promote cyanobacteria is dependent on trophic state and varies among taxa. *Limnology and Oceanography* **59**: 99-114.

Rinta-Kanto, J.M. & S.W. Wilhelm, (2006) Diversity of microcystin-producing cyanobacteria in spatially isolated regions of Lake Erie. *Applied and Environmental Microbiology* **72**: 5083-5085.

Roguet, A., G.S. Laigle, C. Therial, A. Bressy, F. Soulignac, A. Catherine, *et al.*, (2015) Neutral community model explains the bacterial community assembly in freshwater lakes. *FEMS Microbiology Ecology* **91**.

Rohwer, F. & R.V. Thurber, (2009) Viruses manipulate the marine environment. *Nature* **459**: 207-212.

Roux, S., J.R. Brum, B.E. Dutilh, S. Sunagawa, M.B. Duhaime, A. Loy, *et al.*, (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689-693.

Roux, S., F. Enault, B.L. Hurwitz & M.B. Sullivan, (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**: e985.

Roux, S., F. Enault, A. Robin, V. Ravet, S. Personnic, S. Theil, *et al.*, (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**: e33641.

Roux, S., M. Krupovic, D. Debroas, P. Forterre & F. Enault, (2013) Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biology* **3**.

Ruiz-Gonzalez, C., J.P. Nino-Garcia & P.A. del Giorgio, (2015a) Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecology Letters* **18**: 1198-1206.

Ruiz-Gonzalez, C., J.P. Nino-Garcia, J.F. Lapierre & P.A. del Giorgio, (2015b) The quality of organic matter shapes the functional biogeography of bacterioplankton across boreal freshwater ecosystems. *Global Ecology and Biogeography* **24**: 1487-1498.

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

Russell, T.L., K.M. Yamahara & A.B. Boehm, (2012) Mobilization and transport of naturally occurring enterococci in beach sands subject to transient infiltration of seawater. *Environmental Science and Technology* **46**: 5988-5996.

Saarenheimo, J., S.L. Aalto, A.J. Rissanen & M. Tiirola, (2017) Microbial Community Response on Wastewater Discharge in Boreal Lake Sediments. *Frontiers in Microbiology* **8**.

Salcher, M.M., J. Pernthaler & T. Posch, (2011) Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria 'that rule the waves' (LD12). *ISME J* **5**: 1242-1252.

Salcher, M.M., T. Posch & J. Pernthaler, (2013) In situ substrate preferences of abundant bacterioplankton populations in a prealpine freshwater lake. *Isme Journal* **7**: 896-907.

Salka, I., A. Srivastava, M. Allgaier & H.P. Grossart, (2014) The Draft Genome Sequence of Sphingomonas sp. Strain FukuSWIS1, Obtained from Acidic Lake Grosse

Fuchskuhle, Indicates Photoheterotrophy and a Potential for Humic Matter Degradation. *Genome Announcements* **2**.

Schloss, P.D., S.L. Westcott, T. Ryabin, J.R. Hall, M. Hartmann, E.B. Hollister, *et al.*, (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**: 7537-7541.

Schrader, C., A. Schielke, L. Ellerbroek & R. Johne, (2012) PCR inhibitors - occurrence, properties and removal. *Journal of Applied Microbiology* **113**: 1014-1026.

Schriefer, A.E., P.F. Cliften, M.C. Hibberd, C. Sawyer, V. Brown-Kennerly, L. Burcea, *et al.*, (2018) A multi-amplicon 16S rRNA sequencing and analysis method for improved taxonomic profiling of bacterial communities. *Journal of Microbiological Methods* **154**: 6-13.

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

Shanks, O.C., R.J. Newton, C.A. Kelty, S.M. Huse, M.L. Sogin & S.L. McLellan, (2013) Comparison of the Microbial Community Structures of Untreated Wastewaters from Different Geographic Locales. *Applied and Environmental Microbiology* **79**: 2906-2913.

Sharuddin, S.S., N. Ramli, M.A. Hassan, N.A. Mustapha, A. Amran, D. Mohd-Nor, *et al.*, (2017) Bacterial community shift revealed Chromatiaceae and Alcaligenaceae as potential bioindicators in the receiving river due to palm oil mill effluent final discharge. *Ecological Indicators* **82**: 526-529.

Shiklomanov, I.A., (1993) World fresh water resources. In: Water in crisis : a guide to the world's fresh water resources P.H. Gleick (ed). New York: Oxford University Press, pp.

Sinclair, L., O.A. Osman, S. Bertilsson & A. Eiler, (2015) Microbial Community Composition and Diversity via 16S rRNA Gene Amplicons: Evaluating the Illumina Platform. *PLoS One* **10**.

Singer, E., B. Bushnell, D. Coleman-Derr, B. Bowman, R.M. Bowers, A. Levy, *et al.*, (2016) High-resolution phylogenetic microbial community profiling. *Isme Journal* **10**: 2020-2032.

Skvortsov, T., C. de Leeuwe, J.P. Quinn, J.W. McGrath, C.C.R. Allen, Y. McElarney, *et al.*, (2016) Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. *PLoS One* **11**.

Smith, V.H., G.D. Tilman & J.C. Nekola, (1999) Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environmental Pollution* **100**: 179-196.

Spietz, R.L., C.M. Williams, G. Rocap & M.C. Horner-Devine, (2015) A Dissolved Oxygen Threshold for Shifts in Bacterial Community Structure in a Seasonally Hypoxic Estuary. *PLoS One* **10**.

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

Staley, C. & M.J. Sadowsky, (2016) Regional similarities and consistent patterns of local variation in beach sand bacterial communities throughout the Northern Hemisphere. *Applied and Environmental Microbiology* **82**: 2751-2762.

Staley, C., T. Unno, T.J. Gould, B. Jarvis, J. Phillips, J.B. Cotner, *et al.*, (2013) Application of Illumina next-generation sequencing to characterize the bacterial community of the Upper Mississippi River. *Journal of Applied Microbiology* **115**: 1147-1158.

Stein, J.L., T.L. Marsh, K.Y. Wu, H. Shizuya & E.F. DeLong, (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology* **178**: 591-599.

Stoddard, S.F., B.J. Smith, R. Hein, B.R.K. Roller & T.M. Schmidt, (2015) rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research* **43**: D593-D598.

Sullivan, M.B., D. Lindell, J.A. Lee, L.R. Thompson, J.P. Bielawski & S.W. Chisholm, (2006) Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology* **4**: 1344-1357.

Sunagawa, S., L.P. Coelho, S. Chaffron, J.R. Kultima, K. Labadie, G. Salazar, *et al.*, (2015) Structure and function of the global ocean microbiome. *Science* **348**.

Suttle, C.A., (2005) Viruses in the sea. *Nature* **437**: 356-361.

Suttle, C.A., (2007) Marine viruses - major players in the global ecosystem. *Nature Reviews Microbiology* **5**: 801-812.

Szekely, A.J., M. Berga & S. Langenheder, (2013) Mechanisms determining the fate of dispersed bacterial communities in new environments. *Isme Journal* **7**: 61-71.

Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann & E. Willerslev, (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* **21**: 2045-2050.

Tamames, J., J.J. Abellan, M. Pignatelli, A. Camacho & A. Moya, (2010) Environmental distribution of prokaryotic taxa. *BMC Microbiology* **10**: 85.

Teng, F., S.S.D. Nair, P.F. Zhu, S.S. Li, S. Huang, X.L. Li, *et al.*, (2018) Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. *Scientific Reports* **8**.

Tessler, M., J.S. Neumann, E. Afshinnekoo, M. Pineda, R. Hersch, L.F.M. Velho, *et al.*, (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports* **7**.

Thompson, L.R., J.G. Sanders, D. McDonald, A. Amir, J. Ladau, K.J. Locey, *et al.*, (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**: 457-+.

Thurber, R.V., M. Haynes, M. Breitbart, L. Wegley & F. Rohwer, (2009) Laboratory procedures to generate viral metagenomes. *Nature Protocols* **4**: 470-483.

Tran, P., A. Ramachandran, O. Khawasik, B.E. Beisner, M. Rautio, Y. Huot, *et al.*, (2018) Microbial life under ice: Metagenome diversity and in situ activity of

Verrucomicrobia in seasonally ice-covered Lakes. *Environmental Microbiology* **20**: 2568-2584.

Tranvik, L.J., J.J. Cole & Y.T. Prairie, (2018) The study of carbon in inland waters-from isolated ecosystems to players in the global carbon cycle. *Limnology and Oceanography Letters* **3**: 41-48.

Tremblay, J., K. Singh, A. Fern, E.S. Kirton, S. He, T. Woyke, *et al.*, (2015) Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology* **6**: 771.

Truong, D.T., E.A. Franzosa, T.L. Tickle, M. Scholz, G. Weingart, E. Pasolli, *et al.*, (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12**: 902-903.

Tseng, C.H., P.W. Chiang, F.K. Shiah, Y.L. Chen, J.R. Liou, T.C. Hsu, *et al.*, (2013) Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J* **7**: 2374-2386.

Uyaguari-Diaz, M.I., M. Chan, B.L. Chaban, M.A. Croxen, J.F. Finke, J.E. Hill, *et al.*, (2016) A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome* **4**.

Vannini, C., C. Sigona, M. Hahn, G. Petroni & M. Fujishima, (2017) High degree of specificity in the association between symbiotic betaproteobacteria and the host Euplotes (Ciliophora, Euplotia). *European Journal of Protistology* **59**: 124-132.

Ventura, M., C. Canchaya, A. Tauch, G. Chandra, G.F. Fitzgerald, K.F. Chater, *et al.*, (2007) Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiology and Molecular Biology Reviews* **71**: 495-548.

Vetrovsky, T. & P. Baldrian, (2013) The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS One* **8**.

Vijayavel, K., R. Fujioka, J. Ebdon & H. Taylor, (2010) Isolation and characterization of Bacteroides host strain HB-73 used to detect sewage specific phages in Hawaii. *Water Research* **44**: 3714-3724.

Wacklin, P., L. Hoffmann & J. Komarek, (2009) Nomenclatural validation of the genetically revised cyanobacterial genus Dolichospermum (Ralfs ex Bornet et Flahault) comb, nova. *Fottea* **9**: 59-64.

Wang, L., J. Zhang, H. Li, H. Yang, C. Peng, Z. Peng, *et al.*, (2018) Shift in the microbial community composition of surface water and sediment along an urban river. *Science of the Total Environment* **627**: 600-612.

Wang, M.N., X.Y. Ge, Y.Q. Wu, X.L. Yang, B. Tan, Y.J. Zhang, *et al.*, (2015) Genetic diversity and temporal dynamics of phytoplankton viruses in East Lake, China. *Virologica Sinica* **30**: 290-300.

Wang, P., B. Chen, R. Yuan, C. Li & Y. Li, (2016) Characteristics of aquatic bacterial community and the influencing factors in an urban river. *Science of the Total Environment* **569-570**: 382-389.

Wang, Q., G.M. Garrity, J.M. Tiedje & J.R. Cole, (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**: 5261-5267.

Wang, Y., H.F. Sheng, Y. He, J.Y. Wu, Y.X. Jiang, N.F. Tam, *et al.*, (2012) Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. *Applied and Environmental Microbiology* **78**: 8264-8271.

Warnecke, F., R. Amann & J. Pernthaler, (2004) Actinobacterial 16S rRNA genes from freshwater habitats cluster in four distinct lineages. *Environmental Microbiology* **6**: 242-253.

White, R.A., A.M. Chan, G.S. Gavelis, B.S. Leander, A.L. Brady, G.F. Slater, *et al.*, (2016) Metagenomic Analysis Suggests Modern Freshwater Microbialites Harbor a Distinct Core Microbial Community. *Frontiers in Microbiology* **6**.

Whitman, R.L., V.J. Harwood, T.A. Edge, M.B. Nevers, M. Byappanahalli, K. Vijayavel, *et al.*, (2014) Microbes in beach sands: Integrating environment, ecology and public health. In: Reviews in Environmental Science and Biotechnology. pp. 329-368.

Whitman, W.B., D.C. Coleman & W.J. Wiebe, (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci U S A* **95**: 6578-6583.

Wilhelm, S.W., M.J. Carberry, M.L. Eldridge, L. Poorvin, M.A. Saxton & M.A. Doblin, (2006) Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of g20 genes. *Applied and Environmental Microbiology* **72**: 4957-4963.

Wilhelm, S.W., S.E. Farnsley, G.R. LeCleir, A.C. Layton, M.F. Satchwell, J.M. DeBruyn, *et al.*, (2011) The relationships between nutrients, cyanobacterial toxins and the microbial community in Taihu (Lake Tai), China. *Harmful Algae* **10**: 207-215.

Winstel, V., C.G. Liang, P. Sanchez-Carballo, M. Steglich, M. Munar, B.M. Broker, *et al.*, (2013) Wall teichoic acid structure governs horizontal gene transfer between major bacterial pathogens. *Nat Commun* **4**.

Wommack, K.E., J. Bhavsar, S.W. Polson, J. Chen, M. Dumas, S. Srinivasiah, *et al.*, (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **6**: 421-433.

Wommack, K.E., D.J. Nasko, J. Chopyk & E.G. Sakowski, (2015) Counts and sequences, observations that continue to change our understanding of viruses in nature. *Journal of Microbiology* **53**: 181-192.

Wood, D.E. & S.L. Salzberg, (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**.

Wu, J.-Y., X.-T. Jiang, Y.-X. Jiang, S.-Y. Lu, F. Zou & H.-W. Zhou, (2010) Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method. *BMC Microbiology* **10**: 255.

Yau, S., F.M. Lauro, M.Z. DeMaere, M.V. Brown, T. Thomas, M.J. Raftery, *et al.*, (2011) Virophage control of antarctic algal host-virus dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 6163-6168.

Yolken, R.H., L. Jones-Brando, D.D. Dunigan, G. Kannan, F. Dickerson, E. Severance, *et al.*, (2014) Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is

associated with changes in cognitive functions in humans and mice. *Proc Natl Acad Sci U S A* **111**: 16106-16111.

Yoshikawa, G., R. Blanc-Mathieu, C. Song, Y. Kayama, T. Mochizuki, K. Murata, *et al.*, (2019) Medusavirus, a novel large DNA virus discovered from hot spring water. *Journal of Virology*.

Youssef, N.H., I.F. Farag, C. Rinke, S.J. Hallam, T. Woyke & M.S. Elshahed, (2015) In Silico Analysis of the Metabolic Potential and Niche Specialization of Candidate Phylum "Latescibacteria" (WS3). *PLoS One* **10**.

Zawar-Reza, P., G.R. Arguello-Astorga, S. Kraberger, L. Julian, D. Stainton, P.A. Broady, *et al.*, (2014) Diverse small circular single-stranded DNA viruses identified in a freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infection Genetics and Evolution* **26**: 132-138.

Zeder, M., S. Peter, T. Shabarova & J. Pernthaler, (2009) A small population of planktonic Flavobacteria with disproportionally high growth during the spring phytoplankton bloom in a prealpine lake. *Environmental Microbiology* **11**: 2676-2686.

Zhao, G., G. Wu, E.S. Lim, L. Droit, S. Krishnamurthy, D.H. Barouch, *et al.*, (2017) VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **503**: 21-30.

Zhou, J.L., W.J. Zhang, S.L. Yan, J.Z. Xiao, Y.Y. Zhang, B.L. Li, *et al.*, (2013) Diversity of Virophages in Metagenomic Data Sets. *Journal of Virology* **87**: 4225-4236.

Zwart, G., B.C. Crump, M.P.K.V. Agterveld, F. Hagen & S.K. Han, (2002) Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology* **28**: 141-155.

Chapter 2: Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analyses

Mohammad Mohiuddin[1], and Herb E. Schellhorn[1*]
[1]Department of Biology, McMaster University, Hamilton, ON, Canada

*** Correspondence:** Herb E. Schellhorn, Department of Biology, McMaster University, Hamilton, ON, L8S 4L8, Canada.
schell@mcmaster.ca

Reproduced with permission from Mohiuddin, M. and H. E. Schellhorn. 2015. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analyses. *Frontiers in Microbiology*. 6:960. doi: 10.3389/fmicb.2015.00960. Frontiers Media SA.

Abstract

Viruses are the most abundant microorganisms in the aquatic environment, yet the identification of viruses and assessing their diversity still remains a challenge. Here, we present a robust, routinely usable approach to identify viruses from two freshwater lakes of the lower Great Lakes region, Lake Ontario and Lake Erie. We collected water samples from six different beaches of these two lakes during the summer period of 2012 and 2013, and separated into three distinct fractions, namely a bacterial fraction, a virus like particle (VLP) fraction and a fraction of eDNA (environmental DNA). DNA extracted from all three fractions was sequenced and bioinformatic analyses of sequences revealed the presence of viruses from major viral families. The analyzed viral sequences were dominated by bacteriophage sequences, but also contained many plant and animal viruses. Within the context of this study, geographic location does not appear to have a major impact on viral abundance and diversity since virome composition of both lakes were similar. Comparative analyses between eDNA and viral fractions showed that eDNA can be used in combination with VLP fractions to identify viruses from the environment.

1. Introduction

Viruses, including phages outnumber microbial cells ten to one in most aquatic environments (Chibani-Chennoufi et al., 2004). At a concentration of 10^6 to 10^9 viruses per milliliter of sea water (Bergh et al., 1989) and up to 10^6 viral species per kg of marine sediments (Breitbart et al., 2004) and 10^9 viruses per gram of soil in terrestrial environments (Williamson et al., 2003), most likely, viruses are the largest reservoirs of underexplored microbial components of the entire biosphere. These large and diverse viral communities influence a number of processes ranging from global geochemical and ecological cycles (Fuhrman, 1999;Suttle, 2007) to bacterial virulence and pathogenesis (Brussow et al., 2004). However, compared to marine viruses, nature of freshwater viral communities and their impact on aquatic environment still remains largely unexplored (Middelboe et al., 2008). Despite being underexplored, recent studies on the freshwater environment showing that freshwater viral communities are rich, consisting of diverse and novel viruses (Djikeng et al., 2009;Hewson et al., 2012;Roux et al., 2012).

The majority of viruses identified in freshwater environments is bacteriophages along with diverse human and animal viruses (Tseng et al., 2013). To a lesser extent, plant viruses, have also been detected in freshwater environments (Djikeng et al., 2009;Fancello et al., 2013;Tseng et al., 2013). Identification of these viruses, however, is challenging, as most of them are difficult to culture through traditional techniques. Unlike, 16S rRNA sequences of prokaryotes and 18S rRNA sequences of eukaryotes, the signature sequences shared by cellular organisms, viruses lack universally conserved

sequences (Rohwer and Edwards, 2002). The lack of a universal phylogenetic marker limits direct detection of viruses in the environment. To address these limitations, viral fractions were isolated from freshwater and analyzed using shotgun sequencing and metagenomic analysis. Sequencing of whole viral communities has led to the rise of viral metagenomics (Breitbart et al., 2002) and provides an unprecedented insight into viral diversity. This culture-independent approach also overcomes many limitations of the culture based and / or PCR based methods for virus identification and offers an unrestricted access to the type and distribution of viruses in any host environment. More recently, because of the rapid development and cost reduction of sequencing techniques, metagenomic studies are being used regularly to identify viruses from the environment. However, few studies about viral content have been performed in the lower Great Lakes region, particularly in Lake Ontario and Lake Erie. Culture based and PCR based approaches used in these lakes have provided evidence for the presence of pathogenic microorganisms (Edge and Hill, 2007;Fong et al., 2007;Khan et al., 2014). Nevertheless, little is known about the type of viruses with the exception of algal viruses (Short et al., 2011) and coliphages (Fong et al., 2007). These lakes are of long-standing interest as they are frequently used for recreational activities during the summer period and also a primary source of drinking water across the lower Great Lakes region. The study of viruses through metagenomic approach in this region, may provide an important insight into virus type and distribution.

Metagenomic studies applied to the viral populations of discrete environments have expanded our knowledge of viral diversity in recent years. However, comparison between

diverse viromes is difficult because of the lack of a single methodology that allows the targeting of an entire viral population in a given environment. Diversity in viral nucleic acid composition (DNA or RNA, single-stranded or double-stranded) also limits their simultaneous detection. Traditionally, virome studies focus on either DNA or RNA viruses, where particles that pass through the 0.2 µm pore-size filters (Thurber et al., 2009) are considered as viruses. Filtration is an important step as it excludes most cellular fractions from viruses, but it also leads to the exclusion or under-representation of giant viruses with diameter larger than 200 nm (Fischer et al., 2010;Arslan et al., 2011;Philippe et al., 2013). Therefore, it is important to employ a systematic approach which allows targeting all viruses within a given sample. So far, viral metagenomic studies did not consider the fact that viruses release free nucleic acids in the environment after lysis (natural decay or lysis) and extracellular DNA can persist in the environment for up to one month (Dejean et al., 2011). eDNA has been used to identify non-viral organisms from the environment (Ficetola et al., 2008;Thomsen et al., 2012) and therefore has the potential to provide valuable information about all viruses that may present in an environment. Here, we present a simple and systematic approach for the identification of viruses from freshwater sample using VLP DNA and eDNA. To identify and assess the diversity of the lower Great Lakes virome content, we collected water samples from six different beaches of Lake Ontario and Lake Erie from June to September of the year 2012 and 2013. These beaches are of special interest because they are heavily used for recreational activities during the summer period and are subject to frequent closures due

to high *E. coli* counts. The virome content of Lake Ontario was further compared to that of Lake Erie to determine the effect of geographical location on the virome content.

2. Materials and methods

2.1 Sampling sites and sample collection

Water samples were collected in sterile 1.0 L sampling bottles (Nalgene) from 6 different locations across Lake Ontario and Lake Erie. Samples were collected from the surface at 1.0 m depth from Lakeside Beach, Queen's Royal Beach and Fifty Point Beach of Lake Ontario and from Long Beach, Long Beach Conservation Area East and Nickel Beach of Lake Erie during the summer period of 2012 and 2013 (**Figure 2S.1).**

2.2 Sample fractionation

After collection, samples were kept in ice and transferred to the lab within 4 hours and processed within 24 hours. Samples were separated into three distinct fractions: bacterial fraction, VLP (Virus like particles) fraction and eDNA fraction according to the schematic diagram (**Figure 2.1**). eDNA is defined as the DNA released in the environment after lysis (natural decay or lysis) of bacteria, viruses and higher organisms Briefly, 400 ml of water samples were centrifuged for 10 min at 10,000 x *g* at 4^{o}C. After centrifugation, each pellet was resuspended with 2.0 ml of 1x PBS buffer (pH 7.2) and stored in -20^{o}C as bacterial fraction. Of the 400 ml supernatant, 300 ml was used to concentrate the VLPs using PEG (poly-ethylene glycol) and $MgSO_4$ at a final concentration of 4% and 15mM respectively (Colombet et al., 2007;Branston et al., 2012). The supernatant mixed with PEG and $MgSO_4$ was incubated at 4^{o}C for 16 hours

and the mixture was then centrifuged for 30 min at 10,000 x *g* at 4° C to concentrate the VLPs. After concentrating the VLPs, the pellet (VLPs) was resuspended with 8.0 ml of 1x PBS (pH 7.2). The mixture was then filtered through low-protein-binding 0.22-µm-pore-size filter (Millex-GP; Millipore, Etobicoke, ON) and centrifuged at 180,000 x *g* for 1.5 h at 4° C. After centrifugation, the pellet was resuspended with 300 µl 1x PBS (pH 7.2) buffer and stored in -80° C as VLP fraction.

To isolate eDNA, 100 ml of the previously separated supernatant was filtered through 0.22 µm filter. The filtered supernatant was then mixed with a double volume of absolute ethanol and 1/10 th volume of 3M sodium-acetate (pH 5.2) and incubated at -20° C for overnight (Green et al., 2012). The mixture was then centrifuged (11,000 x *g*, 30 min; 4° C) and the supernatant was discarded. The pellet was resuspended with the remaining fluid (2-3 ml) inside the centrifuge bottle. The resuspended mixture was then incubated at -20° C for overnight and centrifuged (11,000 x g, 30 min; 4° C) again. After centrifugation, the supernatant was discarded and the pellet was washed with 1 ml of 70% ethanol (10,000 x *g*, 10 min; 4° C). The pellet was then dried and dissolved in 20 µl of 1x TE (pH 8.0) buffer and stored at -20° C as eDNA fraction.

2.3 Extraction of DNA from bacterial and VLP fraction

Before extraction of the nucleic acids from VLP fraction, the supernatant was treated with DNaseI to remove extracellular DNA as PEG is known to precipitate extracellular DNA (Paithankar and Prasad, 1991). The DNase I treated supernatant was then used to extract DNA using Genomic DNA Isolation Kit (Norgen Biotek, Thorold,

ON). The isolated DNA was further concentrated using a double volume of absolute ethanol and 1/10th volume of 3M sodium-acetate (pH 5.2) as described earlier and dissolved in 20 µl of 1x TE (pH 8.0). The concentrated DNA was then stored at -20^{o} C. The concentration of the DNA samples was measured using Qubit 2.0 Fluorometer (Invitrogen). DNA from bacterial fractions was isolated using Soil DNA Isolation Kit (Norgen Biotek, Thorold, ON) according to the manufacturers' protocol.

2.4 Preparation of DNA samples for sequencing

After extraction, DNA samples were diluted to a concentration of 0.2 ng/µl and libraries were prepared using Nextera XT DNA Sample Prep Kit (Illumina). Fragment size distribution of each library was checked using Bioanalyzer and High Sensitivity DNA Kit (Agilent). Samples were then normalized according to the supplier's instruction using Nextera XT DNA Sample Prep Kit. Five microliters of each library was pooled and concentrated. The final pool of each library was then quantified using qPCR and sequenced using 100 bp paired-end, HiSeq 2000, Illumina platform located at Farncombe Institute (McMaster University, Hamilton, ON, Canada). Sequence statistics listing number of libraries used and number of reads generated are included as supplementary information (**Table 2S.4**). Sequences have been submitted to the NCBI SRA database under the Study Accession SRP060006, Bioproject Accession PRJNA288501 and the accession number of each library has been included as supplementary information (**Table 2S.4**).

2.5 Bioinformatic analyses

Paired-end sequence reads generated from the Illumina HiSeq were assembled into contigs using CLC Genomics Workbench version 6.5.2 (CLC bio, Boston, MA,

USA). Assembly was performed using *de novo* assembly with automatic word and bubble sizes, a minimum of 66 nucleotides in the reads, mismatch cost set to 2, length fraction set to 0.5 and similarity fraction set to 0.8. In addition, colorspace error cost, insertion and deletion cost were set to 3 for assembly. Contigs with a length of less than 200 bp were not considered for analysis. The contigs were then submitted to the Metavir (http://metavir-meb.univ-bpclermont.fr/) pipeline for phylogenetic analysis (Roux et al., 2014). Raw sequences were also uploaded and analyzed using MG-RAST server (http://metagenomics.anl.gov/), an online metagenome annotation pipeline (Meyer et al., 2008). Before uploading, sequences were quality trimmed using MG-RAST QC pipeline, which includes removal of artificial or technical replicates (Gomez-Alvarez et al., 2009) and removal of low quality sequences (Cox et al., 2010). Gene-calling was performed using automated pipeline in MG-RAST which includes the use of FragGeneScan (Rho et al., 2010), clustering of predicted proteins at 90% identity by using uclust (Edgar, 2010) and the use of sBLAT, an implementation of the BLAT algorithm (Kent, 2002) for similarity analysis of each cluster. Sequence reads for which gene prediction tools could not identify a match, are considered as "unknown" which is divided into two categories – unknown sequences (sequences with no similarity to any protein or rRNA sequence) and unknown proteins (predicted proteins of unknown functions) (Meyer et al., 2008). Taxonomic composition of the viromes were obtained through comparison of sequences to the curated NCBI RefSeq complete viral genomes protein sequence database using blastX with an e-value cut-off of $\leq 10^{-3}$. Taxonomic composition was deduced from the best blast hit of each contig.

3. Results

3.1 Efficiency of the fractionation scheme

Efficiency of the fractionation scheme was determined using spiking experiment. A previously isolated bacteriophage MHS-16 (Host Strain: *E. coli* K-12) was spiked into the test water sample (which tested negative for the presence of any *E. coli* K-12 bacteriophage) and percent recovery of the phage was determined through enumeration of phage at every fractionation step (**Figure 2.1**). Briefly, MHS-16 was added to 400 ml water sample at a titer of 7.5 x 10^6 PFU/ml. Five hundred microliters of each fractionated sample (supernatant or resuspended pellet, filtered through 0.22 µm-pore-size filter) was then serially diluted (10-fold dilution) and used to enumerate bacteriophage using double agar overlay method (Green et al., 2012). One hundred microliter of the diluted phage particle was mixed with 500 µl of the logarithmic-phase host cells and 3.5 ml aliquots of 0.4% soft agar. The mixture was then poured on to the nutrient agar plate (LB agar) and incubated overnight. Recovery of phage, calculated by counting the number of plaques formed, was 86 ± 0.6% of the inoculated bacteriophage from spiked VLP fractions. The recovery was slightly low when compared with the recovery from purified phage suspension (~ 95%, data not shown) and this decrease may be due to the presence of inhibitory agents (such as organic molecules, clay minerals or charged particles) in the environment.

3.2 DNA in the environmental water samples

The amount of DNA recovered from isolated fractions was determined from the collected water samples of Lake Ontario and Lake Erie. A total of 12 samples (six

samples from each lake) was used to extract DNA from all three fractions. Recovery of DNA varied from 1.0 – 9.0 µg DNA from 1L water samples (**Figure 2.2 and Table 2S.5**). We also estimated the total DNA of each fraction and found that eDNA accounts for more than half (58.5% ± 5.1%) of the total DNA in a given sample, whereas bacterial and VLP fractions account for 20% ± 3.6% and 21.5% ± 4.8% respectively (**Figure 2.3**).

3.3 Abundance of bacterial and viral genome copies in freshwater

Number of bacterial and viral genome copies in freshwater environment was calculated by summing the DNA content from both bacterial and VLP fractions. Six samples from Lake Ontario (four samples from Lakeside Beach, one sample from Queen's Royal Beach and Fifty Point Beach) and six samples from Lake Erie (four samples from Long Beach and one sample from Long Beach Conservation Area East and Nickel Beach) were used to estimate the DNA amount. Assuming that average viral genome size of 50 kb length (dsDNA) (Steward et al., 2000), we converted the DNA mass to genome copy equivalent and found that the number of viral genome copies vary from 7-15 x 10^6 per milliliter of water sample which is consistent with previously published viral metagenomes (Bergh et al., 1989;Yoshida et al., 2013) (**Table 2.1**). We also estimated the number of bacterial genome copies assuming the average bacterial genome size of 4.7 Mb (Raes et al., 2007;Angly et al., 2009) and found that number of bacterial genomes varies from 1.8 x 10^4 to 1.1 x 10^5 per milliliter of water sample (**Table 2.1**).

3.4 Overview of the viromes

Percentage of sequences that show homology to the sequences in public databases was determined by uploading the sequence reads to MG-RAST and compared against the non-redundant M5NR database (Wilke et al., 2012). Consistent with the previously published freshwater viral metagenomes (Hewson et al., 2012a;Roux et al., 2012a), majority of our sequences did not show any homology to the known protein databases (ORFans) (**Figure 2.4**). The percentage of ORFans was slightly higher in Lakeside Beach viromes than the Long Beach viromes. Sequences with similarity to the known databases are known as ORFs and ORFs from the viral metagenomes were further divided into five major domains, namely, bacteria, viruses, eukaryotes, archaea and other. "Other" refers to sequences originating from mobile genetic elements such as plasmids and cloning vectors. Bacterial sequences predominated the viromes from both Lakeside and Long Beach samples (84% in Lakeside and 62% in Long Beach) (**Figure 2.5**). Viruses accounted for approximately 32% in Long Beach, and 12% in Lakeside Beach samples.

3.5 Taxonomic composition of the viromes

Taxonomic composition of the viral communities was determined through comparison of sequence reads against the curated non-redundant RefSeq Virus database using blastX with an e-value threshold of $\leq 10^{-3}$. Among annotated sequences, viruses of the *Myoviridae* family comprised the greatest proportion (79.7% ± 1.2%) whereas *Podoviridae* and *Siphoviridae* family comprised the second and third largest proportion of sequences, respectively (7.9% ± 1.5% and 4.5% ± 0.4% respectively) (**Figure 2.6, Table 2S.1 and 2S.2**). Viruses belong to these three families are dsDNA phages that infect only bacteria and therefore, majority of the viruses (~ 90%) in both Lakeside and Long Beach viromes are bacteriophages. However, algal viruses (*Phycodnaviridae*) and insect (or animal) viruses of the *Iridoviridae* family were also present (4.3% ± 0.6% and 2.6% ± 0.2% respectively). Apart from these major viral families, viruses belonging to other families such as *Asfarviridae*, *Poxviridae*, *Herpesviridae*, *Mimiviridae* and ssDNA phages of *Microviridae* and *Inoviridae* family are also present in these viromes (**Table 2S.1 and 2S.2**).

3.6 Comparison of the viromes

To determine the impact of geographical location on the distribution of viruses, we compared the virome content between six different beaches of Lake Ontario and Lake Erie. We did not observe any major difference in virome content among the six beaches tested suggesting that geographical location may not be a major determinant (within the context of this study) on the diversity of the viruses in the lower Great Lakes region. However, additional sampling is required to further validate this finding. Viruses

belonging to the *Myoviridae* family were predominant (~ 80%) in all beach sites whereas viruses of the *Podoviridae*, *Siphoviridae*, *Phycodnaviridae* and *Iridoviridae* family comprised the majority of the rest viral sequences (**Table 2.S2**). Relative abundance of viral families in eDNA fractions were consistent with the Lake Ontario and Lake Erie viromes (**Table 2.S1**) except viruses of the *Podoviridae* family which is found in slightly higher concentrations in eDNA fractions.

The distribution of viruses in aquatic environment changes over time (Hewson et al., 2012) and extreme climatic conditions such as typhoons, heavy rainfalls can also change the composition of microbial communities (Tseng et al., 2013). Change in host microbial community likely leads to commensurate changes in the viral community. Therefore, to investigate the impact of the environmental change and seasonal variation on the virome composition of the lower Great Lakes region, we compared the viromes of Lake Ontario and Lake Erie over the year 2012 and 2013. Viruses of the *Myoviridae* family, which still comprises the majority of the viruses in these lakes, their abundance dropped by 10% to 13% from 2012 to 2013 (**Table 2.2**) and viruses of the *Podoviridae* and *Siphoviridae* family increased. The relative abundance of other families remained similar over this period.

To compare the diversity within each lake, biological replication was determined by testing DNA recovery and variation among the top 5 virus families identified in replicate samples taken at three sites within the same beach. Although the recovery of DNA from the fractions (VLP, bacterial and eDNA fractions) varied highly among the replicates

(**Table 2S.5**), this high variability did not affect the abundance of highly abundant viruses (**Table 2.2, 2S.2 and 2S.3**).

3.7 Phylogenetic analysis of the viromes

The genetic diversity of the Lakeside Beach and Long Beach viromes (two representative beaches of Lake Ontario and Lake Erie respectively) were examined with the automated Metavir tool. Based on the Metavir data, two types of phylogenetic markers were used: G20, a well-known marker to assess the diversity of T4-like phages (or cyanophages) (Dorigo et al., 2004;Wilhelm et al., 2006) and TerL to assess the diversity of *Caudovirales* including *Myoviridae, Siphoviridae* and *Podoviridae* as these three families comprise over 90% of total viruses identified in the samples. Phylogenetic trees of the G20 and TerL sequences obtained from Lakeside and Long Beach viromes are shown in **Figure 2.7**, **2.8, 2.9** and **2.10**. Phylogenetic tree analyses did not indicate any major difference between Lakeside and Long Beach viromes, but minor differences were seen in branch length in these viromes (For example, branch length of Cyanophage-S-SM1, Syn30 and S-SSM6 differs in Lakeside and Long Beach viromes, **Figure 2.7** and **2.8**). These differences may be due to the difference in nucleic acid composition of these viruses. However, as full-length genome sequences were not used to construct these trees, it is very difficult to comment on the variation in genetic makeup of these viruses.

4. Discussion

In this pilot survey study, we employed a metagenomic approach to explore the diversity of viruses from representative sites in the lower Great Lakes region. Only a few species of viruses (or phages) have previously been identified in samples from this area (Fong et al., 2007;Short et al., 2011). As viruses infect all living beings, vital players in global geochemical and ecological cycling of key nutrients (C and N) (Suttle, 2005; 2007) and help maintaining balance in aquatic microbial communities (Rohwer and Thurber, 2009), exploring the nature of viral communities in the Great Lakes region may provide important insight into their role in shaping the microbial communities as well as in fixing and cycling of carbon and other nutrients. Here, we used a combinatorial approach using both VLP DNA and eDNA to fully capture the viral diversity. Resultant analyses of the viromes indicated high viral concentration ($> 10^7$ virus like genome equivalent per ml) in water samples collected from six different beaches across Lake Ontario and Lake Erie. To our knowledge, this is the first such study that employs eDNA, in combination with the VLP DNA to examine freshwater viral communities. eDNA has been used to identify non-viral species from freshwater environments (Ficetola et al., 2008;Thomsen et al., 2012), but it has never been used in the identification of viruses from aquatic ecosystem and our analysis suggests that eDNA can represent the major viral groups in a given sample (**Table 2.S1**). We determined the amount of DNA in the water sample and found that eDNA accounts for ~60% of all the DNA, whereas VLP and bacterial fractions account for ~21% and ~20% respectively. Apart from bacterial and VLP DNA, eukaryotes such as plants and animals contribute to the high proportion of eDNA in water

samples (Nielsen et al., 2007) and therefore, eDNA can be used to identify many organisms in the environmental samples.

Sample fractionation facilitates the separation and selective enrichment of viruses from the entire microbial community. We employed a fractionation scheme to identify all microorganisms, including viruses and bacteria from the environmental water samples. In contrast to other viral metagenomic studies, our fractionation scheme does not require large volumes of water samples (300 ml instead of 20 – 100L) (Thurber et al., 2009;Roux et al., 2012). To separate the VLP fraction, only 300 ml of water sample was concentrated and the virome obtained represents the majority of viruses that may be present in a given environment. The majority of the sequences (65% - 75%) from our viral metagenomes did not map to the publicly-available database (M5NR) (**Figure 2.4**). Presence of unknown viruses (or bacterial cells) in aquatic environment results in high percentage of unknown sequences in viral metagenomes (Breitbart et al., 2002b;Lopez-Bueno et al., 2009). However, as sequence read length is also known to influence estimates of the abundance of unknown sequences (Wommack et al., 2008), shorter sequences (~ 100 bp) generated by shotgun sequencing, certainly have contributed to the high percentage of unknown sequences.

The majority of the annotated sequences mapped to bacterial genomes (62% - 84%). Several factors may be responsible for this high percentage of bacterial sequences in viral metagenomes (Angly et al., 2006) including the presence of unknown prophages in bacterial genomes, phages carrying host genes, relatively large size of bacterial genomes

compared to viral genomes and the larger size of the microbial genome database than viral genome database statistically increasing the chance of matching bacterial sequences. The percentage of viral sequences among the annotated sequences was higher in Long Beach viromes (32%) than Lakeside Beach viromes (12%).

The majority of the viruses (90%) in Lake Ontario and Lake Erie viromes are bacteriophages and belong to the three major viral families (*Myoviridae*, *Siphoviridae* and *Podoviridae*) of bacteriophages (**Figure 2.6, Table 2.2, Table 2S.1, 2S.2 and 2S.3**). The high percentage of viruses (or phages) belonging to the *Myoviridae* family (~80%) is mainly due to the abundance of cyanophages and viruses of the abundant SAR11 and SAR16 (Pelagibacter) myoviruses. Genome size of the viruses belonging to these three major viral families ranges from 30 – 170 kb. However, genome size of the viruses belonging to the eukaryotic large NCLDVs (nucleocytoplasmic large DNA viruses) ranges from 100 kb - ~2.5 Mb. The NCLDVs are dsDNA viruses of the eukaryotes and include viruses from six families, including *Poxviridae*, *Asfarviridae*, *Iridoviridae*, *Ascoviridae*, *Mimiviridae* and *Phycodnaviridae*. Among these six viral families, viruses belonging to the *Phycodnaviridae* family are mostly algal viruses and viruses of the other five families are mainly animal (or insect) viruses. All these six viral families share some core genes and likely have originated from a putative common ancestor (Iyer et al., 2001;Iyer et al., 2006). Apart from these large viruses, viruses of relatively small genome size belonging to the *Microviridae* and *Inoviridae* family were also found in the Lake Ontario and Lake Erie beach viromes. Viruses of these two families are ssDNA phages and their genome size ranges from 4 – 9 kb.

Relative abundance of viral genotype changes over time (Hewson et al., 2012) and this was reflected in our 2012 and 2013 virome datasets. Viruses belonging to the *Myoviridae* family dropped from ~80% to ~70% in 2013 whereas, viruses of the *Siphoviridae*, *Podoviridae* and *Iridoviridae* family increased. Viruses belonging to *Myoviridae*, *Siphoviridae* and *Podoviridae* families are mainly bacteriophages. Therefore, the temporal dynamics of viral genotypes in these regions indicate a change in the microbial community as viruses depend on their hosts to survive and replicate. Viruses of the *Iridoviridae* family mainly infect invertebrates such as insects. However, viruses of vertebrates such as fish, reptiles and amphibians also belong to this family. The change in the abundance of these viral families may be due to factors such as climate change and precipitation events (Hewson et al., 2012;Tseng et al., 2013).

Our fractionation scheme, though efficient at capturing the majority of the viruses, may exclude some important viruses. As the efficiency of recovery was 86 ± 0.6%, some viruses may be lost during the concentration step of the virome preparation (**Figure 2.1**). This loss may lead to an under-representation of low abundant viruses or rare viruses in our datasets. Although giant viruses of eukaryotes are present in our prepared viromes, filtration of water samples through 0.22-µm-pore-size filter is known lead to the under-representation of large viruses (Fischer et al., 2010;Arslan et al., 2011;Philippe et al., 2013). Our fractionation scheme also does not consider the RNA viruses that are abundant in the aquatic environments (Djikeng et al., 2009;Culley et al., 2014) and is, therefore, unable to generate a comprehensive profile of all viruses present in a given sample. Furthermore, eDNA, which accounts for approximately 60% of the total DNA

per liter of water sample, the minimum amount of eDNA that is required to identify less-common species from the environment be further investigated.

Although the majority of virus types was captured through our fractionation protocol, more sampling is necessary to generate a comprehensive profile of all virus types. As RNA viruses cannot be detected simultaneously with DNA viruses using currently available techniques of virus discovery, with the advancement of technology, in the future, we hope to identify all viruses from lower Great Lakes water samples. From the metagenome libraries used in this study, we have assembled 205 complete viral genomes (**Table 2S.4**). Detailed characterization of these draft viral genomes, including their distribution and phylogenetic analyses, will be the subject of future investigation. In addition to viruses, identification of microbial populations using amplicon sequencing will provide valuable information regarding virus-host interaction, horizontal gene transfer and the emergence of pathogenic strains in these lakes. Frequent monitoring of seasonal and temporal variation of virus abundance can also provide important information about the nature of change in microbial community over a long period. The DNA content of freshwater samples is *a priori* unknown and may be highly variable. As only a few ng of DNA is needed to generate an Illumina HiSeq library (Binga et al., 2008) and about 1.0 – 9.0 µg DNA can be recovered from 1.0 L of freshwater sample, in future, we aim to use relatively smaller volume of water sample to identify microbial and viral communities from the environment. Complementary to current water analysis techniques, using low sample volume will allow municipal authorities to assess the

quality of freshwater environments especially in regions where regular monitoring is conducted.

In this study, we partnered with municipal water monitoring authorities to assess whether additional information obtained in metagenomic analyses, provides useful information for source tracking and/or estimation of public health risk (through the identification of sequences associated with pathogens). For large or heavily-used sites (e.g. popular beaches), multiple sampling locations (up to 5) are monitored with a given site with only one sample taken at each sampling location At the beginning of this study, we considered samples taken at these sub site locations to be replicates as there is systematic differences among them. However, it is clear from this study that DNA content can vary considerably among sub site samples and, in future, additional sampling, not currently included in municipal sampling programs, will have to be done to obtain good estimates of reproducibility.

5. Acknowledgements

We gratefully acknowledge financial support through the Niagara Region WaterSmart Program and the Natural Sciences and Research Council (NSERC) of Canada's Collaborative Research and Development Program (CRDP). We thank Schellhorn Lab members for comments on the manuscript.

References

Angly, F.E., B. Felts, M. Breitbart, P. Salamon, R.A. Edwards, C. Carlson, *et al.*, (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4**: e368.

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

Arslan, D., M. Legendre, V. Seltzer, C. Abergel & J.M. Claverie, (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci U S A* **108**: 17486-17491.

Bergh, O., K.Y. Borsheim, G. Bratbak & M. Heldal, (1989) High abundance of viruses found in aquatic environments. *Nature* **340**: 467-468.

Binga, E.K., R.S. Lasken & J.D. Neufeld, (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J* **2**: 233-241.

Branston, S., E. Stanley, E. Keshavarz-Moore & J. Ward, (2012) Precipitation of filamentous bacteriophages for their selective recovery in primary purification. *Biotechnol Prog* **28**: 129-136.

Breitbart, M., B. Felts, S. Kelley, J.M. Mahaffy, J. Nulton, P. Salamon, *et al.*, (2004) Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* **271**: 565-574.

Breitbart, M., P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, *et al.*, (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

Brussow, H., C. Canchaya & W.D. Hardt, (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**: 560-602, table of contents.

Chibani-Chennoufi, S., A. Bruttin, M.L. Dillmann & H. Brussow, (2004) Phage-host interaction: an ecological perspective. *J Bacteriol* **186**: 3677-3686.

Colombet, J., A. Robin, L. Lavie, Y. Bettarel, H.M. Cauchie & T. Sime-Ngando, (2007) Virioplankton 'pegylation': use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. *J Microbiol Methods* **71**: 212-219.

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

Culley, A.I., J.A. Mueller, M. Belcaid, E.M. Wood-Charlson, G. Poisson & G.F. Steward, (2014) The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. *MBio* **5**: e01210-01214.

Dejean, T., A. Valentini, A. Duparc, S. Pellier-Cuit, F. Pompanon, P. Taberlet, *et al.*, (2011) Persistence of environmental DNA in freshwater ecosystems. *PLoS One* **6**: e23398.

Djikeng, A., R. Kuzmickas, N.G. Anderson & D.J. Spiro, (2009) Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**: e7264.

Dorigo, U., S. Jacquet & J.F. Humbert, (2004) Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in France, Lake Bourget. *Appl Environ Microbiol* **70**: 1017-1022.

Edgar, R.C., (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.

Edge, T.A. & S. Hill, (2007) Multiple lines of evidence to identify the sources of fecal pollution at a freshwater beach in Hamilton Harbour, Lake Ontario. *Water Res* **41**: 3585-3594.

Fancello, L., S. Trape, C. Robert, M. Boyer, N. Popgeorgiev, D. Raoult, *et al.*, (2013) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *Isme Journal* **7**: 359-369.

Ficetola, G.F., C. Miaud, F. Pompanon & P. Taberlet, (2008) Species detection using environmental DNA from water samples. *Biol Lett* **4**: 423-425.

Fischer, M.G., M.J. Allen, W.H. Wilson & C.A. Suttle, (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci U S A* **107**: 19508-19513.

Fong, T.T., L.S. Mansfield, D.L. Wilson, D.J. Schwab, S.L. Molloy & J.B. Rose, (2007) Massive microbiological groundwater contamination associated with a waterborne outbreak in Lake Erie, South Bass Island, Ohio. *Environ Health Perspect* **115**: 856-864.

Fuhrman, J.A., (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541-548.

Gomez-Alvarez, V., T.K. Teal & T.M. Schmidt, (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME J* **3**: 1314-1317.

Green, M.R., J. Sambrook & J. Sambrook, (2012) *Molecular cloning : a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Hewson, I., J.G. Barbosa, J.M. Brown, R.P. Donelan, J.B. Eaglesham, E.M. Eggleston, *et al.*, (2012) Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes. *Appl Environ Microbiol* **78**: 6583-6591.

Iyer, L.M., L. Aravind & E.V. Koonin, (2001) Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* **75**: 11720-11734.

Iyer, L.M., S. Balaji, E.V. Koonin & L. Aravind, (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* **117**: 156-184.

Kent, W.J., (2002) BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Khan, I.U., V. Gannon, C.C. Jokinen, R. Kent, W. Koning, D.R. Lapen, *et al.*, (2014) A national investigation of the prevalence and diversity of thermophilic Campylobacter species in agricultural watersheds in Canada. *Water Res* **61**: 243-252.

Lopez-Bueno, A., J. Tamames, D. Velazquez, A. Moya, A. Quesada & A. Alcami, (2009) High diversity of the viral community from an Antarctic lake. *Science* **326**: 858-861.

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

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

Nielsen, K.M., P.J. Johnsen, D. Bensasson & D. Daffonchio, (2007) Release and persistence of extracellular DNA in the environment. *Environ Biosafety Res* **6**: 37-53.

Paithankar, K.R. & K.S. Prasad, (1991) Precipitation of DNA by polyethylene glycol and ethanol. *Nucleic Acids Res* **19**: 1346.

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

Raes, J., J.O. Korbel, M.J. Lercher, C. von Mering & P. Bork, (2007) Prediction of effective genome size in metagenomic samples. *Genome Biology* **8**.

Rho, M., H. Tang & Y. Ye, (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* **38**: e191.

Rohwer, F. & R. Edwards, (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**: 4529-4535.

Rohwer, F. & R.V. Thurber, (2009) Viruses manipulate the marine environment. *Nature* **459**: 207-212.

Roux, S., F. Enault, A. Robin, V. Ravet, S. Personnic, S. Theil, *et al.*, (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**: e33641.

Roux, S., J. Tournayre, A. Mahul, D. Debroas & F. Enault, (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**: 76.

Short, C.M., O. Rusanova & S.M. Short, (2011) Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada. *ISME J* **5**: 810-821.

Steward, G.F., J.L. Montiel & F. Azam, (2000) Genome size distributions indicate variability and similarities among marine viral assemblages from diverse environments. *Limnol Oceanogr* **45**: 1697-1706.

Suttle, C.A., (2005) Viruses in the sea. *Nature* **437**: 356-361.

Suttle, C.A., (2007) Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* **5**: 801-812.

Thomsen, P.F., J. Kielgast, L.L. Iversen, C. Wiuf, M. Rasmussen, M.T. Gilbert, *et al.*, (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Mol Ecol* **21**: 2565-2573.

Thurber, R.V., M. Haynes, M. Breitbart, L. Wegley & F. Rohwer, (2009) Laboratory procedures to generate viral metagenomes. *Nat Protoc* **4**: 470-483.

Tseng, C.H., P.W. Chiang, F.K. Shiah, Y.L. Chen, J.R. Liou, T.C. Hsu, *et al.*, (2013) Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J* **7**: 2374-2386.

Wilhelm, S.W., M.J. Carberry, M.L. Eldridge, L. Poorvin, M.A. Saxton & M.A. Doblin, (2006) Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of g20 genes. *Appl Environ Microbiol* **72**: 4957-4963.

Wilke, A., T. Harrison, J. Wilkening, D. Field, E.M. Glass, N. Kyrpides, *et al.*, (2012) The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* **13**: 141.

Williamson, K.E., K.E. Wommack & M. Radosevich, (2003) Sampling natural viral communities from soil for culture-independent analyses. *Appl Environ Microbiol* **69**: 6628-6633.

Wommack, K.E., J. Bhavsar & J. Ravel, (2008) Metagenomics: read length matters. *Appl Environ Microbiol* **74**: 1453-1463.

Yoshida, M., Y. Takaki, M. Eitoku, T. Nunoura & K. Takai, (2013) Metagenomic analysis of viral communities in (hado)pelagic sediments. *PLoS One* **8**: e57271.

Table 2.1 Relative abundance of bacteria and viruses in freshwater

Average genome copies	**Lake Ontario**	**Lake Erie**
Viral genome copies (x 10^6 ml^{-1})	10 (± 2.7)	9 (± 6.5)
Bacterial genome copies (x 10^4 ml^{-1})	3 (± 1.2)	8.9 (± 2.0)

Table 2.2 Temporal change of major viral families in the Lakeside and the Long Beach viromes (VLP fraction)

Virus Family	Primary Host	Relative abundance (% of viral reads)			
		Lakeside Beach		Long Beach	
		2012	2013	2012	2013
Myoviridae	Bacteria	80.76	70.70	79.66	66.82
Podoviridae	Bacteria	4.74	11.88	8.81	14.25
Siphoviridae	Bacteria	4.41	6.13	3.83	7.71
Unclassified (*Caudovirales* order)	Bacteria	0.78	0.84	0.88	0.93
Inoviridae	Bacteria	-	-	0.01	-
Phycodnaviridae	Algae	6.12	4. 83	4.12	4.44
Iridoviridae	Insects, Amphibians, Fish, Invertebrates	2.97	5.50	2.57	4.44
Poxviridae	Humans and other vertebrates, Arthropods	0.05	-	0.04	0.23
Alloherpesviridae	Fish, Amphibians	0.02	-	0.03	0.47
Herpesviridae	Animals, including humans	0.05	0.04	0.03	-
Marseilleviridae	Amoeba	0.06	0.04	0.01	0.23
Baculoviridae	Insects	0.03	0.04	0.02	0.23
Adenoviridae	Humans and other vertebrates	0.02	-	-	-
Nimaviridae	Crustaceans	0.02	-	-	0.23

Table 2S.1 Major viral families in the Lakeside Beach and the Long Beach viromes (VLP fraction)

Virus Family	Primary Host	Relative Abundance (% of viral reads)			
		VLP (Lakeside Beach)	**VLP (Long Beach)**	**eDNA (Lakeside Beach)**	**eDNA (Long Beach)**
Myoviridae	Bacteria	80.76	79.66	81.84	76.43
Podoviridae	Bacteria	4.74	8.81	6.67	11.57
Siphoviridae	Bacteria	4.41	3.83	4.03	5.54
Unclassified (*Caudovirales* order)	Bacteria	0.78	0.88	0.63	1.16
Inoviridae	Bacteria	-	0.01	-	0.01
Microviridae	Bacteria	-	-	-	0.01
Phycodnaviridae	Algae	6.12	4.12	3.79	3.28
Iridoviridae	Insects, Amphibians, Fish, Invertebrates	2.97	2.57	2.85	1.93
Poxviridae	Humans and other vertebrates, Arthropods	0.05	0.04	0.06	0.03
Alloherpesviridae	Fish, Amphibians	0.02	0.03	0.07	0.02
Herpesviridae	Animals including humans	0.05	0.03	0.01	0.01
Asfarviridae	Pigs	-	-	-	0.01
Marseilleviridae	Amoeba	0.06	0.01	-	0.01
Baculoviridae	Insects	0.03	0.02	0.06	-
Adenoviridae	Humans and other vertebrates	0.02	-	-	-
Nimaviridae	Crustaceans	0.02	-	-	-

Table 2S.2 Major viral families in the Lake Ontario and the Lake Erie viromes (VLP fractions)

Virus Family	Primary Host	Relative Abundance (% of viral reads)			
		VLP (Lakeside Beach)	**VLP (Long Beach)**	**eDNA (Lakeside Beach)**	**eDNA (Long Beach)**
Myoviridae	Bacteria	80.76	79.66	81.84	76.43
Podoviridae	Bacteria	4.74	8.81	6.67	11.57
Siphoviridae	Bacteria	4.41	3.83	4.03	5.54
Unclassified (*Caudovirales* order)	Bacteria	0.78	0.88	0.63	1.16
Inoviridae	Bacteria	-	0.01	-	0.01
Microviridae	Bacteria	-	-	-	0.01
Phycodnaviridae	Algae	6.12	4.12	3.79	3.28
Iridoviridae	Insects, Amphibians, Fish, Invertebrates	2.97	2.57	2.85	1.93
Poxviridae	Humans and other vertebrates, Arthropods	0.05	0.04	0.06	0.03
Alloherpesviridae	Fish, Amphibians	0.02	0.03	0.07	0.02
Herpesviridae	Animals including humans	0.05	0.03	0.01	0.01
Asfarviridae	Pigs	-	-	-	0.01
Marseilleviridae	Amoeba	0.06	0.01	-	0.01
Baculoviridae	Insects	0.03	0.02	0.06	-
Adenoviridae	Humans and other vertebrates	0.02	-	-	-
Nimaviridae	Crustaceans	0.02	-	-	-

Table 2S.3 Variation in relative abundance of virus families from Lakeside and Long Beach

Virus Family	% of total viral reads mapped to top five Viral Families									
	Lakeside Beach (VLP Fraction)					Long Beach (VLP Fraction)				
	Lakeside Beach - 1	Lakeside Beach - 2	Lakeside Beach - 3	Mean	Coefficient of Variation	Long Beach - 1	Long Beach - 2	Long Beach - 3	Mean	Coefficient of Variation
Myoviridae	80.76	82.74	81.84	81.78	1.21	79.66	82.69	48.67	70.34	26.77
Podoviridae	4.74	5.12	6.67	5.51	18.60	8.81	6.47	22.16	12.48	67.81
Siphoviridae	4.41	4.05	4.03	4.16	5.16	3.83	3.46	12.34	6.54	76.81
Phycodnaviridae	6.12	3.46	3.79	4.46	32.49	4.12	3.75	10.24	6.03	60.41
Iridoviridae	2.97	3.41	2.85	3.08	9.71	2.57	2.74	1.96	2.42	16.83

Table 2S.4 Metagenomic libraries used in the study

Sample #	Sample Name	Location	Latitude (W)	Longitude (N)	Date	# Reads	# Base pairs (Mbp)	# Contigs	Mean Contig I	GC Conte	# Accession	# Complete Phage Genome
1	Lakeside-1_VD	Lake Ontario	43.2043	79.2669	9-Aug-12	2,756,772	278	28,861	438	41	SRS976488	14
2	Lakeside-2_VD	Lake Ontario	43.2047	79.266	9-Aug-12	1,090,964	110	7,083	372	40.6	SRS976581	0
3	Lakeside-3_VD	Lake Ontario	43.2052	79.2648	9-Aug-12	3,096,970	313	37,085	454	41.1	SRS976582	8
4	Lakeside-4_VD	Lake Ontario	43.2043	79.2669	4-Jul-13	2,793,796	282	21,963	438	44.9	SRS976583	5
5	Queen's Royal_VD	Lake Ontario	43.2579	79.0687	9-Aug-12	3,645,734	368	39,607	455	41.5	SRS976584	14
6	Fifty Point_VD	Lake Ontario	43.2247	79.6183	9-Aug-12	3,647,628	368	35,690	518	46.8	SRS976585	16
7	Long Beach-1_VD	Lake Erie	42.8638	79.3862	12-Aug-12	3,498,802	353	43,035	472	41.3	SRS976586	19
8	Long Beach-2_VD	Lake Erie	42.8698	79.3982	12-Aug-12	1,818,202	275	44,144	508	38.8	SRS976587	26
9	Long Beach-3_VD	Lake Erie	42.8723	79.4187	12-Aug-12	3,283,240	496	81,039	530	40.4	SRS976589	22
10	Long Beach-4_VD	Lake Erie	42.8638	79.3862	17-Jul-13	2,142,442	216	15,935	447	42.7	SRS976590	9
11	Nickel Beach_VD	Lake Erie	42.8746	79.2323	12-Aug-12	2,992,076	302	35,052	468	41	SRS976592	14
12	Long Beach C.E._V	Lake Erie	42.8722	79.4272	12-Aug-12	1,122,618	113	10,876	413	40.5	SRS976593	6
13	Lakeside_ED	Lake Ontario	43.2043	79.2669	9-Aug-12	1,334,080	201	25,944	491	44.6	SRS976594	11
14	Queen's Royal_ED	Lake Ontario	43.2579	79.0687	9-Aug-12	3,136,238	317	20,510	467	46.8	SRS976595	7
15	Fifty Point_ED	Lake Ontario	43.2247	79.6183	9-Aug-12	2,291,674	231	21,848	407	54	SRS976596	2
16	Long Beach_ED	Lake Erie	42.8638	79.3862	12-Aug-12	2,481,398	375	55,296	488	41.9	SRS976597	25
17	Nickel Beach_ED	Lake Erie	42.8746	79.2323	12-Aug-12	832,424	126	18,290	448	43.2	SRS976599	6
18	Long Beach C.E._E	Lake Erie	42.8722	79.4272	12-Aug-12	1,530,500	231	7,210	703	41.4	SRS975751	1
					Total	**43,495,558**	**4,955**	**549,468**				**205**

Table 2S.5 Variation of DNA recovery from Lake Ontario and Lake Erie

Recovery of DNA (µg/L)																					
Fraction																					
	Lake Ontario										Lake Erie										
	Lakeside Beach-1	Lakeside Beach-2	Lakeside Beach-3	Lakeside Beach-4	Queen's Royal	Fifty Point	Mean	Std Dev	Coeffiecient of Variation (%)		Long Beach-1	Long Beach-2	Long Beach-3	Long Beach-4	Nickel Bt	Long Beach Conc	Mean	Std Dev	Coeffiecient of Variation (%)	
VLP DNA	0.77	0.26	0.6	0.81	0.32	0.79	0.59	0.25	41.76		1.19	0.19	0.08	0.15	1.22	0.72	0.59	0.53	88.98		
Bacterial	0.28	0.21	0.08	0.16	1.29	1.35	0.56	0.59	105.28		0.3	0.64	0.42	0.92	0.90	1.24	0.74	0.35	47.46		
eDNA	0.5	1.79	1.84	0.45	1.54	6.19	2.05	2.12	103.25		2.3	1.1	0.81	7.68	3.09	1.10	2.68	2.60	97.00		
Here, Lakeside Beach-4 and Long Beach-4 samples represent samples collected in 2013. The rest of the samples were collected in 2012.																					

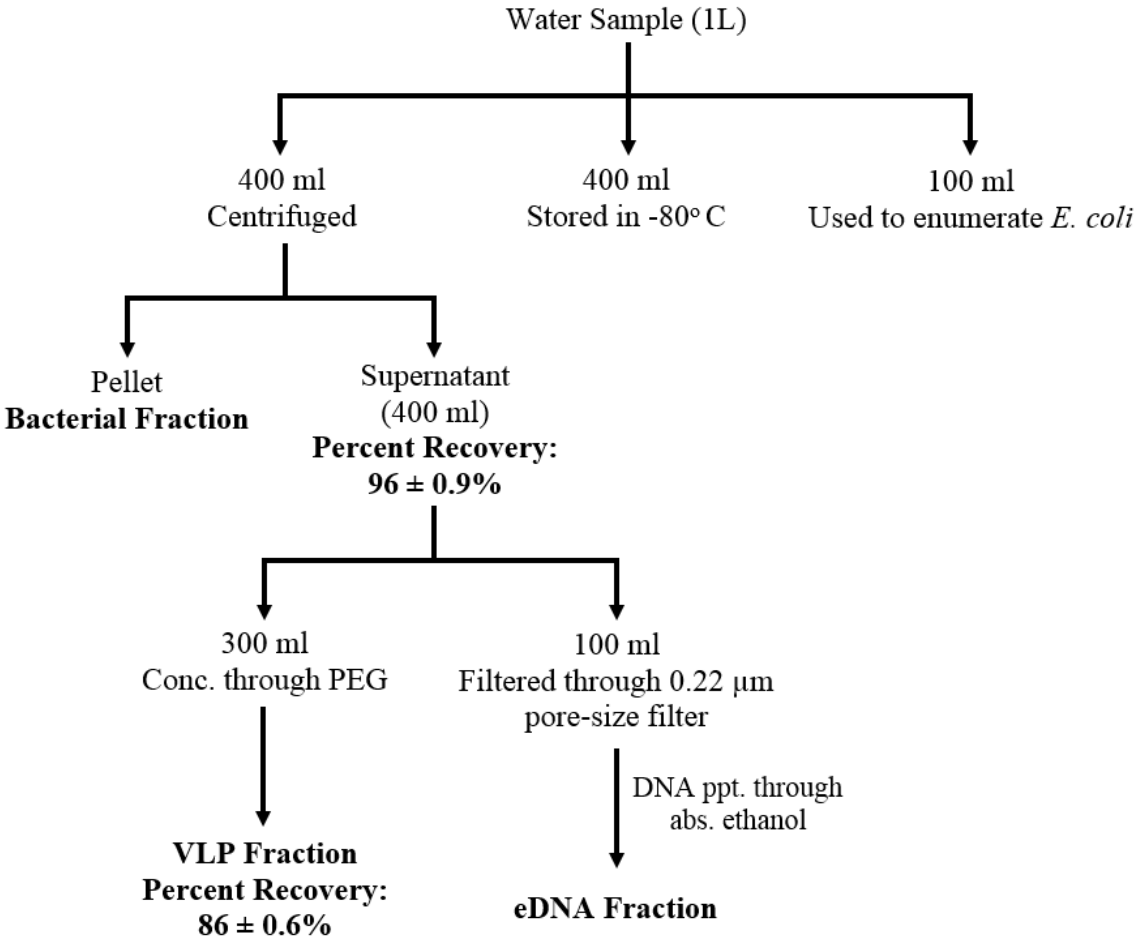

Figure 2.1 Schematic diagram showing water sample fractionation steps and the efficiency of the fractionation scheme. Water samples were separated into three distinct fractions: bacterial, VLP (Virus like particles) and eDNA fraction. eDNA is defined as the DNA released in the environment after lysis (natural decay or lysis) of bacteria, viruses and higher organisms. The efficiency of the fractionation scheme was determined as percent recovery of bacteriophage (MHS 16) mixed with the water sample prior to the fractionation steps.

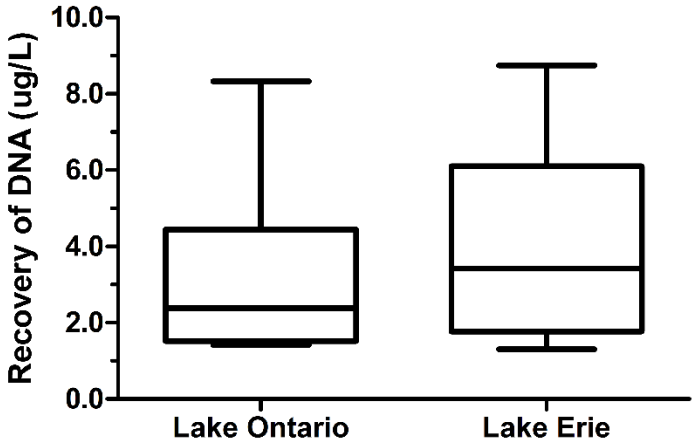

Figure 2.2 Recovery of DNA from freshwater environment. DNA was extracted from three different fractions of twelve individual 1L samples. Six samples from Lake Ontario (four samples from Lakeside Beach, one sample from Queen's Royal Beach and Fifty Point Beach) and six samples from Lake Erie (four samples from Long Beach and one sample from Long Beach Conservation Area East and Nickel Beach) were used to extract DNA and amount of DNA recovered from all three fractions were normalized to 1L water sample.

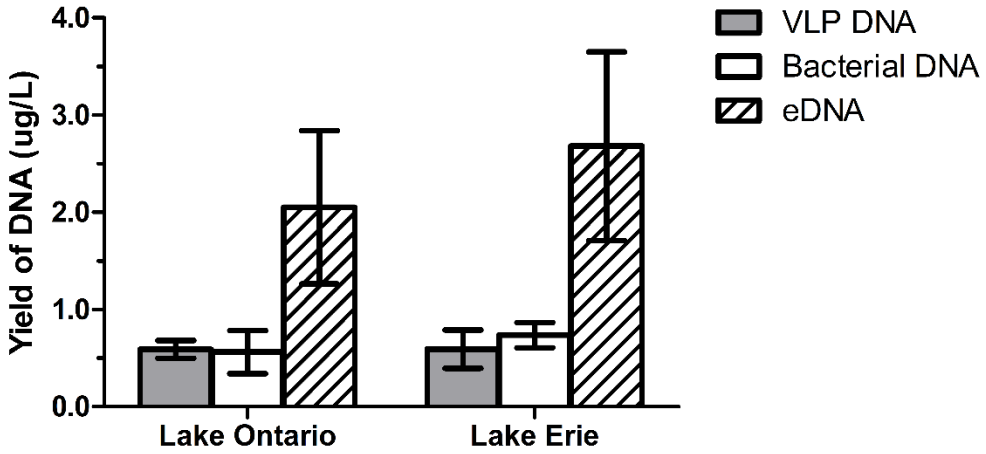

Figure 2.3 DNA yield from VLP, bacterial and eDNA fraction. DNA recovered from six samples from Lake Ontario (four samples from Lakeside Beach, one sample from Queen's Royal Beach and Fifty Point Beach) and six samples from Lake Erie (four samples from Long Beach and one sample from Long Beach Conservation Area East and Nickel Beach) were used to calculate DNA yield from each fraction. eDNA accounts for most of the DNA (~60%) in freshwater environment while VLP DNA and bacterial DNA are present in almost equal amount.

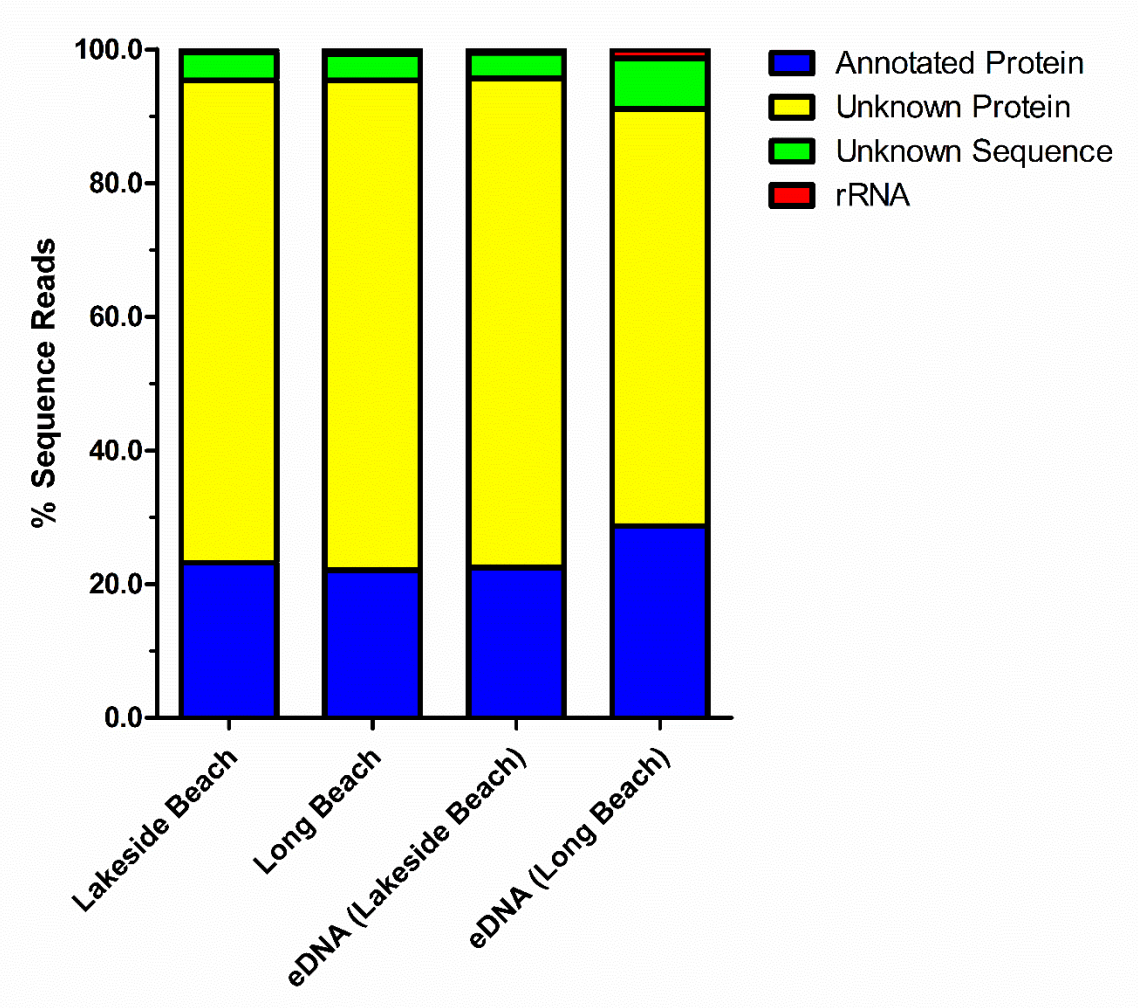

Figure 2.4 Annotation of the virome sequence reads. Sequences from both VLP DNA and eDNA fractions of Lakeside and Long Beach samples (collected in 2012) were compared against the non-redundant M5NR database (E-value $<10^{-3}$ in blastX). About 25% of the sequence reads mapped to the known protein sequences of M5NR database while majority of the sequence reads were categorized as unknown. Among the unknown reads, unknown sequences are sequences for which gene prediction tools could not predict any protein or rRNA sequence and unknown proteins represent predicted proteins with unknown functions.

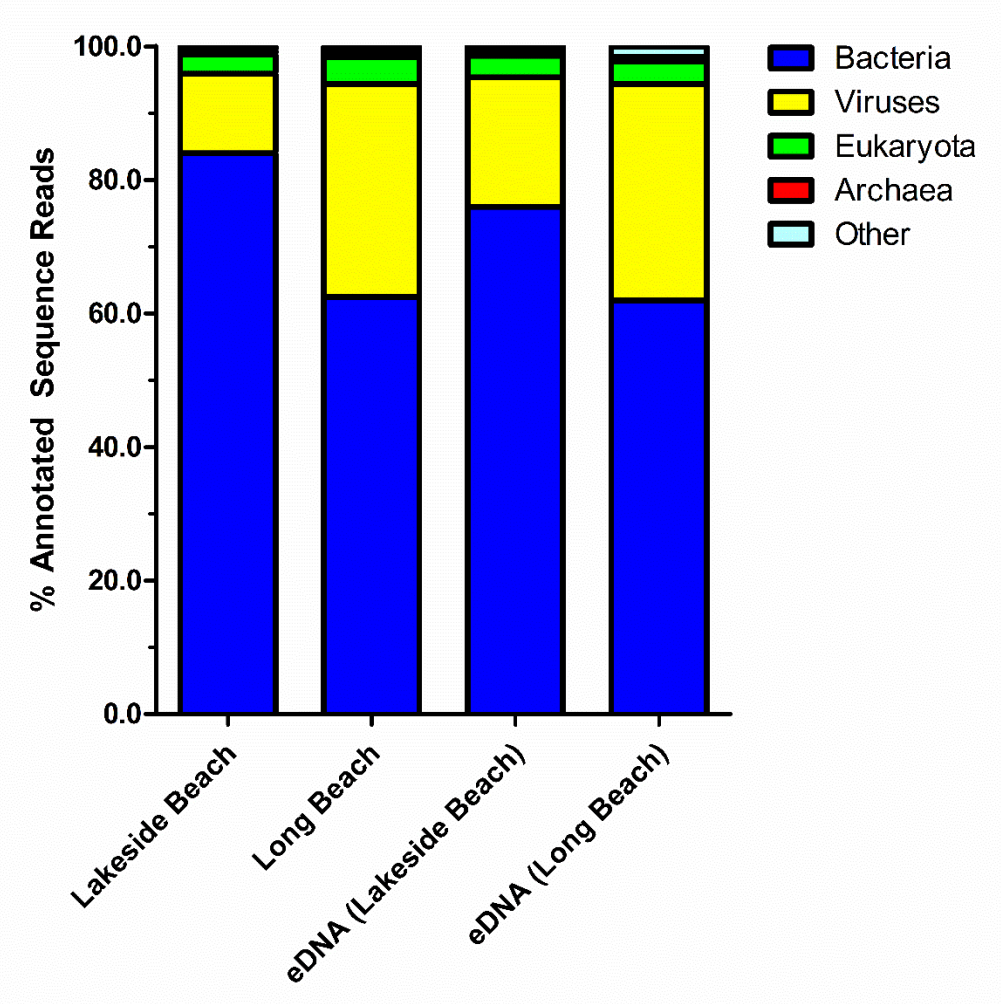

Figure 2.5 Taxonomy of virome domains. Sequence reads from Lakeside Beach and Long Beach VLP DNA and eDNA fractions (of the year 2012) were compared against the M5NR database (E-value $<10^{-3}$ in blastX). Majority of the sequence reads (60% to 80%) mapped to bacterial sequences and 20% - 35% sequence reads mapped to virus sequences of the database.

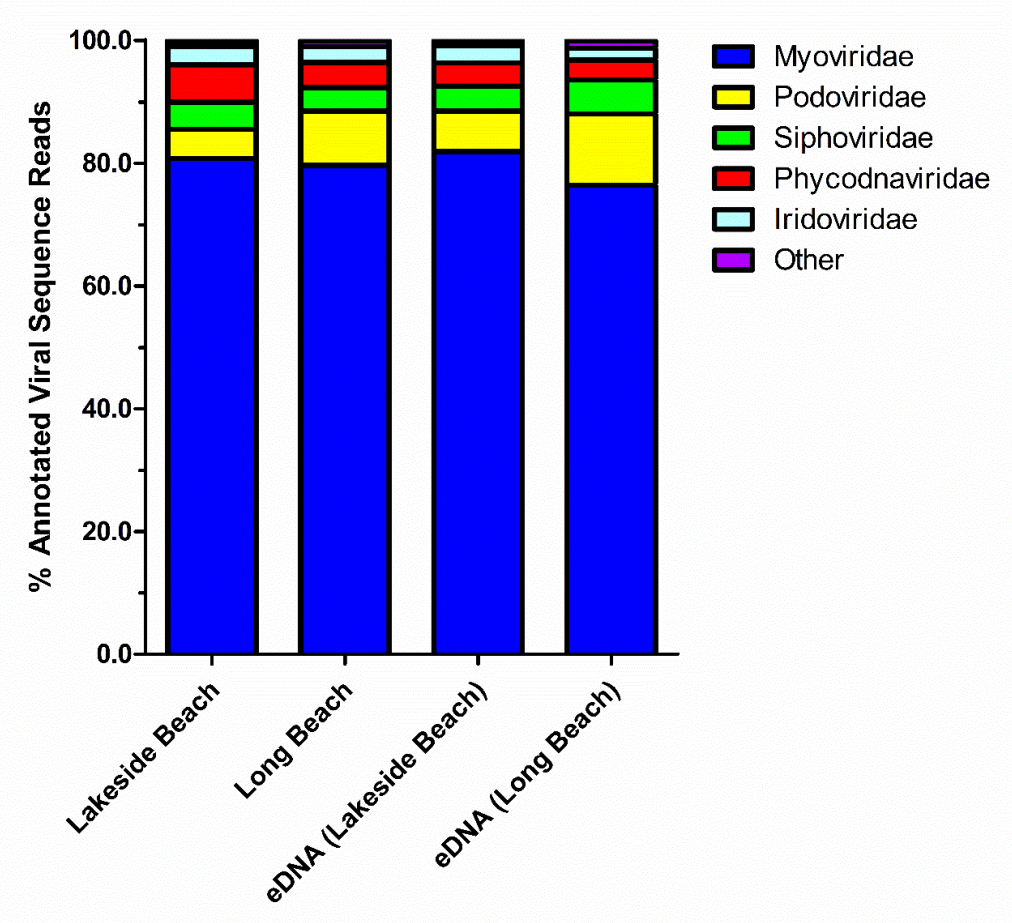

Figure 2.6 Relative abundance of top 5 viral families in the lower Great Lakes water samples. blastX comparison (E-value < 10^{-3}) of VLP DNA and eDNA sequences of Lakeside Beach and Long Beach samples (sampled in 2012). Unclassified viruses and viral families constituting less than 1% of total viral reads are categorized as "Other."

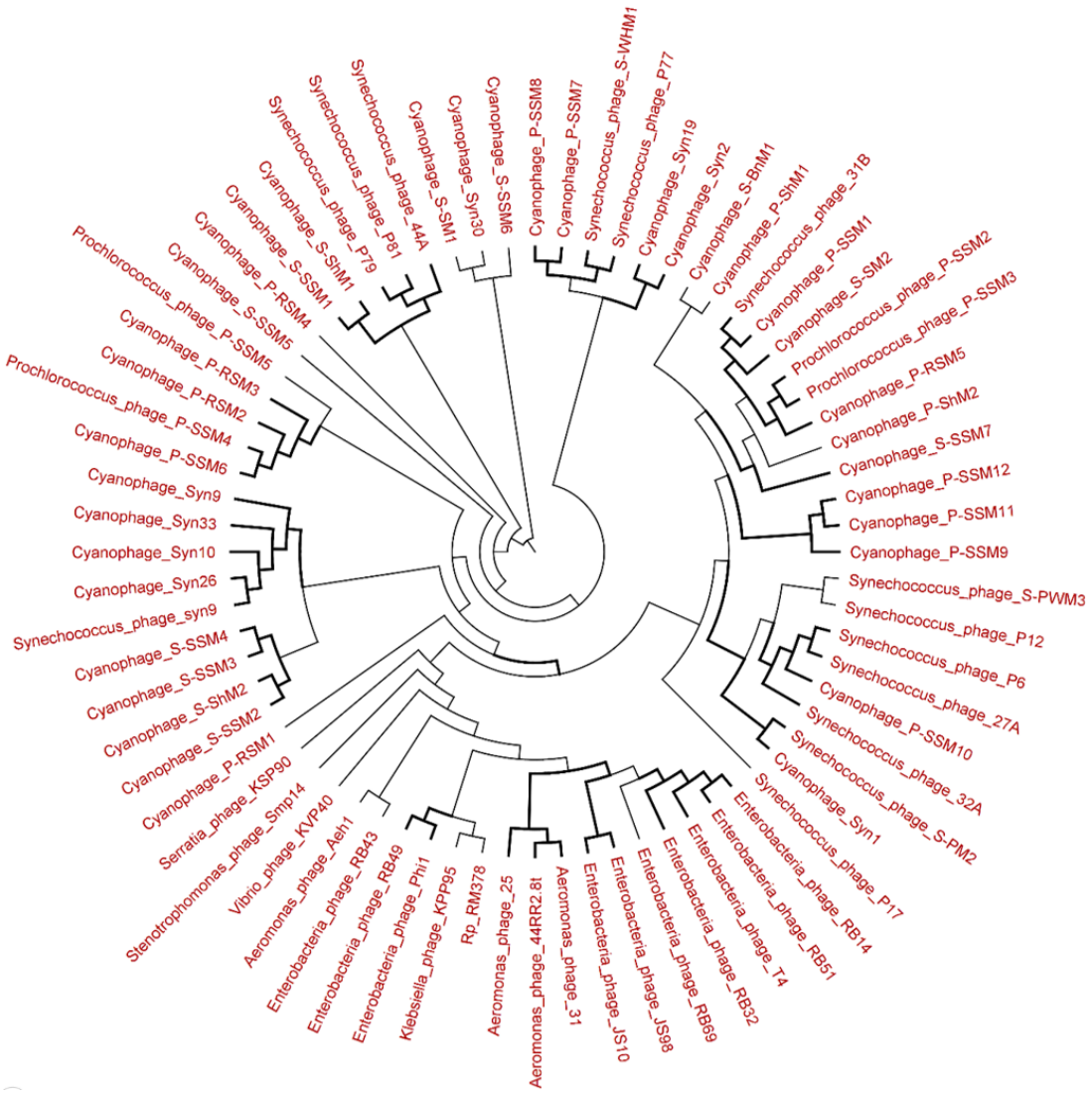

Figure 2.7 Neighbour-joining phylogenetic tree of the capsid assembly protein G20 (T4-like phages or cyanophages) (pfam07230) of Lakeside Beach virome. Lakeside -1 virome library was used as a representative metagenome profile for the Lakeside Beach site. Sequence reads with significant similarity (E value < 10^{-3} using blastX) to the G20 marker sequences were obtained, assembled at 98% identity in 35 bp using Cap3 and used to draw the phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Sequences for which the best blast hit did not correspond to the G20 marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines.

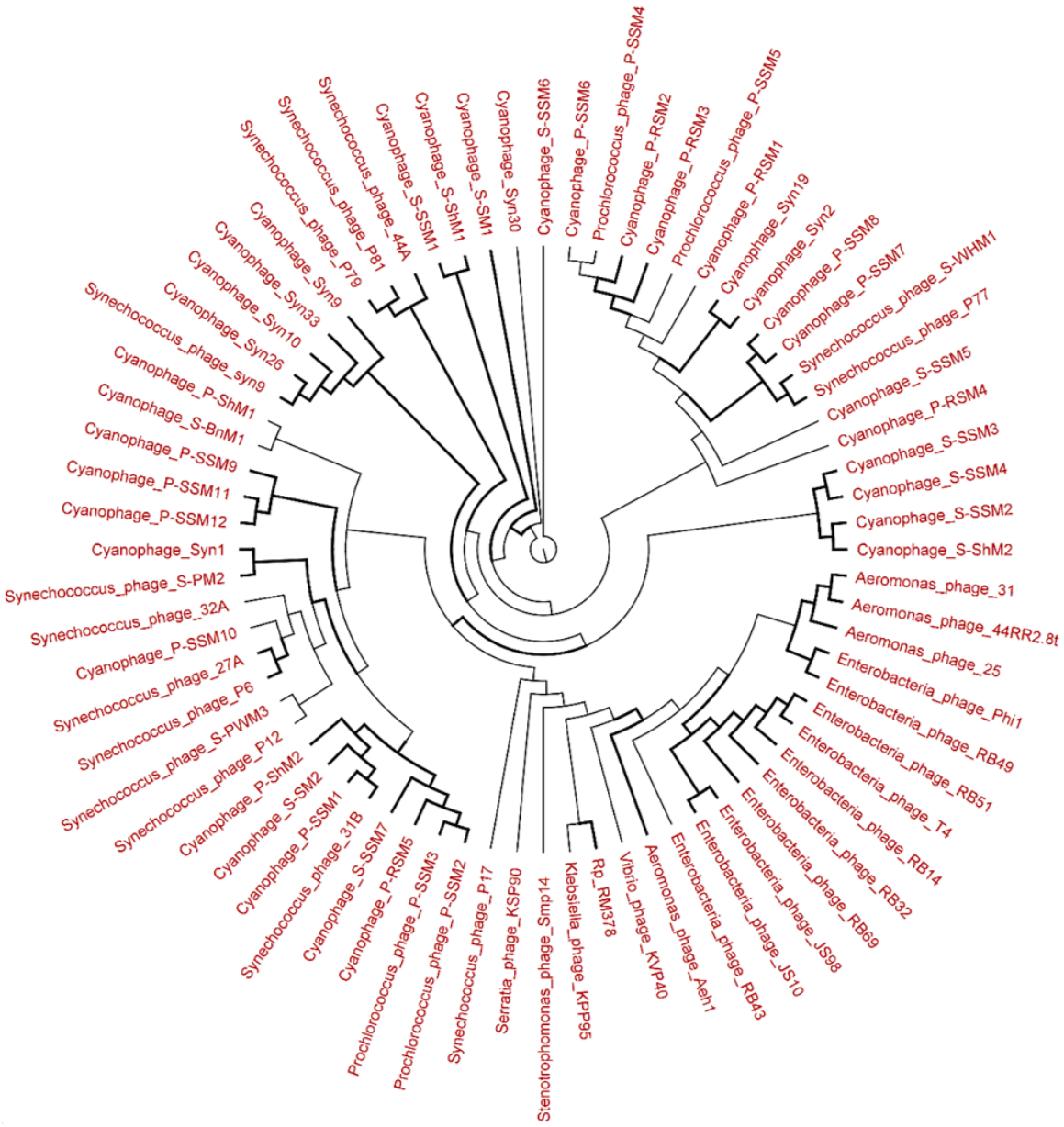

Figure 2.8 Neighbour-joining tree of the capsid assembly protein G20 (T4-like phages or cyanophages) (pfam07230) of Long Beach virome. Long Beach -1 virome library was used as a representative metagenome profile for the Long Beach site. Sequence homologs (E value < 10^{-3} using blastX) to the G20 marker sequences were obtained from the virome library, assembled at 98% identity in 35 bp using Cap3 and used to draw phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Sequences for which the best blast hit did not correspond to the G20 marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines.

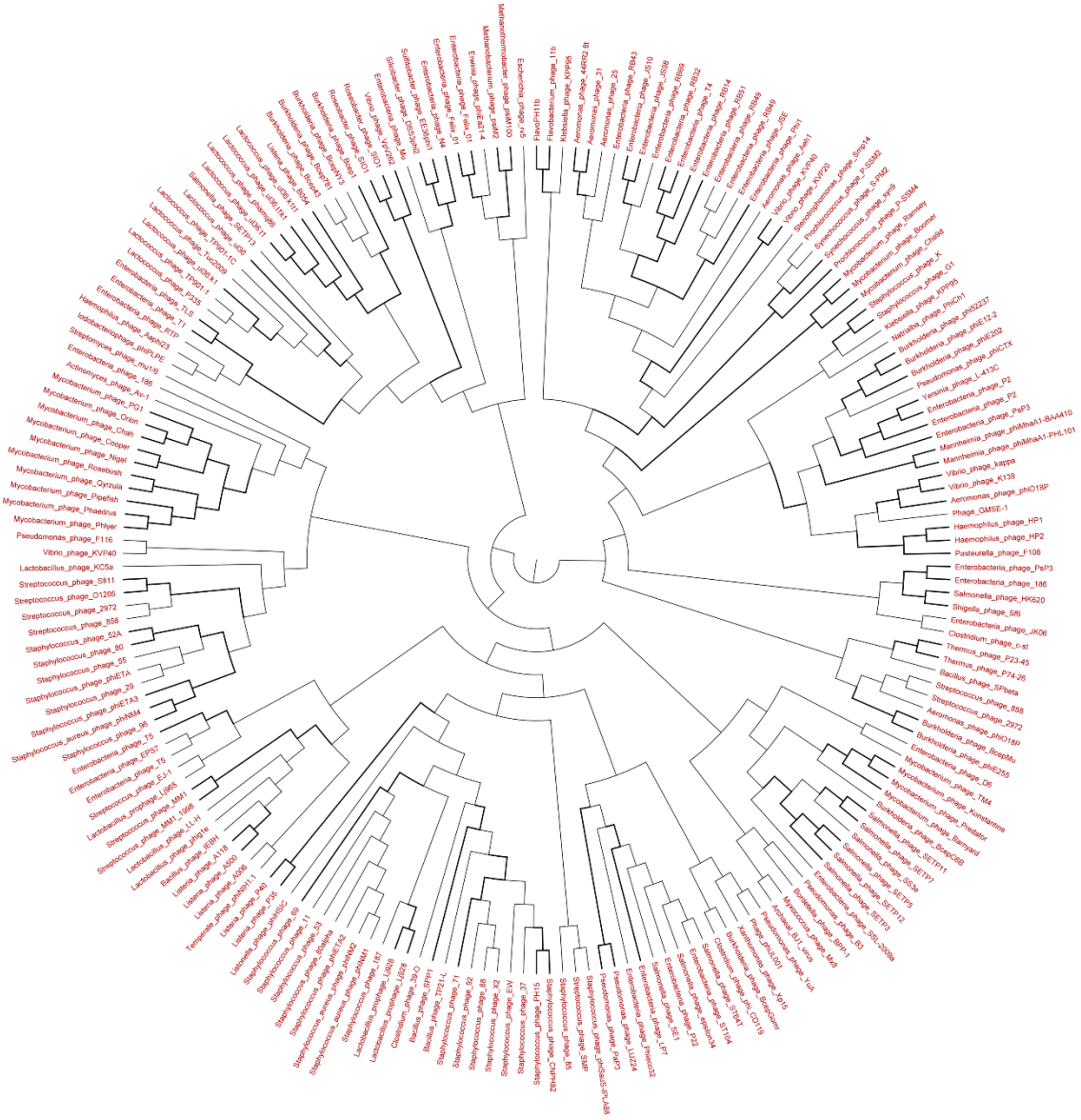

Figure 2.9 Neighbour-joining phylogenetic tree of the TerL (Terminase Large subunit – for *Caudovirales* including *Myoviridae*, *Podoviridae* and *Siphoviridae*) (pfam03237) of Lakeside Beach virome. Lakeside -1 virome library was used as a representative metagenome profile for the Lakeside Beach site. Sequence reads with significant similarity (E value < 10^{-3} using blastX) to the TerL marker sequences were obtained, assembled at 98% identity in 35 bp using Cap3 and used to draw phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Bootstrap values of ≥ 80 are highlighted with black lines. Sequences for which the best blast hit did not correspond to the TerL marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines.

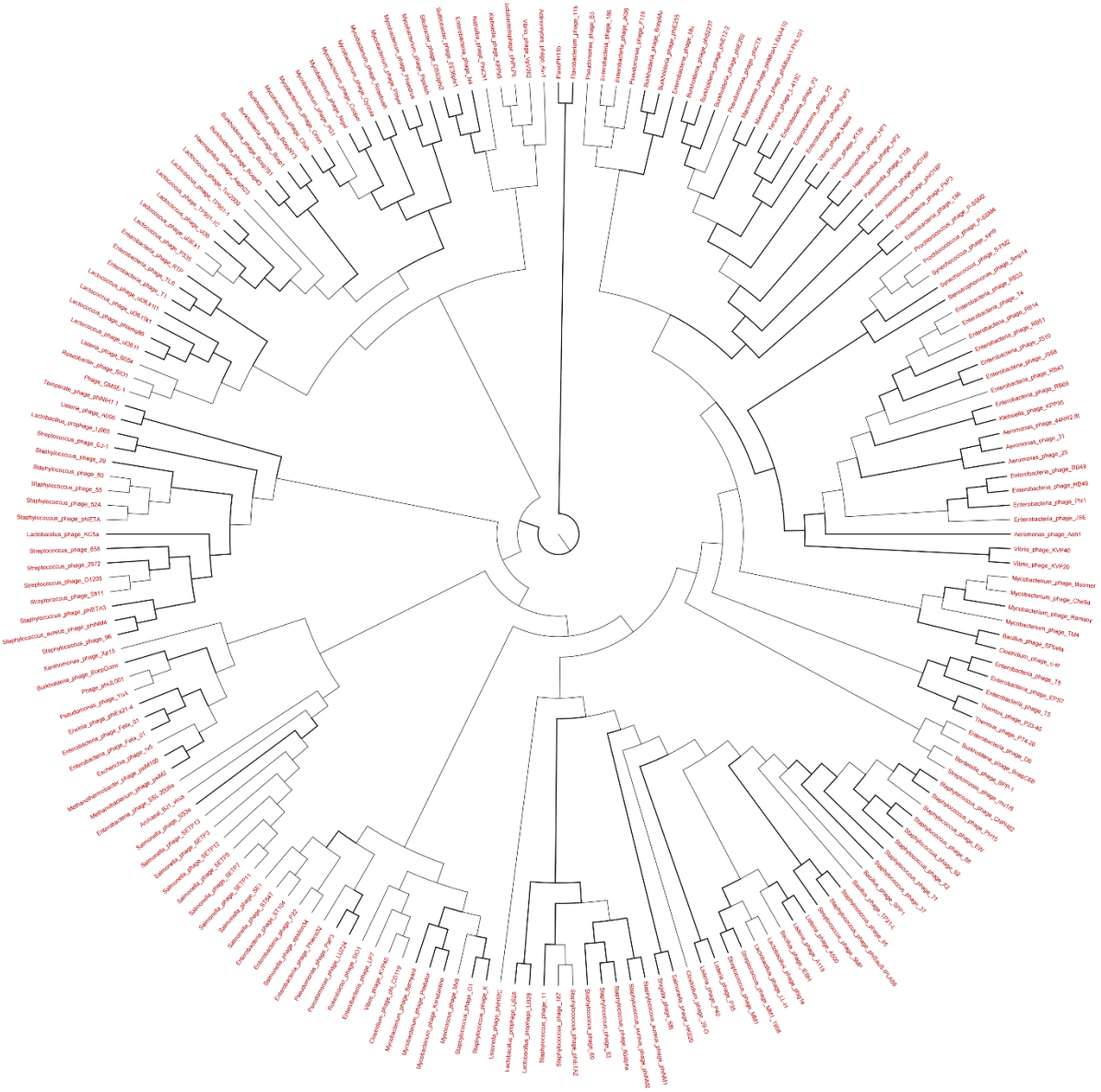

Figure 2.10 Neighbour-joining phylogenetic tree of the TerL (Terminase Large subunit – for *Caudovirales* including *Myoviridae*, *Podoviridae* and *Siphoviridae*) (pfam03237) of Long Beach virome. Long Beach -1 virome library was used as a representative metagenome profile for the Long Beach site. Sequence homologs (E value < 10^{-3} using blastX) to the TerL marker sequences were obtained from the virome library, assembled at 98% identity in 35 bp using Cap3 and used to draw phylogenetic tree alongside reference sequences taken from the protein family (PFAM) database. Sequences for which the best blast hit did not correspond to the TerL marker were excluded from analysis. Bootstrap values of ≥ 80 are highlighted with black lines.

Chapter 3: Shotgun metagenomic sequencing reveals freshwater beach sand as reservoirs of bacterial pathogens

Mahi M. Mohiuddin[1]*, Yasser Salama[1]*, Herb E. Schellhorn[1], G. Brian Golding[1¶]
*These authors contributed equally to this work
[1]Department of Biology, McMaster University, Hamilton, ON, Canada

[¶]Correspondence: G. Brian Golding, LSB 533, Department of Biology, McMaster University, 1280 Main St W, Hamilton, ON, L8S 4k1, Canada.
E-mail: golding@mcmaster.ca

Reproduced with permission from Mohiuddin, M.M., Y. Salama, H.E. Schellhorn & G.B. Golding, (2017) Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Research* 115: 360-369. Elsevier B.V.

Abstract

Recreational waters and adjacent beach sands harbor complex microbial communities which may contain human pathogens that cannot be detected by conventional methods. Here, we investigate the diversity of bacterial populations inhabiting four freshwater beaches of the Great Lakes region using shotgun metagenomic sequencing approach. Our analysis suggests that average taxonomic richness and alpha diversity are significantly higher ($P < 0.001$) in beach sands compared to the corresponding water environments. Compared to the water environments, beach sands harbored taxa from a more diverse range of phyla, including a higher proportion of sequences from unclassified phyla. Unique phyla were also identified in sand which included species from Aquificae, Candidatus Microgenomates, Latescibacteria, and Candidatus Aminicenantes. Sequences originating from pathogens were detected in both sand and water, with some pathogens enriched in both environments. Both lakes exhibited similar community composition suggesting that geographic location did not appear to have any major impact on bacterial diversity. These findings reveal the diversity of bacterial communities of freshwater beaches and highlight the importance of monitoring pathogens in recreational beaches, especially in the sand environment of these beaches.

1. Introduction

Recreational waters may be heavily impacted by adjacent sands which can harbor high levels of bacteria and viruses. Bacterial concentrations in beach sands are often 10-100-fold higher than corresponding recreational waters (Alm et al., 2003;Cui et al., 2013). Beach sand also harbors and supports the growth and persistence of fecal indicator bacteria and other potentially pathogenic species (Byappanahalli et al., 2012;Yamahara et al., 2012), which suggests that this environment is inhabited by numerous pathogens and has important public health impacts. Due to more frequent exposure to human activities and the diverse nature of the inhabiting bacterial populations, marine beaches, compared to freshwater beaches, have received more attention in recent years. However, studies investigating freshwater environments indicate that freshwater bacterial populations are as diverse and as enriched as those from marine environments, suggesting that a more comprehensive characterization of these environments is needed (Tamames et al., 2010;Wang et al., 2012).

The majority of bacterial species in freshwater environments, especially in freshwater lakes and their associated beaches, are indigenous to that particular environment (Staley and Sadowsky, 2016). However, beach sand bacterial species, depending on the distance from the shoreline, differ substantially from aquatic species within the same lake (Staley and Sadowsky, 2016). Microbial communities adapt to water and to the different parts of the beach cross-section, resulting in unique communities between different environments of the beach (Whitman et al., 2014). However, these unique communities are also transported or dispersed between different compartments of

the beach environments, further modifying these communities (Ge et al., 2010;Russell et al., 2012;Boehm et al., 2014) . Bacterial populations harboring fecal indicator bacteria (FIB) (e.g. *Escherichia coli*, fecal coliforms and enterococci) and pathogens are also introduced into the beach environments through point-source (wastewater) (Vijayavel et al., 2010a) and non-point-source pollution (Piggot et al., 2012), direct fecal discharge from birds, humans and other animals (Edge and Hill, 2007;Wright et al., 2009;Khan et al., 2013), or transmission of fecal discharged pathogens from water to the sand (Piggot et al., 2012). Introduction of bacterial species from external sources to the beach environment is important from both an ecological and a public health perspective due to their fate in the new environment and their potential association with human illness. Adequate characterization of these communities is therefore needed to determine their ecological and human health impacts.

While the recent developments in sequencing technologies have allowed robust characterization of microbial communities in different environments (Costello et al., 2009;Kuramae et al., 2012;Staley and Sadowsky, 2016), most studies have employed amplicon sequencing targeting the conserved 16S rDNA and focused mostly on identifying predominant bacterial species. Although amplicon sequencing approaches have some advantages in identifying bacterial populations, a limited resolution at the species level (Poretsky et al., 2014;Ranjan et al., 2016) and the inability to infer functional capacity of the microbial communities are major limitations. Determining the functional potential of bacterial species is a prerequisite as microbial communities shape the global biogeochemical processes (Newman and Banfield, 2002;Sunagawa et al.,

2015). Shotgun metagenomic sequencing, compared to amplicon sequencing, provides better resolution at the species level and analysis of such sequences helps in understanding the structure and diversity of microbial communities as well as their metabolic potential. Few studies have examined the microbial community composition in the lower Great Lakes region, particularly in Lake Ontario and Lake Erie (Bouzat et al., 2013;Staley and Sadowsky, 2016). These studies, however, due to the limitation of the sequencing technology employed, did not consider the functional diversity and the potential role of microbial communities in beach environments. Prior studies also did not investigate the unique characteristics of beach sand and water bacterial populations, presence of rare bacterial taxa, and pathogenic microorganisms. However, studies performed on wave-washed sands using culture-based methods have consistently identified *E. coli* and *Enterococcus* in these environments (Alm et al., 2003;Cui et al., 2013). Additionally, pathogenic species including *Aeromonas* spp. (Khan et al., 2009), *Campylobacter* spp. (Yamahara et al., 2012), *Salmonella* spp. (Yamahara et al., 2012), and *Vibrio* spp. (Shah et al., 2011) are also present in marine and freshwater beaches. Therefore, a more detailed investigation into human pathogens inhabiting these environments is needed.

To provide a robust and comprehensive analysis of the bacterial species that inhabit freshwater beach environments, and to detect pathogenic bacterial species, we collected water and sand pore samples from four different beaches of Lake Ontario and Lake Erie over two years and performed shotgun metagenomic sequencing. Sand pore samples were collected due to their standard use in municipal sampling in the lower Great

Lakes region (Whitman et al., 2014). Both lakes assessed are economically important as they are the primary source of drinking water in this region and the beaches are heavily used for recreational activities during the summer. As a popular attraction for many residents of the area, these beaches have warranted strong municipal monitoring and are subjected to frequent closures due to high concentrations of fecal indicator bacteria. To better characterize the bacterial communities present at the beach water and sands, and to identify pathogens, we employed shotgun metagenomic sequencing. The information obtained from our analyses can aid in future developments in water quality monitoring and further our understanding of the natural bacterial inhabitants at these beaches.

2. Materials and methods

2.1 Sampling and DNA extraction

Water samples were collected in sterile 1.0 L sampling bottles (Nalgene) from four different locations across Lake Ontario and Lake Erie. Samples were collected from Lakeside Beach and Fifty Point Beach of Lake Ontario and from Long Beach and Nickel Beach of Lake Erie during the summer period of 2012 and 2013 (see supplementary data Google Map and **Table 3S.1**). Water samples were collected at the surface level of 1 m-deep water. Sand samples were obtained by digging a 30 x 30 x 30 cm hole 1m into the supratidal region of the beach and subsequent collection of the water which permeated through the sand into the hole. Both water and beach sand samples were collected from the same sampling site at the same time. A total of 32 samples, comprising 16 pairs of water and sand pore samples, were collected from these beaches. Details of the samples

including the geographic location of the sampling sites and date of collection are listed as supplementary data (**Table 3S.1**).

After collection, samples were kept on ice and processed within six hours of collection. Three hundred milliliters of water from each sample was filtered through a 0.45 µm membrane filter (Fischer Scientific, Ottawa, ON) and the filter paper was then stored at -20^{o} C for processing. The retentate of the filter paper enriches for the microbial fraction (APHA, 2012). The filter paper was then cut into fragments (0.25 cm^{2} size) and DNA was extracted from the microbial fraction using the PowerSoil DNA Isolation Kit (MO BIO, Carlsbad, CA) according to manufacturer's instructions and extracted DNA was stored at -20^{o} C until used for library preparation and next generation sequencing.

2.2 Library preparation and sequencing

Library preparation and sequencing was performed as described earlier (Mohiuddin and Schellhorn, 2015c). Briefly, DNA samples were diluted to a concentration of 0.2 ng/µl, and subsequently used to prepare libraries using Nextera XT DNA Sample Prep Kit (Illumina). Bioanalyzer and High Sensitivity DNA Kit (Agilent) were used to check the fragment size distribution of each library. Each library was then quantified using qPCR and sequenced using 100 bp paired-end, HiSeq 2000, Illumina platform located at the Farncombe Metagenomics Facility (McMaster University, Hamilton, ON, Canada). Sequence statistics listing the number of libraries used and the number of reads generated are included as supplementary data (**Table 3S.1**). Sequences are submitted to the NCBI SRA database under the Study Accession SRP081250,

BioProject Accession PRJNA335374 and the accession number of each library are included (**Table 3S.1**).

2.3 Bioinformatic analysis

Sequence reads obtained from Illumina HiSeq 2000 were processed for low-quality bases and adapter sequences using BBDuk from the BBMap package version 35.82 (http://sourceforge.net/projects/bbmap/) and quality-screened using FastQC (version 0.11.3). BBDuk processing parameters were set to right-end and left-end quality trimming with right side adapter trimming using a *k*-mer size of 12. The quality trimming threshold was set to 20 and a minimum sequence length of 40 bp were considered for analysis. Singleton reads whose other pair was discarded were also included in downstream analysis.

Processed reads were compared against the NCBI non-redundant protein database (nr) using the DIAMOND variant of BLASTx (Buchfink et al., 2015). Reads obtained from BLAST comparison were parsed and taxonomic classification was assigned using MEGAN (version 5.10) (Huson et al., 2007) with default classification parameters. Both taxonomic and functional annotations were assigned using NCBI taxonomy (Sayers et al., 2009) and SEED subsystems (Overbeek et al., 2014) within MEGAN. Each member of a paired read was classified independently, and a custom Perl script (Supporting data Perl_Script) was used to arrive at a consensus classification for a read pair using the average bit score of the BLAST hits within 10% of the top bit score for each member of the pair. For pathogen detection, reads were classified at the species level using CLARK

(version 1.1.2) (Ounit et al., 2015) against the RefSeq bacterial database. The CLARK confidence score threshold was set to 95%.

2.4 Statistical analysis

Taxonomic and functional assignments were extracted from MEGAN and were used to generate count matrices for use in R (version 3.2.3). Taxonomic assignments were translated to taxonomic ranks of phylum, class, order, family, genus, and species using the taxize package (version 0.7.4) in R. The metagenomeSeq package (version 1.12.0) (Paulson et al., 2013) in R was used for count aggregation at different taxonomic levels. The vegan package (version 2.3-4) was used for ecological diversity analysis, testing of homogeneity of dispersion, and multivariate variance partitioning with permutational multivariate analysis of variance (permutational MANOVA). DESeq2 (version 1.10.1) was used for differential abundance testing of taxonomic and functional features (Love et al., 2014). Differentially abundant taxonomic features were determined at FDR < 0.05.

3. Results

3.1 Community complexity and diversity

The proportion of sequences assigned to functions or taxa was the same between lakes and beach environments (**Table 3S.2**), indicating that there was no bias in sample processing or classification between these parameters. Rarefaction analysis was performed for both functional and taxonomic assignments to estimate the coverage obtained from sampling (**Fig. 3.1**). Functional assignments did not plateau and showed greater feature discovery as larger subsets were used. In contrast, taxonomic assignments plateaued and sufficiently captured the bacterial diversity, with the exception of one sample which corresponds to a water sample from Long Beach of Lake Erie. Closer examination of this sample revealed a disproportionate increase in richness from other water communities, although alpha diversity was similar to other samples. The presence of many low abundance taxa resulted in an elevated richness, causing the different rarefaction trend seen in the sample. Removal of low abundance taxa (taxa with less than 10 reads assigned) shifted the rarefaction curve to plateau, and decreased the richness estimate similar to that of other samples, confirming that low abundance taxa were responsible for the anomaly.

Also evident from the rarefaction plots is the increased taxonomic richness in sand samples in comparison to the water samples (**Figs. 3.1A and 3.1B**). This is further verified by comparing the total taxonomic features detected in the two types of samples, revealing an increase from 664 to 974 average taxonomic features (**Fig. 3.1C**, $P < 0.001$).

No discernible patterns were seen between Lake Erie and Lake Ontario in the rarefaction and richness analysis.

Taxonomic diversity as measured by the Shannon index revealed significantly greater taxonomic diversity at the species ($P = 6 .51 \times 10^{-4}$) and the phylum level ($P = 3.93 \times 10^{-4}$) in the sand samples in comparison to the water samples (**Fig. 3.2**). There was no significant difference in the diversity index between lakes or between years from the samples analyzed.

Multivariate dispersion did not reveal significant differences in beta diversity between the two beach environments (sand and water). However, permutational multivariate analysis of variance (permutational MANOVA) shows that the beach environment explains the greatest amount of variance in species composition, more so than compositional variance due to years or lakes ($R^2 = 0.2$, $P < 0.001$) (**Table 3S.3**).

3.2 Taxonomic composition

To further examine the bacterial communities of these freshwater beaches, taxonomic composition was examined. All samples contained a small number of reads assigned to viral, eukaryotic and archaeal taxa that were removed bioinformatically. The majority of the samples show a bacterial proportion of 90% or greater, validating the fraction method of extraction (**Fig. 3S.1**). Overall, there are no significant differences in the superkingdom composition between sand and water samples. At the phylum level, Proteobacteria were the most predominant group at most sites in sands (**Fig. 3.3**). A relatively smaller proportion of Actinobacteria in comparison to the Proteobacteria was detected in most sites. The next predominant phylum in sand was Bacteroidetes, making

up between 5% and 20% of the bacteria in the sand samples. Other predominant phyla in the sand samples include Cyanobacteria, Verrucomicrobia, Firmicutes, and Planctomycetes. A large fraction of the reads of the sand samples was assigned to unclassified phyla, which was not observed in the water samples.

In the aquatic environment, two predominant groups are present; Proteobacteria and Actinobacteria. These two phyla occur in relatively equal proportions in the water samples and make up the majority of the bacterial composition. The water samples also show a large proportion of Bacteroidetes. Ten of the sixteen water samples also show a notable presence of Cyanobacteria and Verrucomicrobia (>2% reads assigned). The less abundant phyla in the water samples occur at very low proportions, in contrast with the sand samples which show more evenness among the less abundant phyla. The diversity at the phylum rank is indeed significantly greater in sand than in water ($P = 3.93 \times 10^{-4}$) which supports the observation that the bacterial communities of these freshwater environments are dominated by fewer phyla in water than those in sand.

3.3 Differences between sand and water microbial communities

All the sand samples contain taxa from a much wider array of phyla than the corresponding water environments. Indeed, unique phyla were detected disproportionately in beach sands, with the most abundant of these being Candidatus Microgenomates, Latescibacteria, and Candidatus Aminicenantes (**Table 3.1**).

Table 3.1 Phyla unique to beach sand environment.

Phylum	#Samples		Total Reads	
	Sand	Water	Sand	Water
Aquificae	14	0	1259	0
Candidate division NC10	14	0	1244	0
Latescibacteria	13	0	1829	0
Candidate division Zixibacteria	11	0	902	0
Candidatus Microgenomates	9	0	2868	0
Candidatus Aminicenantes	9	0	1343	0
Nitrospinae	9	0	737	0
Candidatus Hydrogenedentes	9	0	679	0
Deferribacteres	9	0	585	0

*Unique phyla are defined as phyla whose detection occurs in 8 or more samples of one environment but are not detected in any samples in the other environment.

To identify differences in taxonomic composition between the sand and water environments, differential abundance testing was conducted using DESeq2 (see 2.4. Statistical analysis). At the phylum level, significant enrichment was highest in Actinobacteria in water ($Log_2FC > 2$; **Fig. 3.4**). Additionally, enrichment of Proteobacteria, Bacteroidetes, Cyanobacteria, and Verrucomicrobia was observed in water. The majority of these phyla occur at high abundances, in contrast to the many low abundant but enriched phyla present in sand, supporting the greater diversity seen in sand.

Because of the dominance of Proteobacteria across all samples, there was sufficient sequence information to examine the relative abundance of members of this phylum. Betaproteobacteria, Alphaproteobacteria, and Gammaproteobacteria were the most dominant groups within Proteobacteria across both environments (**Fig. 3S.2**). At the family level, Comamonadaceae belonging to Betaproteobacteria occurred at the highest abundance and were more abundant in water. Other high abundance families include

Flavobacteriaceae, Rhodobacteraceae, Streptococcaceae, and Planctomycetaceae (**Fig. 3.5**).

At the species level, over 100 taxa were differentially abundant between the sand and water environments (**Fig. 3S.3**). The distribution of differentially abundant taxa at the species level follows the same distribution seen at the phylum level. With the exception of Bacteroidetes and Actinobacteria species, differentially-abundant taxa were typically enriched in sand environments. A large number of the Proteobacteria species enriched in sand belonged to the sulfur- and sulfate-reducing Deltaproteobacteria (38 out of 56), and belong to genera with functions associated with the reduction of sulfur-containing compounds (**Table 3S.4**). Species enriched in the water environment mostly belong to the Proteobacteria, Bacteroidetes, or Actinobacteria phyla. The majority of the differentially abundant Actinobacteria species are part of the freshwater SAR11 group of bacteria (**Table 3S.5**). The only Actinobacteria species which was more abundant in the sand environment is the typical soil dwelling *Conexibacter woesei*.

Interestingly, differences in phylum composition did not have as large an effect on the functional capacity as determined by the SEED subsystem classifications (**Fig. 3S.3**). Between samples and environment types, broad level functional capacity (defined at subsystem level 2) did not exhibit as much variation as the taxonomic composition. Functions pertaining to macromolecular processing and metabolism were detected with the greatest abundance across all samples, while more specialized functions relating to element acquisition or metabolism occurred in smaller abundances (**Fig. 3S.4**). This

suggests that despite marked differences in phylum composition within and between environments, general functional capacity is maintained.

In spite of a generally homogenous general functional capacity between the two beach environments, differences were seen in the capacity for more specific functions (**Fig. 3.6**). One interesting disparity between the two beach environments is the enrichment of genes associated with sulfur reduction in sand and sulfur oxidation in water, suggesting a dichotomy in sulfur utilization between the two communities. Differences were also seen in the genetic capacity for spore formation, nitrosative stress, and hydrogenases, among other functions (**Fig. 3.6**).

3.4 Potential human pathogens and fecal indicators in freshwater beaches

Pathogens and fecal indicator bacteria were detected in both water and beach sands. Pathogens and indicator bacteria were selected based on their relevance and potential impact to public health in recreational waters. A total of 34 pathogenic (and indicator) bacterial species were detected in both water and sand environments (**Fig. 3.7**). The most abundant species of those examined was *E. coli* (an important fecal indicator bacteria), which did not show significant differences in abundance between the beach environments. *Pseudomonas mendocina* and *Pseudomonas aeruginosa* were also relatively more abundant in both environments compared to other pathogens, and were significantly elevated in sand. Several low abundance pathogens were also detected from the *Clostridium* genus, with *Clostridium botulinum* occurring at the highest abundance

and exhibiting significant enrichment in water. All *Vibrio* spp. detected were also enriched in water and exhibited asymmetric presence; most sites examined had a presence of *Vibrio* spp. in water but not in sand.

4. Discussion

Beach sands may heavily impact the microbial communities found in recreational waters (Alm et al., 2003;Cui et al., 2013). However, comprehensive profiling of environmental samples has only recently been possible due to advances in sequencing technology and only a few studies have compared the beach microbial communities to the corresponding water environments (Staley and Sadowsky, 2016). Here, we investigated the microbial communities of four freshwater beaches using a shotgun sequencing approach. Our findings provide direct evidence that the beach environment (sand or water) is a major determinant in the bacterial composition of these communities. Our analysis also suggests that shotgun metagenomic sequencing can potentially be used to detect pathogens as evidenced by the detection of sequence homology to key pathogens, not easily detected by other traditional means.

Effect of different beach environments including beach waters and beach sands on bacterial communities of marine beaches were previously examined using 16s rDNA sequencing (Cui et al., 2013). The backshore sand samples investigated in that study (Cui et al., 2013) are comparable to sand pore samples in this study and, our data confirm that, richness and diversity were significantly elevated in backshore sands compared to beach waters. Additionally, ordination revealed a strong clustering of samples originating from backshore sand which was distinct from beach waters. In general, our findings suggest

that the trends between sand and water environments of beaches are common in both marine and freshwater beaches. However, we identified more taxonomic features in both beach water and sand environments, suggesting that a shotgun metagenomic approach may be more sensitive in capturing bacterial diversity than 16S rDNA analysis (or that freshwater environments are more diverse).

Sand pore samples were used as a representative of sand habitats. Although the sand pore method has previously been validated (Whitman et al., 2014), a potential bias may apply when capturing the microbial communities of one environment (the sand) that is dependent on the other (the water). This could potentially lead to underestimation of differences between the two environments due to the interdependency of the sand pore samples on the beach water. This may have impacted our comparative investigation of these two environments. In addition, the inability of the flow of water to mobilize biofilms strongly attached to sand grains could lead to an underestimation of the diversity in sand. Implementing a method for direct release of bacteria from sand would be useful in capturing the additional diversity associated with sand particles.

Major differences were observed between the beach sand and water in terms of taxonomic composition at a broad level. These differences include the decreased diversity in water as Proteobacteria and Actinobacteria exhibited greater abundance relative to sand, while the decreased abundance of the predominant phyla in sand permitted increased abundance of less dominant phyla. Although Proteobacteria, Actinobacteria, and Bacteroidetes were enriched in water, they were the dominant phyla along with the Firmicutes in sand, a finding which supports previous 16S rDNA analysis of beach sand

(Whitman et al., 2014;Staley and Sadowsky, 2016). However, the most abundant families detected in sand of these studies were Rhodobacteraceae, Flavobacteriaceae, Flammeovirgaceae, and Campylobacteraceae. In contrast, we found that Rhodobacteraceae, Flavobacteriaceae, and Comamonadaceae were most abundant. The Comamonadaceae are aerobic motile organisms and therefore it is not surprising that they are abundant in water, although their relatively high abundance in both environments may reflect the transfer of bacterial species between water and sand.

The high levels of Actinobacteria observed in water relative to beach sand was unexpected because Actinobacteria are often considered soil-dwelling bacteria. However, these levels are due to a subpopulation of Actinobacteria species belonging to a group of planktonic organisms prevalent in freshwater environments (**Table 3S.3**). These seem to be related to a highly abundant group of freshwater Actinobacteria present in many freshwater environments (Newton et al., 2011;Ghai et al., 2014).

Unique phyla could be identified in sand that were absent in water (**Table 3.1**). Interestingly, all unique phyla were present only in sand, further providing evidence for an enriched diversity and richness of bacterial communities in sand relative to water. Many of these unique phyla correspond to uncultured microorganisms or poorly characterized phyla grouped by organisms detected only through metagenomic methods, suggesting a plethora of unexplored bacterial constituents present in sand that are absent in water. This corresponds with higher abundance of taxa belonging to unclassified phyla in sand as well (**Fig. 3.3**). Candidatus Microgenomates (previously called OP11), found only in sand in our analysis, is an extremely diverse phylum comprised of species with a

widespread environmental distribution and they are often found in methanogenic environments (Hu et al., 2016). The Aquificae, also unique to sand, naturally occur in thermophilic environments with optimal growth temperatures at 65°C or higher (Spear et al., 2005). Although aerobic, they can employ anaerobic respiration through the use of thiosulfate or sulfur and are often found in sulfur pools (Hou et al., 2013). Candidatus Aminicenantes is typically abundant in aquatic non-marine and hydrocarbon-impacted environments, and are also found in anoxic environments more than hypoxic or oxic environments (Farag et al., 2014). The candidate division NC10 are a known aquatic group of organisms initially detected through 16S rDNA sequencing and includes an anaerobic methane oxidizer (Ettwig et al., 2010). The Latescibacteria are another group which can be found in anoxic conditions and sediment environments (Youssef et al., 2015). These phyla have important environmental impacts and despite the proximity of the two beach environments investigated in our study, they were found only in sand. Additionally, many of these unique phyla have not previously been detected in sand habitats, and may possess niche functions that result in specialized adaptation to the sand environment.

Many of the Proteobacteria species enriched in sand belonged to the sulfur- and sulfate-reducing Deltaproteobacteria (38 out of 56) and belong to genera that reduce sulfur-containing compounds (**Table 3.S3**). The remaining species that were found to be enriched in sand belong to uncharacterized bacterium with non-informative identifiers (e.g. bacterium UASB14 and UASB270), identified previously in anaerobic sludge blankets through 16S rDNA sequencing (Sekiguchi et al., 2015).

Metagenomic analysis can be used to assess the functional capacities and enrichment of functions by examining the abundance of genes with relation to categorized functions. This type of information can reveal what the communities have adapted to and what functions are more prevalent. Although overall functional profiles were almost similar across beach environments (**Fig. 3S.3**), there was differential abundance in genes relating to nutrient metabolism, such as nitrogen and sulfur, as well as nutrient stress (**Fig. 3.6**). Importantly, functions such as spore coat formation, acid stress response regulators, and sigmaB stress regulation genes are enriched in sand. The enrichment of flavohaemoglobin (nitric oxide dioxygenase) and nitrosative stress genes in addition to nitrite reductase and nitrate/nitrite ammonification genes also suggests the abundance of nitrogen-containing compounds. The enrichment of sulfate reduction-associated complexes in sand also coincides with the enrichment of anaerobic Deltaproteobacteria, and other sulfur- and sulfate-reducing bacteria. Enrichment of cytochrome c oxidase biogenesis genes in water communities also further provides support for the greater capacity of aerobic respiration in water, whereas anaerobic respiration likely dominates in sand.

An important goal of this work was to identify potential pathogens and FIB within freshwater beach environments. Many sequences were identified as originating from pathogens and FIBs (**Fig. 3.7**). While several pathogens and FIB such as *Aeromonas* spp., *E. coli*, *Salmonella enterica*, and *Pseudomonas aeruginosa* were detected in almost all samples, others were not detected at all (e.g. many *Campylobacter* spp.). Many of the *Vibrio* species examined were detected in both beach water and sand or just the beach

water, but rarely detected only in and at any given site. This was also the case for *Staphylococcus aureus*. This information is important for understanding disparities in pathogen and FIB presence at recreational beaches and can help shape and optimize water quality monitoring processes.

In this study, we partnered with municipal water monitoring authorities and investigated the potential use of shotgun metagenomic sequencing in assessing water quality. The successful detection of pathogens using DNA-based methods in this study suggests that this approach may be used to augment traditional means of water monitoring. Traditional methods have many limitations, such as a limited profile of pathogens and FIBs that can feasibly be monitored. The approach implemented in this study may be a useful tool in augmenting traditional monitoring programs. However, standardization of sampling, sample processing and sequencing must be developed and correlated with traditional methods before metagenomic methods can be implemented to monitor pathogen loads.

5. Conclusions

1. Beach environment (sand or water) is a major determinant in shaping beach bacterial communities. At the phylum level, Proteobacteria, Actinobacteria, and Bacteroides are the most predominant groups in both water and sand. However, sand samples exhibit greater taxonomic diversity at the species level in comparison to water.
2. Bacterial phyla corresponding to uncultured microorganisms or poorly characterized phyla including the Aquificae, Candidatus Microgenomates, and Candidatus Aminicenantes are present only in sand. Many of these bacteria are commonly found in methanogenic and hydrocarbon impacted environments

suggesting that sand supports anaerobic respiration (Aquificae, although aerobic, can employ anaerobic respiration through the use oxidation of reduced sulfur compounds).

3. Pathogens and fecal indicator bacteria including *E. coli*, *Pseudomonas* spp., *Aeromonas* spp., *Legionella* spp., and *Bacteroides* spp. are present in both water and sand indicating that metagenomic analyses can be used to detect pathogens and trace fecal contamination of recreational beaches.

4. Enrichment of anaerobic metabolism associated pathways such as sulfate and sulfur-reduction complexes validate the increased abundance of anaerobic Deltaproteobacteria and other sulfur-reducing bacteria in sand. Enrichment of cytochrome c oxidase biogenesis genes in water bacterial communities may indicate greater capacity of aerobic respiration in water.

6. Acknowledgement

We thank members of the Schellhorn Lab and Golding Lab for their help and comments on the manuscript. We gratefully acknowledge financial support through the Niagara Region WaterSmart Program, the Natural Sciences and Research Council (NSERC) of Canada's Collaborative Research and Development Program (CRDP), and NSERC operating grants to HES and to GBG. The authors declare no conflict of interest.

7. Author contributions

M.M.M., and H.E.S. designed research; M.M.M. performed sampling and laboratory experiments; M.M.M., and Y.S. analyzed data; M.M.M., Y.S., H.E.S., and G.B.G. wrote the paper; H.E.S., and G.B.G. supervised research.

References

Alm, E.W., Burke, J., and Spain, A. (2003). Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water Res* 37**,** 3978-3982.

Apha (2012). *Standard methods for the examination of water and wastewater.* American Public Health Association, Washington D.C.

Boehm, A.B., Yamahara, K.M., and Sassoubre, L.M. (2014). Diversity and transport of microorganisms in intertidal sands of the California coast. *Appl Environ Microbiol* 80**,** 3943-3951.

Bouzat, J.L., Hoostal, M.J., and Looft, T. (2013). Spatial patterns of bacterial community composition within Lake Erie sediments. *Journal of Great Lakes Research* 39**,** 344-351.

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

Byappanahalli, M.N., Roll, B.M., and Fujioka, R.S. (2012). Evidence for occurrence, persistence, and growth potential of *Escherichia coli* and enterococci in Hawaii's soil environments. *Microbes Environ* 27**,** 164-170.

Costello, E.K.Lauber, C.L.Hamady, M.Fierer, N.Gordon, J.I.Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science* 326**,** 1694-1697.

Cui, H., Yang, K., Pagaling, E., and Yan, T. (2013). Spatial and temporal variation in enterococcal abundance and its relationship to the microbial community in Hawaii beach sand and water. *Appl Environ Microbiol* 79**,** 3601-3609.

Edge, T.A., and Hill, S. (2007). Multiple lines of evidence to identify the sources of fecal pollution at a freshwater beach in Hamilton Harbour, Lake Ontario. *Water Res* 41**,** 3585-3594.

Ettwig, K.F.Butler, M.K.Le Paslier, D.Pelletier, E.Mangenot, S.Kuypers, M.M., et al. (2010). Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. *Nature* 464**,** 543-548.

Farag, I.F., Davis, J.P., Youssef, N.H., and Elshahed, M.S. (2014). Global patterns of abundance, diversity and community structure of the Aminicenantes (candidate phylum OP8). *PLoS One* 9**,** e92139.

Ge, Z., Nevers, M.B., Schwab, D.J., and Whitman, R.L. (2010). Coastal loading and transport of *Escherichia coli* at an embayed beach in Lake Michigan. *Environ Sci Technol* 44**,** 6731-6737.

Ghai, R., Mizuno, C.M., Picazo, A., Camacho, A., and Rodriguez-Valera, F. (2014). Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Mol Ecol* 23**,** 6073-6090.

Hou, W.Wang, S.Dong, H.Jiang, H.Briggs, B.R.Peacock, J.P., et al. (2013). A comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing. *PLoS One* 8**,** e53350.

Hu, P.Tom, L.Singh, A.Thomas, B.C.Baker, B.J.Piceno, Y.M., et al. (2016). Genome-Resolved Metagenomic Analysis Reveals Roles for Candidate Phyla and Other

Microbial Community Members in Biogeochemical Transformations in Oil Reservoirs. *MBio* 7**,** e01669-01615.

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

Khan, I.U., Hill, S., Nowak, E., and Edge, T.A. (2013). Effect of incubation temperature on the detection of thermophilic campylobacter species from freshwater beaches, nearby wastewater effluents, and bird fecal droppings. *Appl Environ Microbiol* 79**,** 7639-7645.

Khan, I.U., Loughborough, A., and Edge, T.A. (2009). DNA-based real-time detection and quantification of aeromonads from fresh water beaches on Lake Ontario. *J Water Health* 7**,** 312-323.

Kuramae, E.E.Yergeau, E.Wong, L.C.Pijl, A.S.Van Veen, J.A.Kowalchuk, G.A. (2012). Soil characteristics more strongly influence soil bacterial communities than land-use type. *FEMS Microbiol Ecol* 79**,** 12-24.

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

Mohiuddin, M., and Schellhorn, H.E. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* 6**,** 960.

Newman, D.K., and Banfield, J.F. (2002). Geomicrobiology: how molecular-scale interactions underpin biogeochemical systems. *Science* 296**,** 1071-1077.

Newton, R.J., Jones, S.E., Eiler, A., Mcmahon, K.D., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* 75**,** 14-49.

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.Olson, R.Pusch, G.D.Olsen, G.J.Davis, J.J.Disz, T., et al. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res* 42**,** D206-214.

Paulson, J.N., Stine, O.C., Bravo, H.C., and Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 10**,** 1200-1202.

Piggot, A.M., Klaus, J.S., Johnson, S., Phillips, M.C., and Solo-Gabriele, H.M. (2012). Relationship between enterococcal levels and sediment biofilms at recreational beaches in South Florida. *Appl Environ Microbiol* 78**,** 5973-5982.

Poretsky, R., Rodriguez, R.L., 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.

Ranjan, R., Rani, A., Metwally, A., Mcgee, H.S., and Perkins, D.L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun* 469**,** 967-977.

Russell, T.L., Yamahara, K.M., and Boehm, A.B. (2012). Mobilization and transport of naturally occurring enterococci in beach sands subject to transient infiltration of seawater. *Environ Sci Technol* 46**,** 5988-5996.

Sayers, E.W.Barrett, T.Benson, D.A.Bryant, S.H.Canese, K.Chetvernin, V., et al. (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 37**,** D5-15.

Sekiguchi, Y.Ohashi, A.Parks, D.H.Yamauchi, T.Tyson, G.W.Hugenholtz, P. (2015). First genomic insights into members of a candidate bacterial phylum responsible for wastewater bulking. *PeerJ* 3**,** e740.

Shah, A.H.Abdelzaher, A.M.Phillips, M.Hernandez, R.Solo-Gabriele, H.M.Kish, J., et al. (2011). Indicator microbes correlate with pathogenic bacteria, yeasts and helminthes in sand at a subtropical recreational beach site. *J Appl Microbiol* 110**,** 1571-1583.

Spear, J.R., Walker, J.J., Mccollom, T.M., and Pace, N.R. (2005). Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem. *Proc Natl Acad Sci U S A* 102**,** 2555-2560.

Staley, C., and Sadowsky, M.J. (2016). Regional Similarities and Consistent Patterns of Local Variation in Beach Sand Bacterial Communities throughout the Northern Hemisphere. *Appl Environ Microbiol* 82**,** 2751-2762.

Sunagawa, S.Coelho, L.P.Chaffron, S.Kultima, J.R.Labadie, K.Salazar, G., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348**,** 1261359.

Tamames, J., Abellan, J.J., Pignatelli, M., Camacho, A., and Moya, A. (2010). Environmental distribution of prokaryotic taxa. *BMC Microbiol* 10**,** 85.

Vijayavel, K., Fujioka, R., Ebdon, J., and Taylor, H. (2010). Isolation and characterization of Bacteroides host strain HB-73 used to detect sewage specific phages in Hawaii. *Water Res* 44**,** 3714-3724.

Wang, Y.Sheng, H.F.He, Y.Wu, J.Y.Jiang, Y.X.Tam, N.F., et al. (2012). Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. *Appl Environ Microbiol* 78**,** 8264-8271.

Whitman, R.Harwood, V.J.Edge, T.A.Nevers, M.Byappanahalli, M.Vijayavel, K., et al. (2014). Microbes in Beach Sands: Integrating Environment, Ecology and Public Health. *Rev Environ Sci Biotechnol* 13**,** 329-368.

Wright, M.E., Solo-Gabriele, H.M., Elmir, S., and Fleming, L.E. (2009). Microbial load from animal feces at a recreational beach. *Mar Pollut Bull* 58**,** 1649-1656.

Yamahara, K.M., Sassoubre, L.M., Goodwin, K.D., and Boehm, A.B. (2012). Occurrence and persistence of bacterial pathogens and indicator organisms in beach sand along the California coast. *Appl Environ Microbiol* 78**,** 1733-1745.

Youssef, N.H.Farag, I.F.Rinke, C.Hallam, S.J.Woyke, T.Elshahed, M.S. (2015). In Silico Analysis of the Metabolic Potential and Niche Specialization of Candidate Phylum "Latescibacteria" (WS3). *PLoS One* 10**,** e0127499.

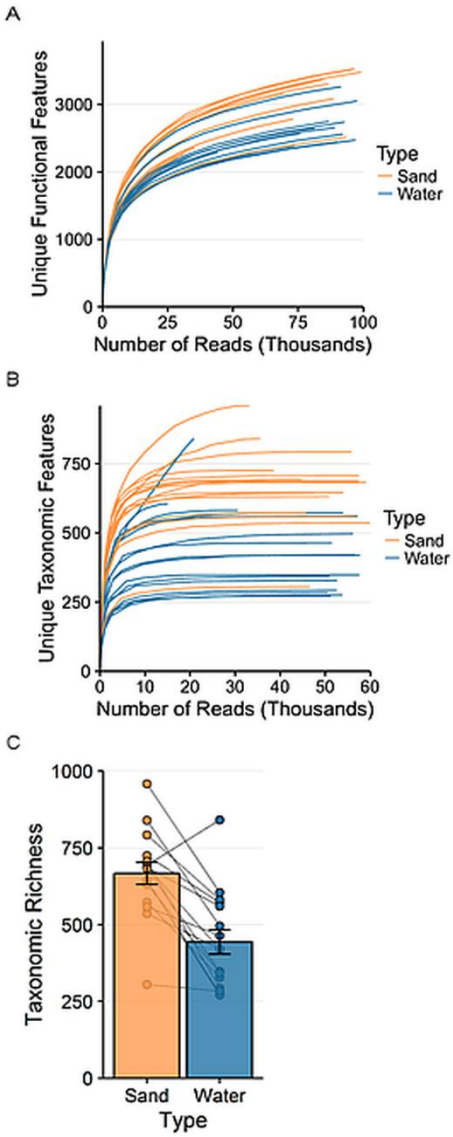

Figure 3.1 Rarefaction and richness analysis of sequence data. Rarefaction analysis was performed on all reads with both functional (A) and taxonomic assignments (B). Each subset used for rarefaction was repeated 10 times. Sand and water samples are distinguished by color. Unique features refer to distinct and non-redundant taxonomic or functional features. (A) Functional assignments did not saturate whereas (B) taxonomic assignments saturated for all but one of the samples. (C) Average taxonomic features showing pore samples are more taxonomically rich than water samples. Standard error of the mean is indicated by the error bar.

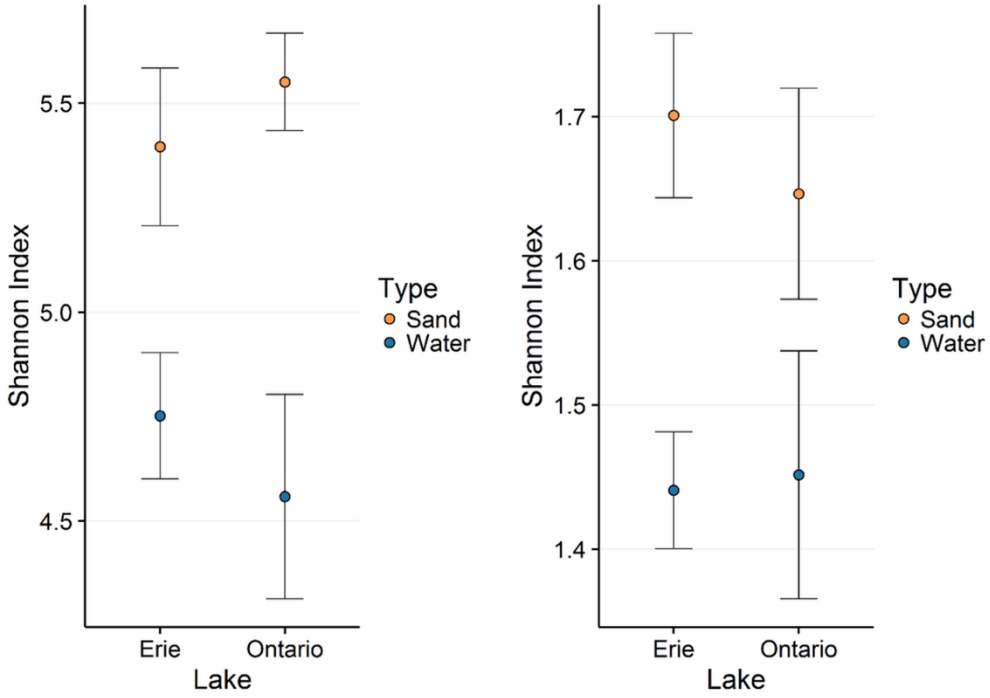

Figure 3.2 Shannon diversity of samples from both Lake Ontario and Lake Erie. (A) Species level Shannon index and (B) phylum level Shannon index. Standard errors of the mean are indicated by the error bars. Shannon diversity is enriched in sand environments.

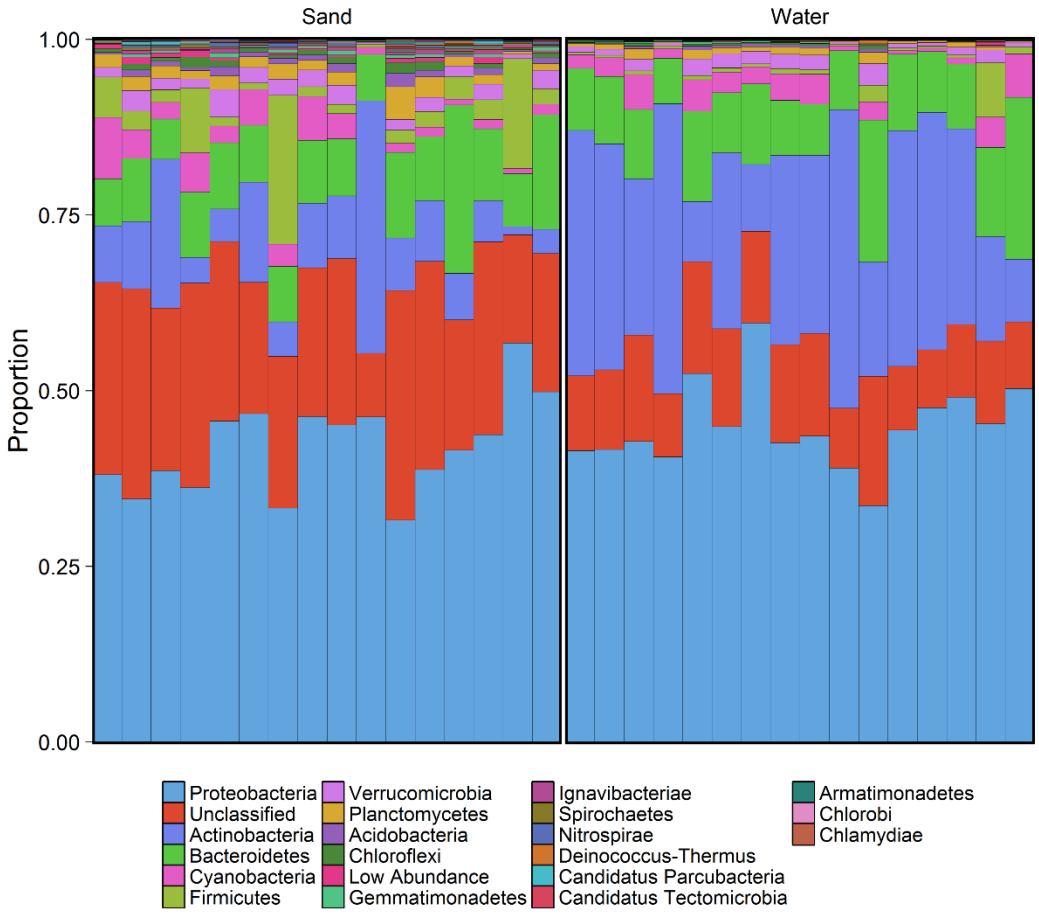

Figure 3.3 Taxonomic composition at the phylum level. Reads assigned to bacterial taxa were aggregated to the phylum level and plotted as proportions. Reads assigned to bacterial taxa whose phylum classification was unknown are aggregated and labelled as unclassified. Sites are distinguished by environment.

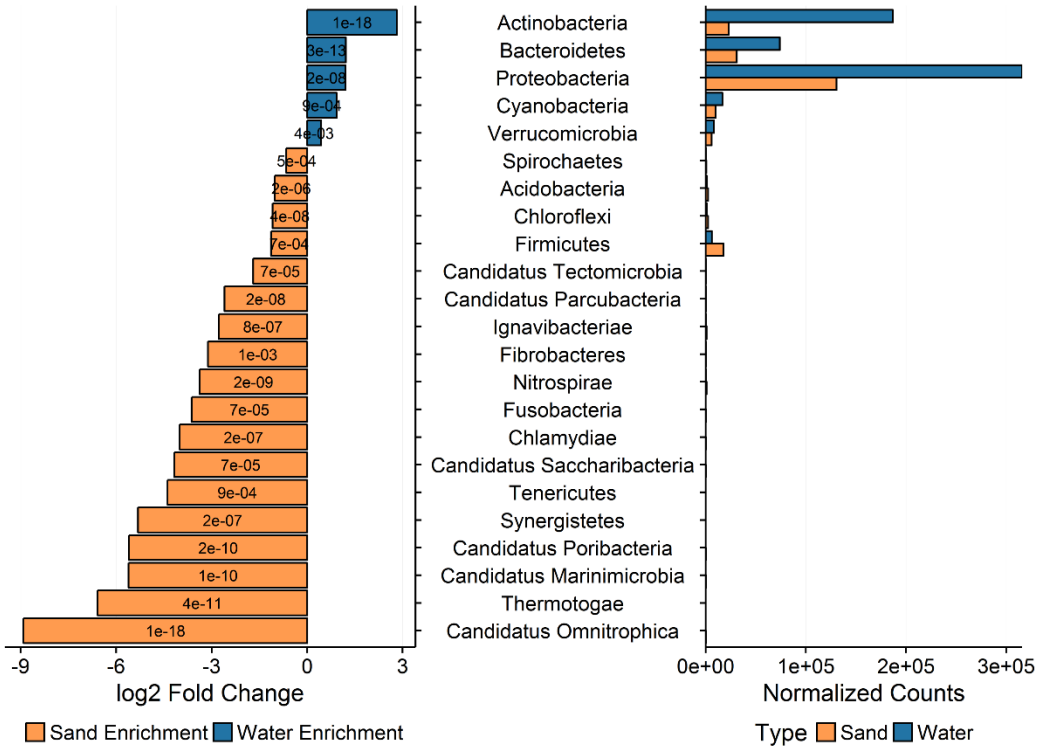

Figure 3.4 Differentially abundant phyla in sand and water environments. Numerous phyla are differentially abundant between beach environments. Differential abundance of phyla was determined by DESeq2 using a paired model. log2 fold changes are indicated on the left with the associated adjusted p-values of significance (cutoff of 0.01). The normalized counts for each phylum in each environment are plotted on the right.

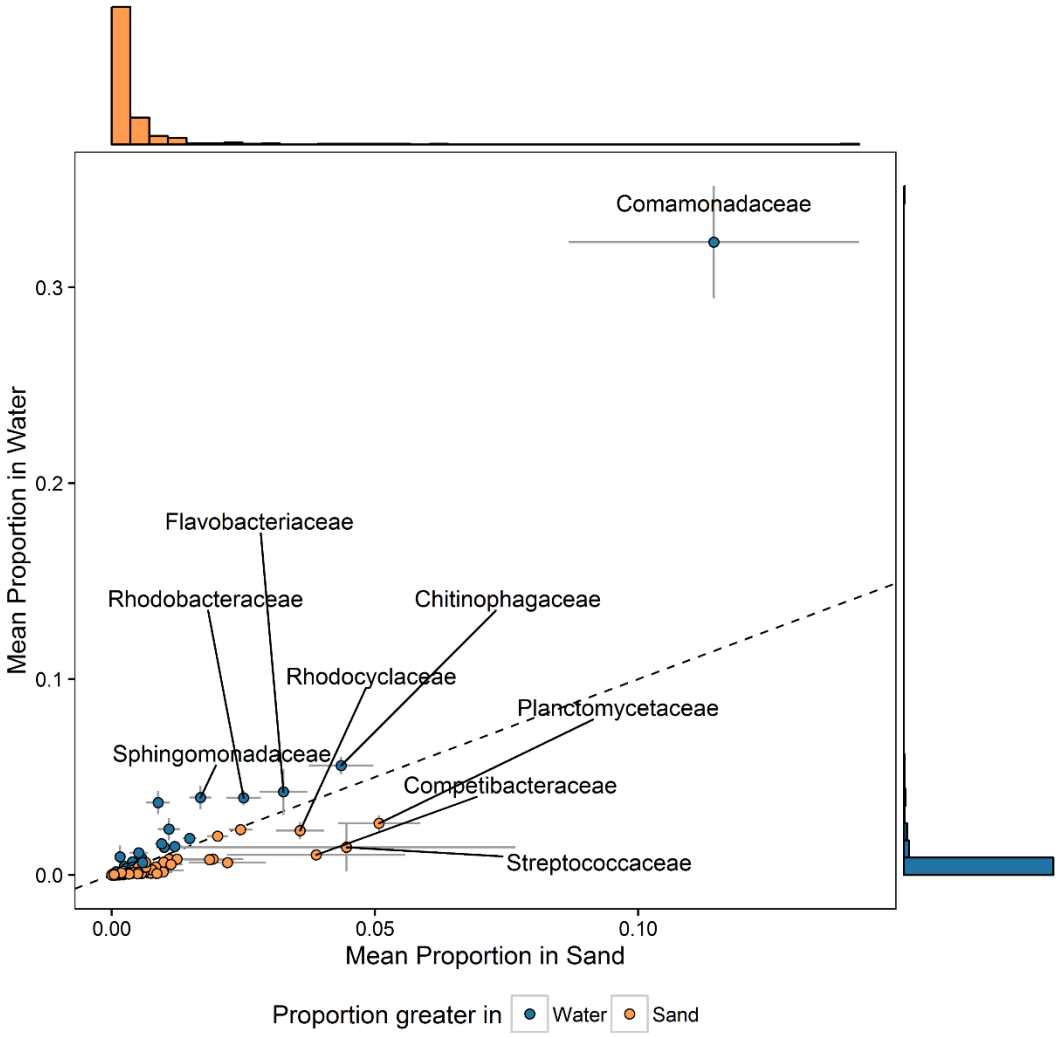

Figure 3.5 Mean proportions of bacterial taxa at the family level in beach sand and water. The mean proportion of families in both sand and water are plotted against each other with standard error of the mean proportion indicated for each environment. The marginal histogram describes the frequency of phyla at the given proportions. Families labelled as unclassified were removed prior to determining mean proportions and are not plotted.

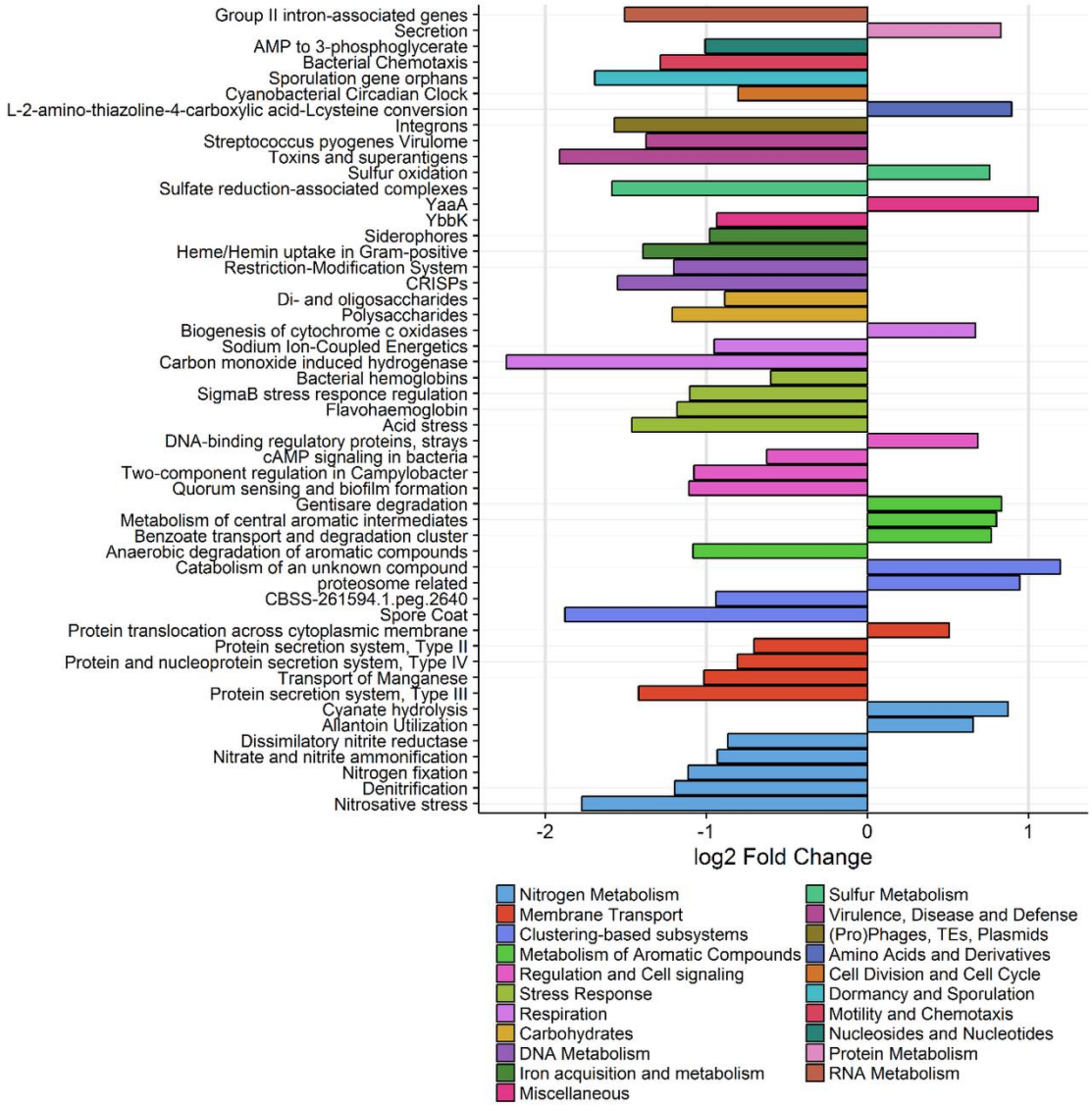

Figure 3.6 Functions exhibiting differential abundance between beach environments. Differential abundance of functions was determined by DESeq2 using a paired model. log2 fold changes are indicated, with negative fold changes signifying enrichment in sand and positive fold changes signifying enrichment in water. Functions depicted are in level three of the SEED subsystem hierarchy and are colored by their broader level two categorization and significance was determined at an adjusted *p*-value of 0.01.

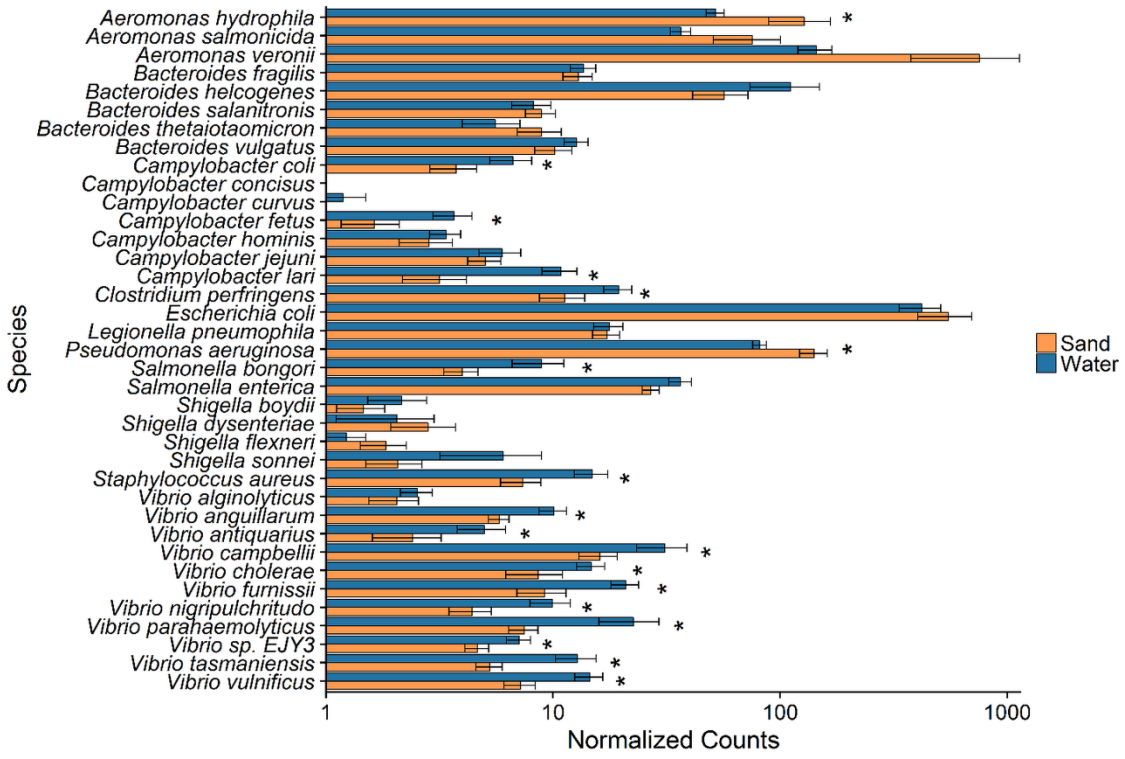

Figure 3.7 Many pathogens exhibit differential abundance between beach environments. Reads were classified using CLARK as described in the methods. Counts were normalized using DESeq2's size factor estimations and mean normalized counts for each environment were plotted on a $\log_{10}$ scale with the error bars indicating standard error of the mean. Significance between environments was determined at an adjusted *p*-value of 0.05 using DESeq2's negative binomial model testing and signified by the asterisks. A confidence score determined by CLARK of 95% was used as a threshold for classification.

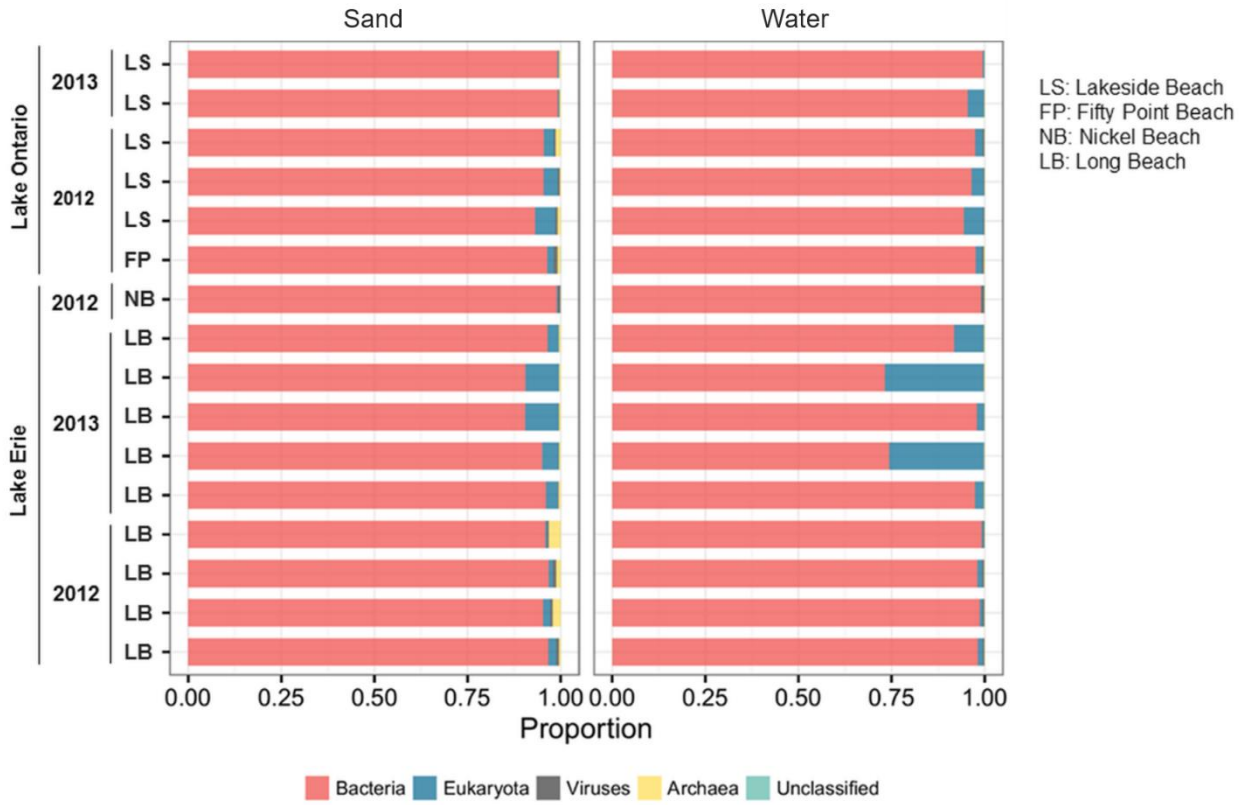

Figure 3S.1 Taxonomic composition at the superkingdom level. The proportion of reads assigned to each superkingdom is indicated. Reads belonging to any node below the superkingdom nodes were counted as belonging to that superkingdom. The proportions were calculated by dividing the number of reads belonging to a superkingdom by all reads assigned to any node in the NCBI reference tree. The "Unclassified" entry refers to sequences which have been entered in the NCBI database but have not been assigned to a superkingdom (e.g. environmental sequences).

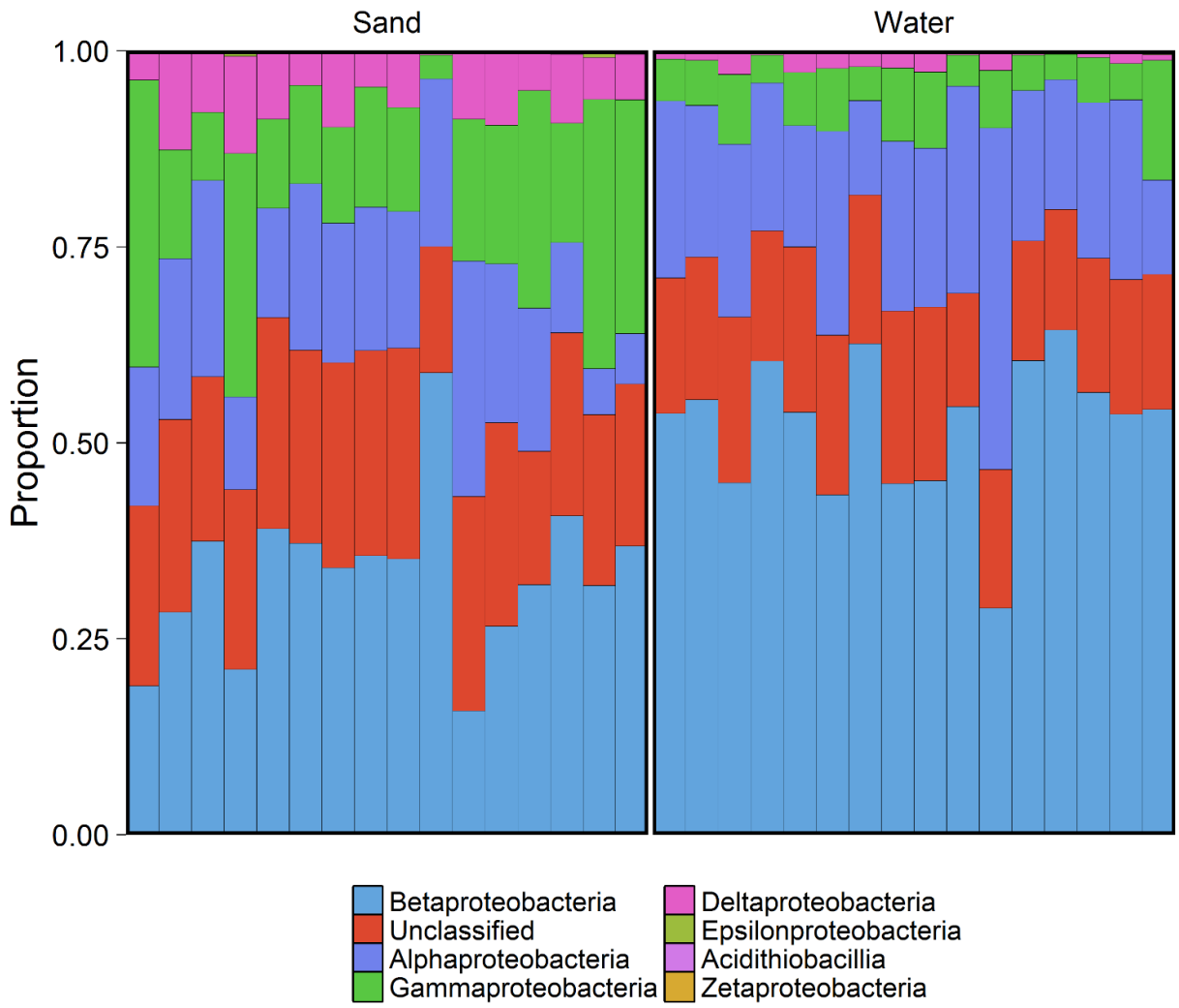

Figure 3S.2 Taxonomic composition of the proteobacteria at the class rank. Proportions were determined by the number of reads assigned to each proteobacterial class divided by the total number of proteobacterial hits. The "Unclassified" label refers to reads whose assignments occurred to the proteobacterial phylum and have no classification at the class level.

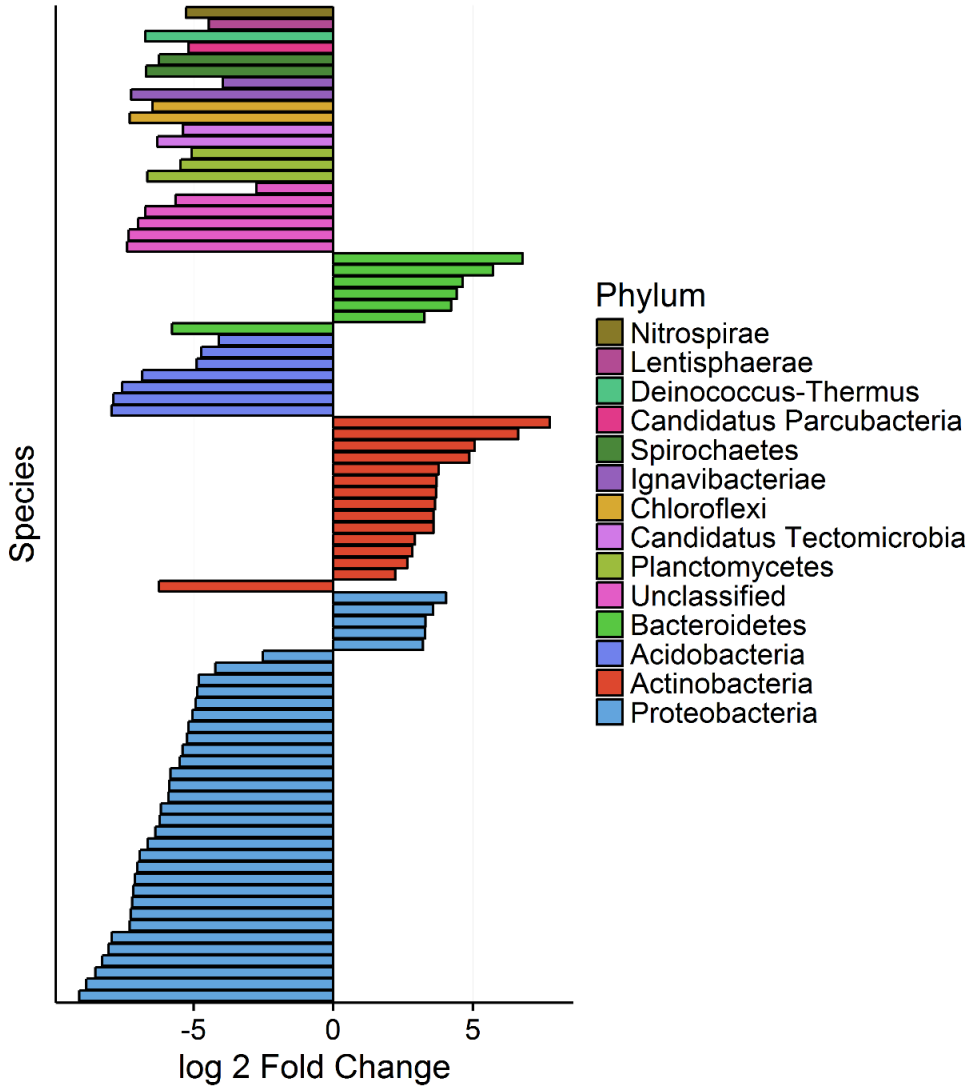

Figure 3S.3 Differentially abundant species between beach environments. Differential abundance of species was determined by DESeq2 using a paired model. log2 fold changes are indicated, with negative l2FC signifying enrichment in the sand and positive l2FC signifying enrichment in the water. Species are coloured according to their phylum classification and significance was determined at an adjusted *p*-value of 0.05.

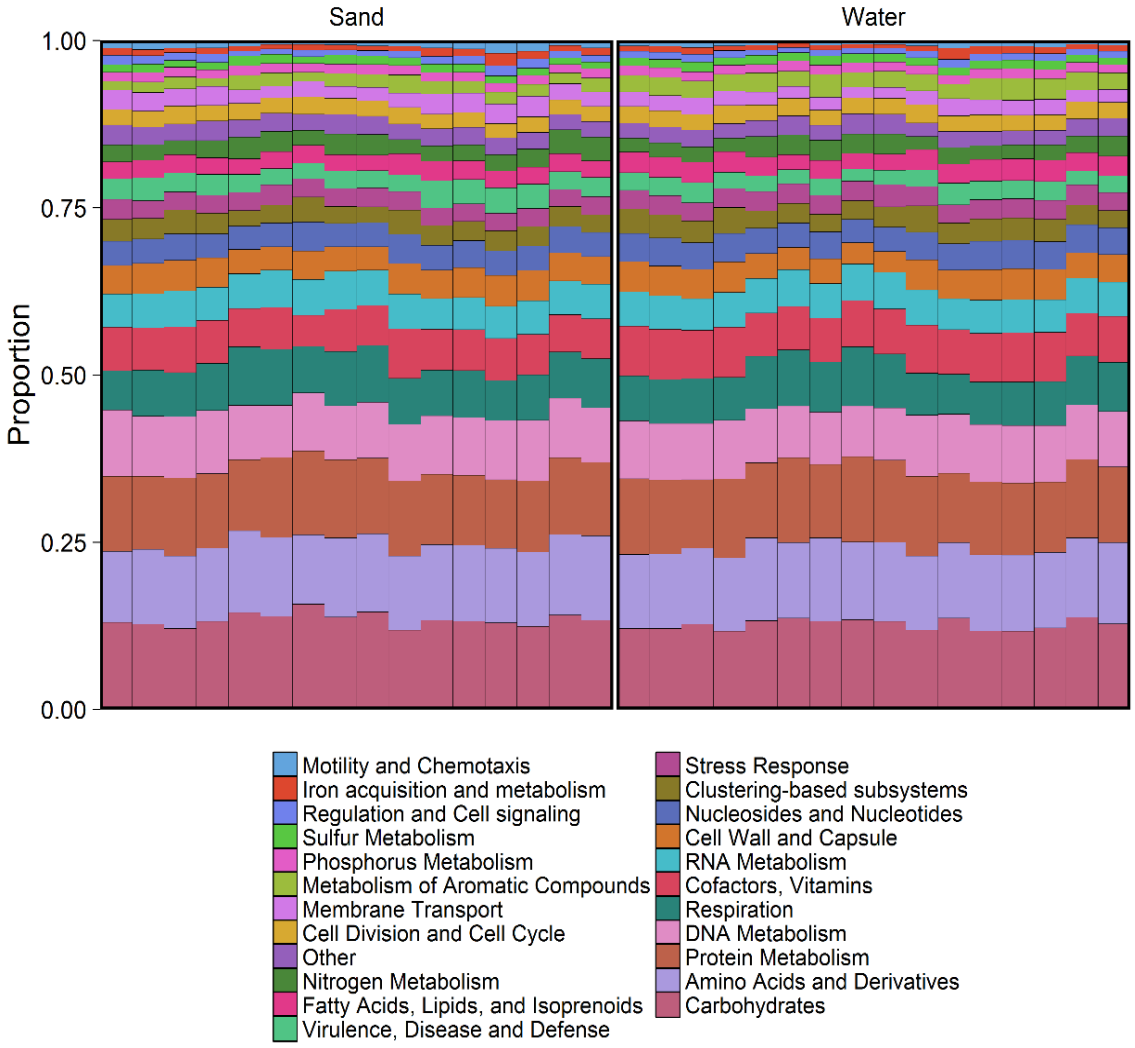

Figure 3S.4 Functional capacity is stable between samples and environments. Functional classification to the SEED subsystem hierarchy aggregated to level two is plotted as proportions of the number of reads with a functional assignment for each sample. Samples corresponding to sand and water sites are separated.

Chapter 4: Temporal and spatial changes in bacterial diversity in mixed use watersheds of the Great Lakes region

Mahi M. Mohiuddin, Steven R. Botts, Athanasios Paschos, Herb E. Schellhorn*
Department of Biology, McMaster University, Hamilton, ON, Canada

***Correspondence:** Herb E. Schellhorn, LSB 433, Department of Biology, McMaster University, 1280 Main St W, Hamilton, ON, L8S 4K1, Canada.
E-mail: schell@mcmaster.ca

Reproduced with permission from Mohiuddin, M.M., S. R. Botts, A. Paschos & H.E. Schellhorn, (2019) Temporal and spatial changes in bacterial diversity in mixed-use watersheds of the Great Lakes region. *Journal of Great Lakes Research* 45(1): 109-118. Elsevier B.V.

Abstract

Environmental water monitoring is an important responsibility of municipal governments. In this study, we partnered with several municipalities in an extensive sampling program to investigate the effects of spatiotemporal and environmental factors on bacterial diversity in a complex watershed ecosystem containing specific environments including creeks, a river, canals, stormwater outfalls and freshwater lakes of the Niagara Peninsula. Samples were collected using standard municipal protocols and bacterial DNA extracted from these samples was sequenced using high-throughput DNA sequencing targeting the V3-V4 region of the 16S rRNA gene. Average taxonomic richness and alpha diversity differed significantly between samples collected from lakes and creeks ($P < 0.05$), and between lakes and stormwater outfalls ($P < 0.05$). Beta diversity also differed significantly ($P < 0.0001$) between habitats suggesting that each of these habitats harbours distinct bacterial groups. Among the environmental factors examined, dissolved oxygen (DO) level was strongly associated ($P < 0.001$) with bacterial diversity. Using a Bayesian source tracking method, the proportional contribution of creeks, river, canals and stormwater outfall habitats in shaping lake bacterial community structure was quantified. Sequences associated with genera known to contain pathogens as well as fecal indicator bacteria were found in every habitat. This study demonstrates that DNA sequence analysis can augment traditional methods of watershed monitoring and management by providing additional information on bacteria of interest to water quality policy makers. Future work may integrate taxonomic and functional analyses to obtain a

greater understanding of pathogen survival, nutrient cycling and microbial interactions in freshwater ecosystems.

1. Introduction

Many urban watersheds are routinely monitored by municipal agencies using standardized sampling programs that include tests for fecal contaminants, heavy metals, BOD (biological oxygen demand) and COD (chemical oxygen demand). Considerable additional and comprehensive information can potentially be obtained using DNA sequencing technologies. These include source tracking of indicator bacteria (Staley et al., 2018), identifying pathogenic bacteria and viruses (Weidhaas et al., 2018) and examining the diversity of bacterial communities (Qu et al., 2017). Such information cannot be readily determined using traditional culture-based assays employed in municipal water quality monitoring programs. DNA analyses may therefore be used to augment traditional water monitoring programs by providing additional important information of interest to municipal authorities.

It is well established that freshwater microbial communities are enriched in bacteria belonging to the phyla Proteobacteria (mainly of the classes Alpha, Beta and Gammaproteobacteria), Actinobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia and Firmicutes (Tamames et al., 2010;Lee et al., 2016;Horton et al., 2018;Olapade, 2018). However, relative abundance of these groups is dependent on the type of freshwater habitat, with surface water and ground water enriched in class Betaproteobacteria and phylum Bacteroidetes (Read et al., 2015;Braun et al., 2016). Firmicutes, on the other hand, are not abundant in surface and ground water but are fairly common in wastewater

(Tamames et al., 2010). Additional factors including chemical and nutrient inputs associated with different environments can also shape bacterial community structure. For example, Verrucomicrobia is enriched in potassium-rich habitats and Acidobacteria in organic carbon contaminated environments (Staley et al., 2014;Newton and McLellan, 2015). Therefore, the presence of specific bacterial groups in a particular habitat may provide useful information about physicochemical characteristics of aquatic environments, which may be important to public health officials monitoring water quality.

Watersheds are complex ecosystems composed of a mixture of linked freshwater habitats including lakes, rivers, canals, stormwater outfalls, wetlands and forests. These linked habitats provide water for consumption, household, agricultural and industrial use, recreation and routes for transportation. Watersheds of Niagara Peninsula consist of complex networks of lands drained by rivers, canals, stormwater outfalls, creeks and lakes. These lakes include Lake Ontario and Lake Erie, two of the largest freshwater reservoirs in North America. Many of these watersheds contain high levels of fecal indicator bacteria such as *Escherichia coli* (NPCA, 2014). Using both microbial source tracking (MST) and high-throughput sequencing (HTS) approaches, pathogenic bacteria including *Campylobacter* spp., *Pseudomonas* spp., *Clostridium* spp., *Staphylococcus* spp. (Edge and Hill, 2007;Fong et al., 2007;Khan et al., 2014;Berry et al., 2017;Mohiuddin et al., 2017), and viruses were identified in these habitats (Mohiuddin and Schellhorn, 2015). While single marker MST-based approaches can be used to identify sources of select pathogens in lake environments (Edge and Hill, 2007;Fong et al., 2007),

environmental determinants that shape the lake microbial community structure are not clear. In addition, lakes receive inputs from many sources including creeks, river, canals and stormwater outfalls and, therefore, contribution from source habitats into lake habitat may influence lake microbial community structure. Therefore, identifying the contribution of each of these source habitats in shaping lake microbial community structure may provide a better understanding of bacterial community structure and the transport of microbes in inter-connected watersheds.

In this study, we partnered with a large municipal sampling program. We chose the Niagara Peninsula as a test region as this 10,000 km^2 area includes multiple habitats within agricultural and urban watershed. To assess the diversity and to provide a robust and comprehensive analysis of major bacterial groups, microbial communities were characterized using next-generation sequencing approach targeting the conserved 16S rRNA gene. We employed a Bayesian approach to identify the inner link between inter-connected watersheds. We also assessed the effects of environmental parameters on the distribution of microbial communities and identified habitat-specific bacterial species in each habitat.

2. Materials and methods

2.1 Sample collection and processing

Water samples were collected from twenty-five creeks, two canals, one river, six stormwater outfalls and two lakes of Niagara Peninsula during the 2014 sampling season (Fig. 4S.1). For creeks, the river, canals and stormwater outfalls, samples were collected

weekly over a five-month period (July to November). For the lakes, samples were collected from Lakeside Beach of Lake Ontario and Long Beach of Lake Erie over a three-month period (June to August). A total of 414 water samples were collected during the sampling period. Surface water samples were collected at 1.0 m depth in sterile 1.0 L sampling containers. After collection, the samples were kept on ice, transported to the lab, and processed within 6 h as previously described (APHA, 2012;Mohiuddin et al., 2017). Physicochemical parameters examined were either measured on-site using calibrated YSI equipment (YSI Inc., Yellow Springs, Ohio, USA) or were supplied as metadata by Ministry of the Environment and Climate Change (Canada) laboratory from their standard sampling program (https://www.ontario.ca/data/provincial-stream-water-quality-monitoring-network). Physicochemical parameters including temperature, specific conductivity, total dissolved solids (TDS), nitrite plus nitrate (NOx), total phosphorous (TP), and dissolved oxygen (DO) were measured for all samples (Table 4.1) except for lake for which TP and NOx measurements were unavailable. For DNA extraction, water samples were passed through a 0.45-µm-pore-size 47-mm-diameter sterile mixed cellulose ester membrane filter (Fisher Scientific, Ottawa, ON, Canada). The filters were then cut into fragments (1 cm^2 size) with sterile scissors and the fragments were aseptically transferred with sterile forceps into PowerBead tubes (MO BIO, Carlsbad, CA, USA) for DNA extraction. DNA was extracted on the same day using PowerSoil DNA Isolation Kit (MO BIO) according to manufacturer's instruction and purified DNA was stored at -20^{o}C until further analysis.

2.2 Library preparation for 16S rRNA gene amplicons and sequencing

Bacterial diversities were investigated by targeting the V3-V4 hypervariable region of the 16S rRNA gene. Amplicons were generated using the optimized primer pair (S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21) evaluated elsewhere (Klindworth et al., 2013). PCR amplicons from all samples were pooled in equal mass amounts and sent to the Farncombe Metagenomics Facility at McMaster University for further processing. BioAnalyzer and High Sensitivity DNA Kit (Agilent) were used to check the product size and the presence of primer dimers in each library. Finally, library yields were quantified by qPCR using primers complementary to the Illumina adaptors. Sequencing was performed on the Illumina MiSeq using v3 chemistry which generated 2x300 bp paired-end sequence reads.

2.3 Bioinformatic analyses

Sequences generated by Illumina MiSeq were de-multiplexed and quality filtered using established criteria (Bokulich et al., 2013). Briefly, sequences were quality filtered using a minimum quality score (Q score) of 25 over at least 75% of the sequence read. Sequences containing more than 10 consecutive low-quality base pairs, ambiguous bases, errors in barcode sequences, and more than 2 nt mismatches from the primer sequences were removed. The quality-filtered sequences were then analyzed using QIIME version 1.9.0 with default parameters (Caporaso et al., 2010). Sequences were aligned with the Greengenes reference database (v13_8) (McDonald et al., 2012) with a sequence similarity threshold of 97% and a minimum query alignment length of 50% using UCLUST (Edgar, 2010). A biological observation matrix (BIOM) summary table was generated from the resulting operational taxonomic unit (OTU) table which provides OTU count/sample statistics. A relatively high sampling depth was employed (25,000 OTU counts/sample) for OTU table rarefaction and subsequent bacterial diversity (alpha and beta diversity) analyses.

2.4 Statistical analyses

Statistical analysis was conducted in R (version 3.3.3) (www.R-project.org). The VEGAN package (version 2.4-2) within R was used to measure ecological diversity and testing of multivariate variance partitioning with permutational multivariate analysis of variance (PERMANOVA). Parametric and non-parametric diversity were determined using Shannon index and Abundance-based Coverage Estimates (ACE index) respectively. Analysis of Similarity (ANOSIM) was used to compare beta diversity. To

test multivariate homogeneity of group dispersions (variances) we performed PERMDISP (Anderson, 2006;Anderson et al., 2006) alongside ANOSIM. Ordination was performed via principal coordinates analysis using Bray-Curtis dissimilarity matrices. Differences in the abundance of OTUs were determined using the Kruskal-Wallis test. To identify the source of microbial communities among inter-connected habitats, we used SourceTracker v2.0 (Knights et al., 2011). SourceTracker uses a Bayesian-based method, comparing microbial community profiles between "source" and "sink", and predicts the contribution of each source to the sink. In the present study, creek, river, canal and stormwater outfall habitat samples were considered as "sources" and lake samples were considered as "sink". To identify bacterial OTUs (agglomerated by species) that are habitat-specific in habitats, we used "INDICSPECIES" package within R (Dufrene and Legendre, 1997;De Caceres and Legendre, 2009). This analysis uses an Indicator Value index (IndVal) to measure the strength of association between a species and a habitat type (Dufrene and Legendre, 1997). The significance of the association was then assessed using permutation tests with Benjamini-Hochberg False Discovery Rate (FDR) correction. Potentially erroneous rare species as defined by species without at least two sequences in 20% of the samples (Henson et al., 2018a) were removed using PhyloSeq package within R. Therefore, a species was considered to be habitat-specific only if the species is present more than two times in at least 20% of the samples. All statistical tests were evaluated at an α value of 0.05. Analysis of variance (ANOVA) with Tukey's *post hoc* statistical test was performed using XLSTAT software version 19.02.44369 (Addinsoft, Belmont, MA, USA).

2.5 Nucleotide sequence accession numbers

The sequence data are deposited in the NCBI Sequence Read Archive (SRA) database under the BioProject Accession Number PRJNA393519.

3. Results

3.1 Physicochemical characteristics of water

Five different habitats including creek, river, canal, stormwater outfall, and lake were examined. These habitats differed significantly in all physicochemical parameters measured, except for NOx content ($P = 0.43$). Water temperature of the lake habitat was significantly greater ($P < 0.0001$) than other habitats. Compared to the lake habitat, TDS was significantly higher in creek ($P < 0.0001$) and stormwater outfall habitats ($P < 0.01$). Salinity as assessed by specific conductivity also varied among the habitats ($P < 0.0001$), with creek and stormwater outfall habitats having higher salinity concentrations than river, canal and lake habitats. While dissolved oxygen content was similar between river, canal and stormwater outfall habitats, significant differences in DO content was observed between creek and lake habitats ($P < 0.05$).

Table 4. 1 Physicochemical parameters of habitats examined[a]

Parameter	Habitat				
	Creek	River	Canal	Stormwater Outfall	Lake
n	227	58	11	19	99
Water Temperature (°C)	13.8 ± 6.2	14.4 ± 6.8	16.7 ± 6.3	15.6 ± 3.4	20.5 ± 1.9
Specific Conductivity (µs . cm^{-1})	611.7 ± 208.7	499.4 ± 207.8	291.5 ± 37.9	607.8 ± 500.5	304.7 ± 8.4
TDS (mg . $liter^{-1}$)	0.6 ± 0.5	0.5 ± 0.4	0.2 ± 0.0	0.7 ± 0.5	0.3 ± 0.0
pH	7.8 ± 0.3	7.8 ± 0.3	7.9 ± 2.3	9.0 ± 0.2	8.6 ± 0.3
NOx (µM)	0.9 ± 2.2	0.5 ± 0.3	0.3 ± 0.2	0.9 ± 1.8	ND
Total Phosphorous (µM)	0.2 ± 0.3	0.1 ± 0.1	0.1 ± 0.0	0.1 ± 0.1	ND
DO (mg . $liter^{-1}$)	9.8 ± 2.8	9.5 ± 2.5	10.5 ± 2.3	9.1 ± 2.1	8.8 ± 2.2

[a]Values are mean ± standard deviation. Parameters include total dissolved solids (TDS), nitrite plus nitrate (NOx), and dissolved oxygen (DO), ND (not determined)

3.2 Bacterial community complexity and diversity

Sequencing of 414 samples resulted in a total of 86,824,144 reads. The sequence reads were then quality trimmed and aligned against Greengene reference database which generated 45,490,051 OTU counts with an average of 111,728 (±79,804) OTU counts per sample. To investigate alpha diversity, we calculated Shannon indices and ACE richness estimator. Mean Shannon indices for all habitats analyzed ranged from 3.41 to 3.92 (Fig. 4.1a and Table 4S.1). The ranking of diversity between the habitats was as follows: Lake > Canal > River > Creek > Stormwater Outfall. Differences in Shannon diversity as measured by the Kruskal-Wallis test in all habitats were significant ($P < 0.0001$). *Post hoc* statistical tests revealed that only the difference between lake and creek ($P < 0.0001$), lake and stormwater outfall ($P < 0.0001$), and between lake and river ($P = 0.0002$), were significant. The ACE index, a non-parametric estimator of species richness, determined the majority of differences (Fig. 4.1b) between habitats were not significant ($P = 0.15$). *Post hoc* statistical tests revealed that only the difference between the creek and lake habitats was significant ($P < 0.05$).

Bacterial community composition (beta diversity), as evaluated by ANOSIM, showed significant differences ($P < 0.0001$) between habitats (Table 4S.2). Bacterial communities from creek did not differ significantly from river ($P = 0.10$), canal ($P = 0.31$), and stormwater outfall ($P = 0.24$). However, significant differences were observed between the creek and lake bacterial communities ($P < 0.0001$). The community composition of river differed significantly from lake communities ($P < 0.0001$), whereas

no significant differences were observed between river and canal ($P = 0.59$), and between river and stormwater outfall ($P = 0.17$). Communities from the canal habitat differed significantly from stormwater outfall ($P < 0.05$) and lake communities ($P < 0.0001$). To identify the homogeneity of multivariate dispersions between groups of samples, we have performed PERMDISP and our result indicates that the dispersions of groups of samples (habitats) are significantly different ($P < 0.01$). Ordination of samples by principal-coordinate analyses using non-Euclidean distance measures (Bray-Curtis dissimilarity) revealed clustering by habitat type (Fig. 4.2), except for the creek habitats. PERMANOVA also supported this separation and showed that sampling period and habitat type explain the greatest amount of variance ($R^2 = 0.26$, $P < 0.001$ and $R^2 = 0.12$, $P < 0.001$, respectively) in diversity (Table 4.S3). However, when the samples collected in each month were analyzed separately, differences in bacterial diversity was observed and PERMANOVA analyses suggest that DO level, in addition to the habitat type, explains this difference ($R^2 = 0.15$, $P > 0.001$). The concentration of DO ranged between 3.7 – 23.0 mg/L and samples with lower DO conc. (below 7.0 mg/L) showed less diversity (alpha diversity) when compared with samples of higher DO level ($\geq$ 7.0 mg/L). Similar diversity profile was observed for sample with higher DO. The 7.0 mg/L cutoff was chosen based on a previous study that examined the shift of bacterial communities in response to changes in DO level (Spietz et al., 2015).

3.3 Taxonomic composition of bacterial communities

To identify the underlined association between bacterial communities with habitat type in greater detail, taxonomic composition was examined. Among all habitats, at the phylum level, members of Proteobacteria, Bacteroidetes, Actinobacteria, Cyanobacteria, and Firmicutes (Fig. 4.3) were dominant bacterial communities. As expected, Proteobacteria was the most abundant phylum in all habitats. Bacteroidetes was the second most abundant phylum in creek, and stormwater outfall habitats, whereas, in river, canals and lake habitats, Actinobacteria was the second most abundant group. Cyanobacteria was the third most abundant phylum (> 2.0 % of OTUs) in creek, canal and stormwater outfall habitats, whereas, in river, and lake habitats, Firmicutes was the third most predominant phylum (> 4% of the OTUs). Among the five habitats examined, abundance of Verrucomicrobia was highest in canals and lowest in lake habitat. Chloroflexi, another major phylum was present at a higher abundance in lake habitats (> 4.0 % of OTUs) than in other habitats (<1.0 % of OTUs). To identify differences in bacterial communities at a higher taxonomic resolution, the top 20 most abundant families were selected (Table 4S.4) and assessed for relative abundance. No significant difference (P = 0.84) was observed between the habitats.

Since Proteobacteria was the most abundant phylum in every habitat, we reexamined the sequences to obtain a detailed profile of this particular group. Within Proteobacteria, Betaproteobacteria and Alphaproteobacteria were the two most predominant classes (Fig. 4.3), comprising 65% to 85% of all OTUs corresponding to

Proteobacteria. However, compared to other habitats, Betaproteobacteria exhibited a relatively higher abundance in both creek and stormwater outfall. Gammaproteobacteria was the next most predominant class in every habitat except for canals, where Epsilonproteobacteria predominated (approx. 20% of all the Proteobacteria OTUs). Only a few OTUs (less than 1% of Proteobacteria) belonging to Zetaproteobacteria, TA18, and unclassified Proteobacteria (combined and referred to as Other Proteobacteria) were present in all habitats. At the family level, Commamonadaceae was the most abundant family within Betaproteobacteria (Table 4S.4). The most abundant families in other Proteobacteria classes include Pelagibacteraceae, Xanthomonadaceae, Campylobacteraceae, and, Myxococcaceae within Alphaproteobacteria, Gammaproteobacteria, Epsilonproteobacteria, and Deltaproteobacteria respectively. In addition, Flavobacteriaceae (phylum Bacteroidetes) was present in high abundance in all habitats, except for lakes.

Differences in creek samples were primarily attributable to a disproportionate increase in the abundance of bacterial species from two bacterial phyla. In some sites, Bacteroidetes was the most predominant phylum, while in others Cyanobacteria was the predominant group. However, the alpha diversity of these samples was similar to other creek samples. Samples with low DO (< 7.0 mg/L) were similar in taxonomic composition at the phylum level to that of samples with high DO level (≥ 7.0 mg/L), however, some differences were observed at a higher taxonomic resolution. For example, Bacteroidetes order Sphingobacteriales was enriched in samples with low DO level

whereas members of the order Flavobacteriales were enriched in samples with high DO level. Members of the order Acidomicorbiales (phylum Actinobacteria), family Rhodobacterales (phylum Proteobacteria), and family Pelagibacteraceae (phylum Proteobacteria) were also present in high abundance in samples with low DO level compared to samples with high DO level.

3.4 Microbial community connectivity between habitats

Among the habitats examined, lakes serve as receiving reservoirs for creeks, river, canals and stormwater outfalls. Therefore, lake bacterial community structure is heavily influenced by microbial dispersal from other habitats. To test the influence of microbial dispersal from other habitats in shaping lake bacterial communities, we used SourceTracker and our analysis showed that approx. 40% of the sequences in Lake Ontario water samples derived from creek habitats, whereas, in Lake Erie, the highest contribution was from canal habitat (approx. 50%) (Fig. 4.4). While river was the second most contributor in shaping Lake Ontario bacterial community structure (approx. 25%), creek was the second most contributor in Lake Erie (approx. 20%). A high proportion of sequences from unidentified sources (referred as ‘other’, Fig. 4.4) were present in both lakes. These may have been due to the presence of bacteria that are indigenous to the lakes or were transported into the lakes through other sources including ballast water, human and animals (see discussion).

3.5 Habitat-specific bacterial species in watersheds

We wish to identify potential habitat-specific bacterial species based on differential abundance. For this purpose, a species was considered to be habitat-specific if it was found two times in at least 20% of the samples of a particular habitat. Sixty-two bacterial species were found to be habitat-specific (FDR-corrected $P < 0.05$ and IndVal > 0.5) (Table 4.S5). The majority of the habitat-specific bacterial species were found in the lake habitat (forty-six bacterial species), followed by stormwater outfall, canal and river habitats (eight, seven, and one species respectively).

3.6 Fecal indicators and potential human pathogens

Watersheds of the Niagara Peninsula are routinely being monitored for fecal indicator bacteria using MST-based approaches (NPCA, 2014). While single marker MST-based approaches are useful in identifying individual microbes at the species level, HTS-based methods can serve as basis for MST-based approaches by providing information about a wide range of bacterial species. Therefore, we examined all the habitats for the presence of fecal indicators and pathogen containing genera. Because of the limitation of 16S sequencing based approaches in resolving taxonomy at the species level for many bacteria, only OTUs resolved at the genus level, were considered for this analysis. Pathogen-containing genera and fecal indicators were selected based on their source and potential impact on public health in freshwater environments (Whitman et al., 2014). OTUs from a total of nine indicators and pathogen-containing genera were found, although, their abundance differed among habitats (Fig. 4.5). For example, *Pseudomonas*

spp. was the most abundant genera found in the lake environment followed by *Mycobacterium*, whereas *Clostridium* and *Mycobacterium* were the two most abundant genera present in stormwater outfall habitat. *Mycobacterium* was also the most abundant group found in creek and river habitats. Compared to other habitats, the lake environment showed a relatively higher abundance of *Staphylococcus*. Among all the habitats examined, river exhibited the lowest abundance of all pathogen-containing genera. Because of the high abundance of *Mycobacterium* spp. in all habitats, we examined the subpopulations of this particular group. While approx. 70% of all the *Mycobacterium* OTUs mapped to nontuberculosis Mycobacteria (NTM) that includes *M. arupense*, *M. celatum*, *M. gordonae*, *M. llatzerense*, *M. vaccae,* others could not be mapped to a finer taxonomic level.

3.7 Temporal changes of bacterial diversity in watersheds

To determine the differences in diversity amongst bacterial communities across temporal scale, we compared the Shannon diversity and ACE index (Fig. 4.6). Only the creek, river, and lake habitats were considered for this analysis. For canal, and stormwater outfall, sufficient samples were not available for the considered sampling period. Both Shannon diversity and ACE index differed significantly ($P < 0.01$, and $P < 0.005$, respectively) for the creek habitat during the sampling period. However, the differences in Shannon diversity among individual sampling months were not consistent with the differences in ACE index. For the creek habitat, Shannon diversity significantly differed between samples collected in July and August and July and October ($P < 0.05$ and $P <$

0.005 respectively), whereas, for ACE index, the difference was significant between August and November, September and November, and between October and November ($P < 0.05$; all comparisons). Unlike the creek habitat, Shannon diversity did not differ significantly ($P = 0.49$) for the river habitat. However, significant differences ($P < 0.001$) were observed in the ACE index. Samples collected in November differed significantly from samples collected in July, August, September, and October ($P < 0.05$; all comparisons). Beta diversity also differed significantly ($P < 0.05$) between the samples collected in November and July for both creek and river habitats as determined by ANOSIM. The lake habitat, in contrast to creek and river habitats, did not differ significantly in either Shannon diversity ($P = 0.06$) or ACE index ($P = 0.96$) during the sampling period. We have also examined the temporal distribution of major bacterial groups in creek, river and lake habitats (Fig. 4S.2). While the relative abundance of the majority of the top 10 phyla decreased in winter in creek and river habitats, a sharp increase in abundance was observed for Firmicutes in both these habitats. The majority of the top 10 phyla exhibited a higher abundance in late summer to early fall in creek and river. For lakes, samples were collected only in summer and fall. While Proteobacteria, Actinobacteria and Verrucomicrobia abundance increased in fall, a decrease in abundance was observed for Bacteroidetes, Firmicutes and Chloroflexi in lake habitat.

4. Discussion

Recent investigations of freshwater bacterial communities have either focused on the identification of major taxonomic groups in individual habitats (Savio et al.,

2015;Wang et al., 2016) or compared the abundance of bacterial taxa on a spatial scale (Staley and Sadowsky, 2016;Mohiuddin et al., 2017). Understanding the microbial diversity of inter-connected habitats may provide important insight into the influence of one habitat upon another. The present study examined the microbial communities of five inter-connected freshwater habitats of the Niagara Region using next-generation sequencing. Within the context of our study, lake habitats showed highest bacterial diversity and richness, while the diversity was lowest in stormwater outfall habitat. Compared to other habitats, lakes exhibited a higher abundance of habitat-specific bacterial species. Among the environmental parameters tested in the study, temperature and DO level appear to have more influence on the structure of bacterial communities in all habitats. Presence of pathogen containing genera in each habitat suggest that future studies can be modified to identify pathogens at a fine taxonomic level.

Similar to the major bacterial groups identified in other freshwater habitats of this region, Proteobacteria, Bacteroidetes, Actinobacteria, Verrucomicrobia, and Firmicutes were the most abundant groups in the examined watersheds, with Proteobacteria being the most predominant group in all habitats (Fig. 4.3). However, the similarity in abundance of major taxonomic phyla was not recapitulated at the genus level, as *Flavobacterium* and *Fluviicola* spp., members of the Bacteroidetes phylum, predominated in the river and lake habitats. Interestingly, a large proportion of sequences from each habitat were identified as Actinobacteria. Actinobacteria, although primarily considered soil bacteria, are often found in freshwater environments (Newton et al., 2011). This is primarily due to the

prevalence of two orders of Actinobacteria, Actinomycetales and Acidimicrobiales (Ghai et al., 2014) which are photoheterotrophic and planktonic and, therefore it is not surprising that these bacterial orders are present in freshwater.

Members of the Proteobacteria phylum, such as Alpha-, Beta- and Gammaproteobacteria were abundant in all habitats and Epsilonproteobacteria were enriched in canal habitats. Although Epsilonproteobacteria includes a wide range of bacterial species, this increase was due to the enrichment of members from the families Campylobacteraceae and Helicobacteraceae. Both of these families harbor a wide range of pathogens including *Campylobacter* spp., *Helicobacter* spp., and *Arcobacter* spp., which are often found in runoffs from agricultural farms and wastewater treatment plants (Engberg et al., 2000;Hutchison et al., 2005;Lehner et al., 2005). Consistent with this association, the majority of the sampling sites for canals were located in close proximity to agricultural farms which likely account for the high abundance of Epsilonproteobacteria in this habitat. In addition, members of the family Flavobacteriaceae (phylum Bacteroidetes) were enriched in all habitats, except for lakes. Members of this family may be associated with algal particles and play important roles in degrading organic materials (Buchan et al., 2014).

We have identified several bacterial species that may be considered habitat-specific. Using indicator species analysis, we identified 62 habitat-specific bacterial species (Table 4S.5). The majority of the habitat-specific bacterial species (forty-six species) were identified in lake habitats, which is not surprising since they are the largest

watersheds in Southern Ontario. Relative to other habitats, both these watersheds experience higher anthropogenic activities during the summer period and are also used as reservoirs of wastewater and agricultural runoffs, which may further contribute to the higher bacterial diversity in these two lakes. The majority of the habitat-specific bacterial species identified in lake habitats (27, out of 46) belong to the phylum Proteobacteria while the remaining species belong to the phyla Firmicutes, Bacteroidetes, and Chlamydiae. For stormwater outfall, the majority of the habitat-specific bacterial species belong to Firmicutes, followed by species belonging to Proteobacteria and Actinobacteria. Stringent criteria (FDR-corrected $P < 0.05$ and IndVal > 0.5) were used to identify habitat-specific bacterial species as it was used in several other studies (Seedorf et al., 2014;Planer et al., 2016). However, this analysis may have some limitations. For example, the bacterial communities were characterized in samples collected only at particular sites which do not necessarily represent the entire watersheds. Therefore, potential biases may be introduced. The absence of habitat-specific species in creeks and their low abundance in habitats aside from lakes could also be due to OTU sampling depth limitations. Creeks, canals, rivers, and stormwater outfalls feed into lakes and therefore, the presence of habitat-specific bacterial species was expected in all of these habitats. In lakes, there could be a change in growth of microbial communities compared to their source habitats, increasing the abundance of relatively rare species to detectable levels. A more robust OTU sampling depth may address these potential issues in other habitats.

Microbial communities are sensitive to changes in environmental parameters and water chemistry (Allison and Martiny, 2008;Van Rossum et al., 2015). Strong correlation between microbial community complexity and abiotic factors including water temperature (Allison and Martiny, 2008), dissolved oxygen (Spietz et al., 2015;Aldunate et al., 2018), pH (Liu et al., 2015), and conductivity (Zhang et al., 2015) were examined. Among the environmental parameters tested in this study, DO level and temperature (sampling period) had strong impact on microbial community structure. Compared to samples with high DO level, members of the Sphingobacteriales, Acidomicrobiales, Rhodobacterales and Pelagibacteraceae were more abundant in samples with low DO levels. Similar bacterial profiles were also reported in previous studies investigating the bacterial communities in low DO environment (Spietz et al., 2015;Aldunate et al., 2018).

We have identified a possible link between inter-related aquatic environments. We showed that the adjacent creeks and canals heavily impact lake microbial community structure (Fig. 4.4). Many of these creeks and canals drain from agricultural sites into the lakes, potentially carrying a heavy load of nutrients and microbes. The differences in the contribution of each of the habitat in shaping lake microbial community composition is likely impacted by the type of agricultural practices in this region. Compared to Lake Ontario, Lake Erie region is characterized by more intense agricultural practices, which result in the increased release of agricultural farm runoffs into the canals which ultimately impact the Lake Erie microbial community composition. Similar results were obtained in studies that extensively investigated the effects of agricultural practices on the microbial

community diversity (Cloutier et al., 2015;Qu et al., 2017;Seuradge et al., 2017). A high contribution of unidentified source was observed in both Lake Ontario and Lake Erie. This unidentified source may contain microbes that are indigenous to lake habitats or are transported into lakes from sources including ballast water (MacIsaac et al., 2002), humans, birds and animals (Edge and Hill, 2007;Wright et al., 2009;Khan et al., 2013). While our analysis indicate a possible link between inter-connected habitats, our study only included one site in each lake and, therefore, this result must be taken with a degree of caution. In future, municipal sampling programs may be redesigned to include additional representative sites within type environments to allow more robust analysis of bacterial diversity.

Fecal indicators and pathogen-containing genera were also identified in this study. While some pathogenic genera such as *Mycobacterium*, *Pseudomonas* and *Clostridium* were identified in all habitats, others such as *Legionella* and *Staphylococcus* were found only in particular habitats (Fig. 4.5). Because of a high abundance of *Mycobacterium* genus in all habitats, we were able to resolve *Mycobacterium* to a finer taxonomic level. All the *Mycobacterium* spp. identified in this study, are often present in freshwater environments and cause human infections (Chilima et al., 2006;Mazumder et al., 2010;Liu et al., 2012;Delafont et al., 2017), except for *M. vaccae*. However, due to the limited resolution of 16S sequencing approach at high taxonomic levels, we were unable to distinguish most pathogens from naturally occurring non-pathogenic bacterial species. In this study, we used the V3 (partial) and V4 (complete) region of the 16S rRNA gene,

which have been validated for comprehensive profiling of bacterial communities from complex environmental samples (Klindworth et al., 2013). Here, we have limited our analysis at the genus level and referred pathogenic bacteria as pathogen-containing genera. Many of the pathogen-containing genera identified in this study were also reported in a previous study performed in Lake Ontario and Lake Erie (Mohiuddin et al., 2017), both of which are subjects of the current study (lake habitat). The limited capacity of 16S sequencing approach in resolving taxonomy at a fine level prevents us from reporting the presence of pathogenic bacteria at the species level. While the methodology employed in this study cannot accurately identify pathogens, information obtained from this study may serve as baseline for future studies investigating other freshwater habitats. Different methodology such as shotgun metagenomic sequencing and/or qPCR can be employed to identify bacterial pathogens at a fine taxonomic level (species or strain level).

Freshwater habitats undergo changes in physical parameters and resource availability which contribute to changes in bacterial community structure (Kent et al., 2004). In this work, we monitored changes in creek, river and lake bacterial communities over a five-month period. For creek and river habitats, Shannon diversity did not differ significantly during the sampling period while an increase in species richness was evident at the end of sampling period (November) (Fig. 4.6). Beta diversity also differed significantly (ANOSIM; $P < 0.05$) between the samples collected in July and November for both creek and river habitats. This change is mostly due to the change in seasonality

(summer to winter) as evident by the changes in abundance of the major bacterial groups (Fig. 4S.2). This finding is consistent with the previous studies investigating temporal diversity of bacterial communities in freshwater environments (Kent et al., 2004;Fujimoto et al., 2016;Ma et al., 2016).

Although 16S rRNA gene sequencing captures a large breadth of bacterial diversity, it does not provide robust information regarding the functional capacity of microbial communities. Despite these limitations, 16S rRNA gene amplicon sequencing, compared to shotgun metagenomic sequencing, is considered better in identifying broad levels of community composition (Poretsky et al., 2014;Tessler et al., 2017). In addition, primer pairs targeting the hypervariable region of the 16s rRNA gene may shift abundance estimates towards certain phyla (Tremblay et al., 2015). To address this, we used a primer pair that targets the V3-V4 hypervariable region of the 16S rRNA gene. This primer pair captures a broad range of bacterial phyla and was shown to be effective for microbial community analyses from complex environments with sequencing performed on Illumina MiSeq (Klindworth et al., 2013;Tremblay et al., 2015).

In this study, we partnered with three water monitoring authorities including federal, provincial and regional agencies and investigated the potential of a next-generation sequencing approach for water quality monitoring. With advancements in sequencing technologies, a reduction in associated costs, increased sensitivity in the detection of pathogens, and use of adequate quantification standards, a next-generation

sequencing approach can be implemented to augment existing water quality monitoring programs.

5. Conclusions

Here, we investigated the diversity of bacterial communities in five interconnected freshwater habitats. We also examined the influence of environmental factors on the distribution of bacterial species. Our analyses suggest that, although interconnected, lakes exhibit greater taxonomic diversity than any other habitats. At the phylum level, Proteobacteria, Bacteroidetes, and Actinobacteria were the three most dominant groups. While Proteobacteria was the most abundant phylum in all habitats, Bacteroidetes was the second most abundant group in creeks and stormwater outfalls. In other habitats, Actinobacteria was the second most abundant group. Unlike other habitats, canals exhibited a higher abundance of Epsilonproteobacteria. This increase was due to the enrichment of two families, Campylobacteraceae and Helicobacteraceae, which are frequently found in agricultural farm runoffs and in wastewater. Since the canals serve as reservoirs for agricultural farm runoffs, the increase in Epsilonproteobacteria in this habitat is expected. Strong association between dissolved oxygen levels and bacterial community composition was observed suggesting that bacterial communities are sensitive to changes in abiotic factors. Using SourceTracker, we have identified a possible link between inter-connected habitats. Habitat-specific bacterial species were identified in all habitats except for creeks. Among these habitats, lakes harbor the widest range of habitat-specific bacterial species. Although, creeks, rivers, canals and stormwater outfalls drain

into lakes, the increase of habitat-specific bacterial species in lakes could be due to increased bacterial growth (due to increased nutrient availability) in lakes, which results in detectable levels following sequencing. Sequences associated with a wide range of pathogen-containing-genera as well as fecal indicator bacteria were identified in all habitats suggesting that metagenomic approach can be used to identify bacterial groups that are not currently included in traditional monitoring programs.

6. Acknowledgement

We thank members of the Schellhorn Lab for their help and comments on the manuscript. We thank Joshua Diamond from Niagara Peninsula Conservation Authority, Glen Hudgin from Niagara Region Public Health and Thomas A. Edge from Environment Canada, for advice on the study. We gratefully acknowledge the financial support through the Niagara Region WaterSmart Program, the Natural Sciences and Engineering Research Council (NSERC) of Canada's Collaborative Research and Development Program (CRDP), and NSERC operating grants to HES. The authors declare no conflict of interest.

7. Author contributions

M.M.M., and H.E.S. designed study; M.M.M., and A.P. conducted laboratory experiments; M.M.M. performed bioinformatic analyses; S.R.B. wrote the script for Qiime usage; M.M.M. analyzed data and performed statistical analyses; H.E.S. supervised study; M.M.M. wrote the manuscript; All authors provided edits and approved the final version of the manuscript.

References

Aldunate, M., R. De la Iglesia, A.D. Bertagnolli & O. Ulloa, (2018) Oxygen modulates bacterial community composition in the coastal upwelling waters off central Chile. *Deep Sea Res Part 2 Top Stud Oceanogr*.

Allison, S.D. & J.B.H. Martiny, (2008) Resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci U S A* **105**: 11512-11519.

Anderson, M.J., (2006) Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62**: 245-253.

Anderson, M.J., K.E. Ellingsen & B.H. McArdle, (2006) Multivariate dispersion as a measure of beta diversity. *Ecol Lett* **9**: 683-693.

APHA, (2012) *Standard methods for the examination of water and wastewater*. American Public Health Association, Washington D.C.

Berry, M.A., T.W. Davis, R.M. Cory, M.B. Duhaime, T.H. Johengen, G.W. Kling, *et al.*, (2017) Cyanobacterial harmful algal blooms are a biological disturbance to Western Lake Erie bacterial communities. *Environ Microbiol* **19**: 1149-1162.

Bokulich, N.A., S. Subramanian, J.J. Faith, D. Gevers, J.I. Gordon, R. Knight, *et al.*, (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* **10**: 57-59.

Braun, B., J. Schroder, H. Knecht & U. Szewzyk, (2016) Unraveling the microbial community of a cold groundwater catchment system. *Water Res* **107**: 113-126.

Buchan, A., G.R. LeCleir, C.A. Gulvik & J.M. González, (2014) Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat Rev Microbiol* **12**: 686-698.

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

Chilima, B.Z., I.M. Clark, S. Floyd, P.E.M. Fine & P.R. Hirsch, (2006) Distribution of environmental mycobacteria in Karonga District, Northern Malawi. *Appl Environ Microbiol* **72**: 2343-2350.

Cloutier, D.D., E.W. Alm, S.L. McLellan & K.E. Wommack, (2015) Influence of land use, nutrients, and geography on microbial communities and fecal indicator abundance at Lake Michigan beaches. *Appl Environ Microbiol* **81**: 4904-4913.

De Caceres, M. & P. Legendre, (2009) Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**: 3566-3574.

Delafont, V., A. Samba-Louaka, E. Cambau, D. Bouchon, L. Moulin & Y. Héchard, (2017) *Mycobacterium llatzerense*, a waterborne *Mycobacterium*, that resists phagocytosis by *Acanthamoeba castellanii*. *Sci Rep* **7**.

Dufrene, M. & P. Legendre, (1997) Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecol Monogr* **67**: 345-366.

Edgar, R.C., (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.

Edge, T.A. & S. Hill, (2007) Multiple lines of evidence to identify the sources of fecal pollution at a freshwater beach in Hamilton Harbour, Lake Ontario. *Water Res* **41**: 3585-3594.

Engberg, J., S.L. On, C.S. Harrington & P. Gerner-Smidt, (2000) Prevalence of *Campylobacter*, *Arcobacter*, *Helicobacter*, and *Sutterella* spp. in human fecal samples as estimated by a reevaluation of isolation methods for Campylobacters. *J Clin Microbiol* **38**: 286-291.

Fong, T.T., L.S. Mansfield, D.L. Wilson, D.J. Schwab, S.L. Molloy & J.B. Rose, (2007) Massive microbiological groundwater contamination associated with a waterborne outbreak in Lake Erie, South Bass Island, Ohio. *Environ Health Perspect* **115**: 856-864.

Fujimoto, M., J. Cavaletto, J.R. Liebig, A. McCarthy, H.A. Vanderploeg & V.J. Denef, (2016) Spatiotemporal distribution of bacterioplankton functional groups along a freshwater estuary to pelagic gradient in Lake Michigan. *J Great Lakes Res* **42**: 1036-1048.

Ghai, R., C.M. Mizuno, A. Picazo, A. Camacho & F. Rodriguez-Valera, (2014) Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Mol Ecol* **23**: 6073-6090.

Henson, M.W., J. Hanssen, G. Spooner, P. Fleming, M. Pukonen, F. Stahr, *et al.*, (2018) Nutrient dynamics and stream order influence microbial community patterns along a 2914 kilometer transect of the Mississippi River. *Limnol Oceanogr*.

Horton, D.J., K.R. Theis, D.G. Uzarski & D.R. Learman, (2018) Microbial community structure and microbial networks correspond to nutrient gradients within coastal wetlands of the Laurentian Great Lakes. *bioRxiv*

Hutchison, M.L., L.D. Walters, S.M. Avery, F. Munro & A. Moore, (2005) Analyses of livestock production, waste storage, and pathogen levels and prevalences in farm manures. *Appl Environ Microbiol* **71**: 1231-1236.

Kent, A.D., S.E. Jones, A.C. Yannarell, J.M. Graham, G.H. Lauster, T.K. Kratz, *et al.*, (2004) Annual patterns in bacterioplankton community variability in a humic lake. *Microb Ecol* **48**: 550-560.

Khan, I.U.H., V. Gannon, C.C. Jokinen, R. Kent, W. Koning, D.R. Lapen, *et al.*, (2014) A national investigation of the prevalence and diversity of thermophilic *Campylobacter* species in agricultural watersheds in Canada. *Water Res* **61**: 243-252.

Khan, I.U.H., S. Hill, E. Nowak & T.A. Edge, (2013) Effect of incubation temperature on the detection of thermophilic Campylobacter species from freshwater beaches, nearby wastewater effluents, and bird fecal droppings. *Appl Environ Microbiol* **79**: 7639-7645.

Klindworth, A., E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, *et al.*, (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**: e1.

Knights, D., J. Kuczynski, E.S. Charlson, J. Zaneveld, M.C. Mozer, R.G. Collman, *et al.*, (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* **8**: 761-763.

Lee, C.S., M. Kim, C. Lee, Z. Yu & J. Lee, (2016) The microbiota of recreational freshwaters and the implications for environmental and public health. *Front Microbiol* **7**: 1826.

Lehner, A., T. Tasara & R. Stephan, (2005) Relevant aspects of *Arcobacter* spp. as potential foodborne pathogen. *Int J Food Microbiol* **102**: 127-135.

Liu, R., Z. Yu, H. Zhang, M. Yang, B. Shi & X. Liu, (2012) Diversity of bacteria and mycobacteria in biofilms of two urban drinking water distribution systems. *Can J Microbiol* **58**: 261-270.

Liu, S., H. Ren, L. Shen, L. Lou, G. Tian, P. Zheng, *et al.*, (2015) pH levels drive bacterial community structure in sediments of the Qiantang River as determined by 454 pyrosequencing. *Front Microbiol* **6**.

Ma, L., G. Mao, J. Liu, G. Gao, C. Zou, M.G. Bartlam, *et al.*, (2016) Spatial-temporal changes of bacterioplankton community along an exhorheic river. *Front Microbiol* **7**: 250.

MacIsaac, H.J., T.C. Robbins & M.A. Lewis, (2002) Modeling ships' ballast water as invasion threats to the Great Lakes. *Can J Fish Aquat Sci* **59**: 1245-1256.

Mazumder, S.A., A. Hicks & J. Norwood, (2010) *Mycobacterium gordonae* pulmonary infection in an immunocompetent adult. *North Am J Med Sci* **2**: 205-207.

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

Mohiuddin, M. & H.E. Schellhorn, (2015) Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* **6**.

Mohiuddin, M.M., Y. Salama, H.E. Schellhorn & G.B. Golding, (2017) Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Res* **115**: 360-369.

Newton, R.J., S.E. Jones, A. Eiler, K.D. McMahon & S. Bertilsson, (2011) A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* **75**: 14-49.

Newton, R.J. & S.L. McLellan, (2015) A unique assemblage of cosmopolitan freshwater bacteria and higher community diversity differentiate an urbanized estuary from oligotrophic Lake Michigan. *Front Microbiol* **6**: 1028.

NPCA, (2014) *NPCA Water Quality Monitoring Program: 2014 Report*. Niagara Peninsula Conservation Authority, Welland, ON.

Olapade, O.A., (2018) Community composition and diversity of coastal bacterioplankton assemblages in Lakes Michigan, Erie, and Huron. *Microb Ecol* **75**: 598-608.

Planer, J.D., Y. Peng, A.L. Kau, L.V. Blanton, I.M. Ndao, P.I. Tarr, *et al.*, (2016) Development of the gut microbiota and mucosal IgA responses in twins and gnotobiotic mice. *Nature* **534**: 263-266.

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

Qu, X., Z. Ren, H. Zhang, M. Zhang, Y. Zhang, X. Liu, *et al.*, (2017) Influences of anthropogenic land use on microbial community structure and functional potentials of stream benthic biofilms. *Sci Rep* **7**.

Read, D.S., H.S. Gweon, M.J. Bowes, L.K. Newbold, D. Field, M.J. Bailey, *et al.*, (2015) Catchment-scale biogeography of riverine bacterioplankton. *ISME J* **9**: 516-526.

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

Seedorf, H., N.W. Griffin, V.K. Ridaura, A. Reyes, J. Cheng, F.E. Rey, *et al.*, (2014) Bacteria from diverse habitats colonize and compete in the mouse gut. *Cell* **159**: 253-266.

Seuradge, B.J., M. Oelbermann, J.D. Neufeld & P. Baldrian, (2017) Depth-dependent influence of different land-use systems on bacterial biogeography. *FEMS Microbiol Ecol* **93**: fiw239.

Spietz, R.L., C.M. Williams, G. Rocap & M.C. Horner-Devine, (2015) A dissolved oxygen threshold for shifts in bacterial community structure in a seasonally hypoxic estuary. *PLoS One* **10**: e0135731.

Staley, C., T.J. Gould, P. Wang, J. Phillips, J.B. Cotner & M.J. Sadowsky, (2014) Bacterial community structure is indicative of chemical inputs in the upper Mississippi river. *Front Microbiol* **5**.

Staley, C., T. Kaiser, A. Lobos, W. Ahmed, V.J. Harwood, C.M. Brown, *et al.*, (2018) Application of SourceTracker for accurate identification of fecal pollution in recreational freshwater: A double-blinded study. *Environ Sci Technol* **52**: 4207-4217.

Staley, C. & M.J. Sadowsky, (2016) Regional similarities and consistent patterns of local variation in beach sand bacterial communities throughout the northern hemisphere. *Appl Environ Microbiol* **82**: 2751-2762.

Tamames, J., J.J. Abellan, M. Pignatelli, A. Camacho & A. Moya, (2010) Environmental distribution of prokaryotic taxa. *BMC Microbiol* **10**: 85.

Tessler, M., J.S. Neumann, E. Afshinnekoo, M. Pineda, R. Hersch, L.F.M. Velho, *et al.*, (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci Rep* **7**.

Tremblay, J., K. Singh, A. Fern, E.S. Kirton, S.M. He, T. Woyke, *et al.*, (2015) Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* **6**.

Van Rossum, T., M.A. Peabody, M.I. Uyaguari-Diaz, K.I. Cronin, M. Chan, J.R. Slobodan, *et al.*, (2015) Year-long metagenomic study of river microbiomes across land use and water quality. *Front Microbiol* **6**.

Wang, P., B. Chen, R.Q. Yuan, C.Q. Li & Y. Li, (2016) Characteristics of aquatic bacterial community and the influencing factors in an urban river. *Sci Total Environ* **569**: 382-389.

Weidhaas, J., A. Anderson, R. Jamal & D.W. Schaffner, (2018) Elucidating waterborne pathogen presence and aiding source apportionment in an impaired stream. *Appl Environ Microbiol* **84**: e02510-02517.

Whitman, R.L., V.J. Harwood, T.A. Edge, M.B. Nevers, M. Byappanahalli, K. Vijayavel, *et al.*, (2014) Microbes in beach sands: integrating environment, ecology and public health. *Rev Environ Sci Bio* **13**: 329-368.

Wright, M.E., H.M. Solo-Gabriele, S. Elmir & L.E. Fleming, (2009) Microbial load from animal feces at a recreational beach. *Mar Pollut Bull* **58**: 1649-1656.

Zhang, H.-H., S.-N. Chen, T.-L. Huang, W.-X. Ma, J.-L. Xu & X. Sun, (2015) Vertical distribution of bacterial community diversity and water quality during the reservoir thermal stratification. *Int J Env Res Public Health* **12**: 6933-6945.

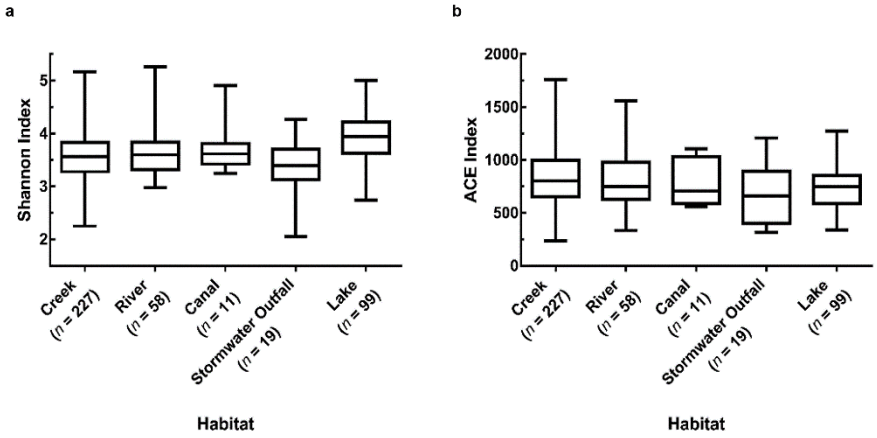

Figure 4.1 Shannon index (a) and bacterial richness as measured by abundance-based coverage estimate (ACE index) (b) in all the habitats at the genus level. The whiskers represent minimum and maximum values. Among all the habitats, taxonomic diversity was highest in lakes and lowest in stormwater outfalls.

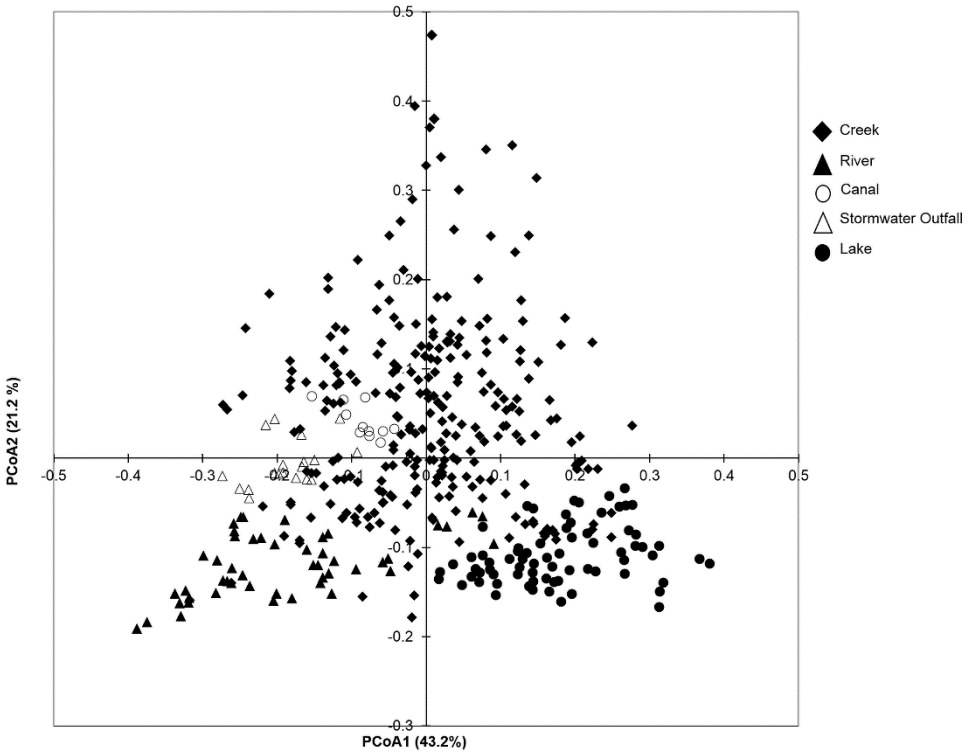

Figure 4.2 Principal coordinates analysis of bacterial communities (Bray-Curtis dissimilarities) from all the habitats. Samples were clustered based on the habitats.

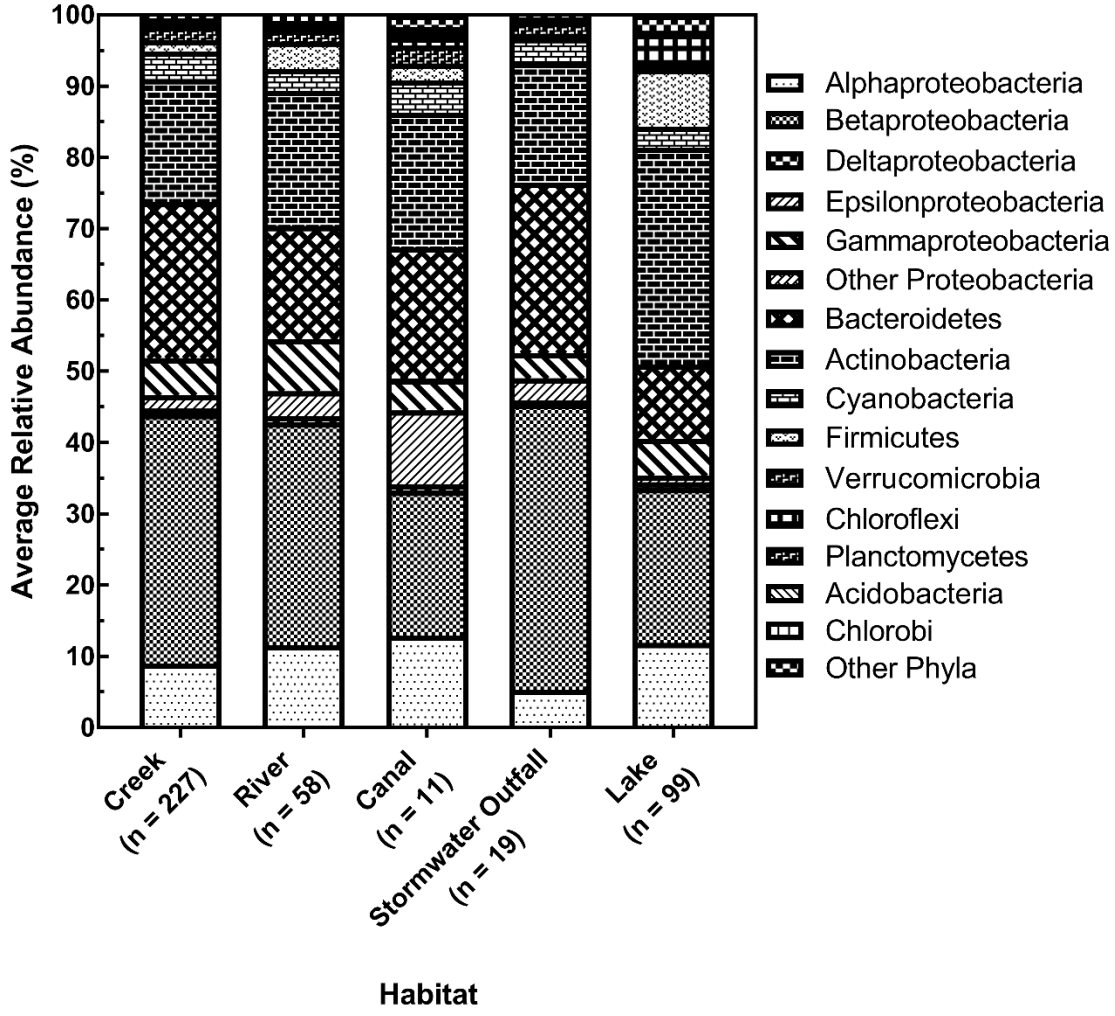

Figure 4.3 Distribution of major phyla and classes across all habitats. Because of the higher abundance of Proteobacteria phylum among all samples, taxonomic composition of Proteobacteria phylum is shown at the class level. Only top five classes are included. "Other Proteobacteria" refer to OTUs from TA18 and Zetaproteobacteria. OTUs from remaining phyla are labeled as other phyla.

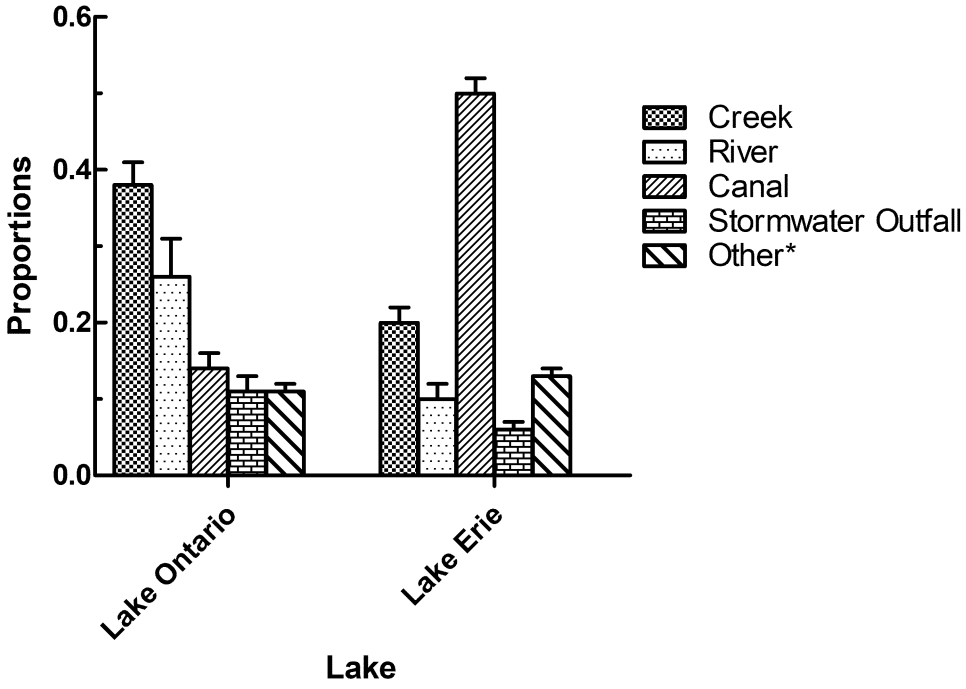

Figure 4.4 Proportional contribution of source habitats (creek, river, canal and stormwater outfall habitats) in shaping lake microbial community structure. SourceTracker, a Bayesian-based approach was used to quantify the contribution of each source habitat. Here, other refers to sequence proportion that was found only in lake habitat.

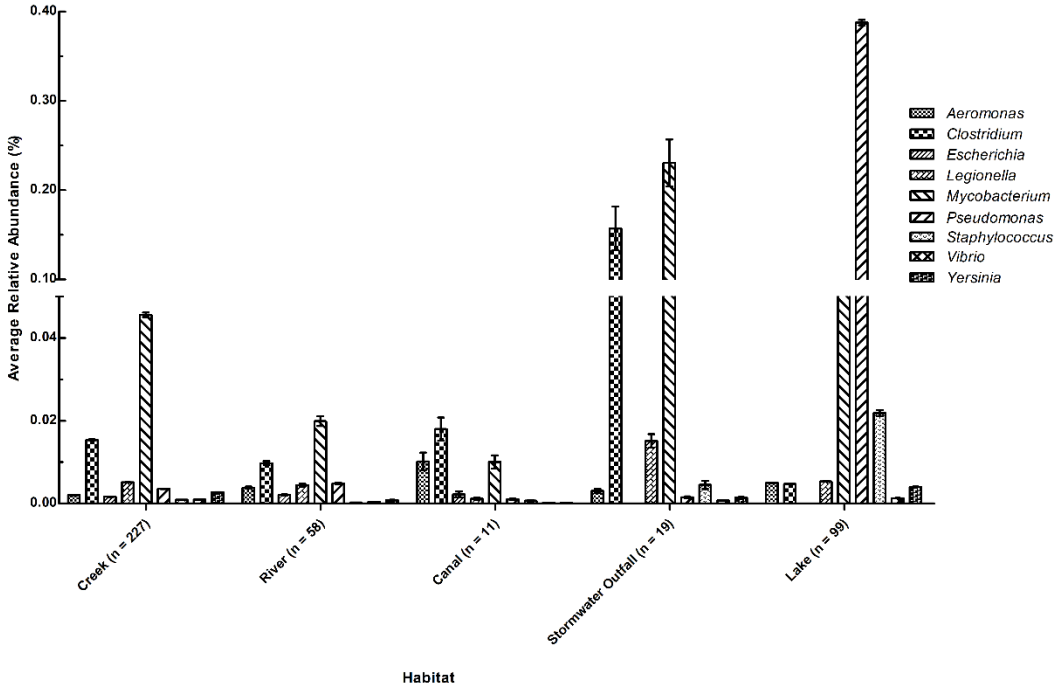

Figure 4.5 Relative abundance of potential human pathogen containing genera and fecal indicators in watersheds. OTUs corresponding to pathogens and FIBs were resolved at the genus level. Error bars indicate standard deviations of the mean.

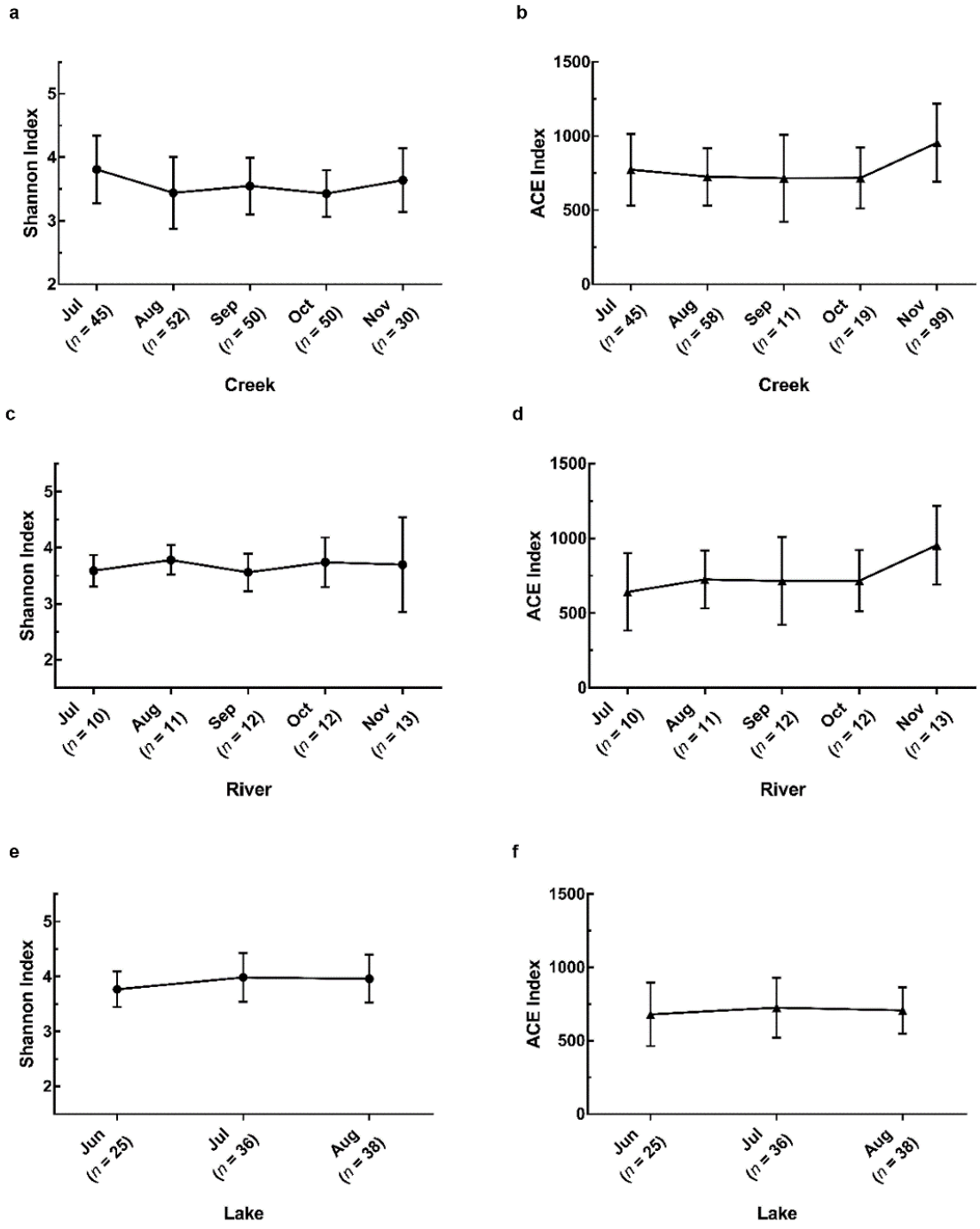

Figure 4.6 Temporal changes of microbial communities in creek (a, b), river (c, d), and lake (e, f) habitats based on Shannon diversity and species richness (ACE index).

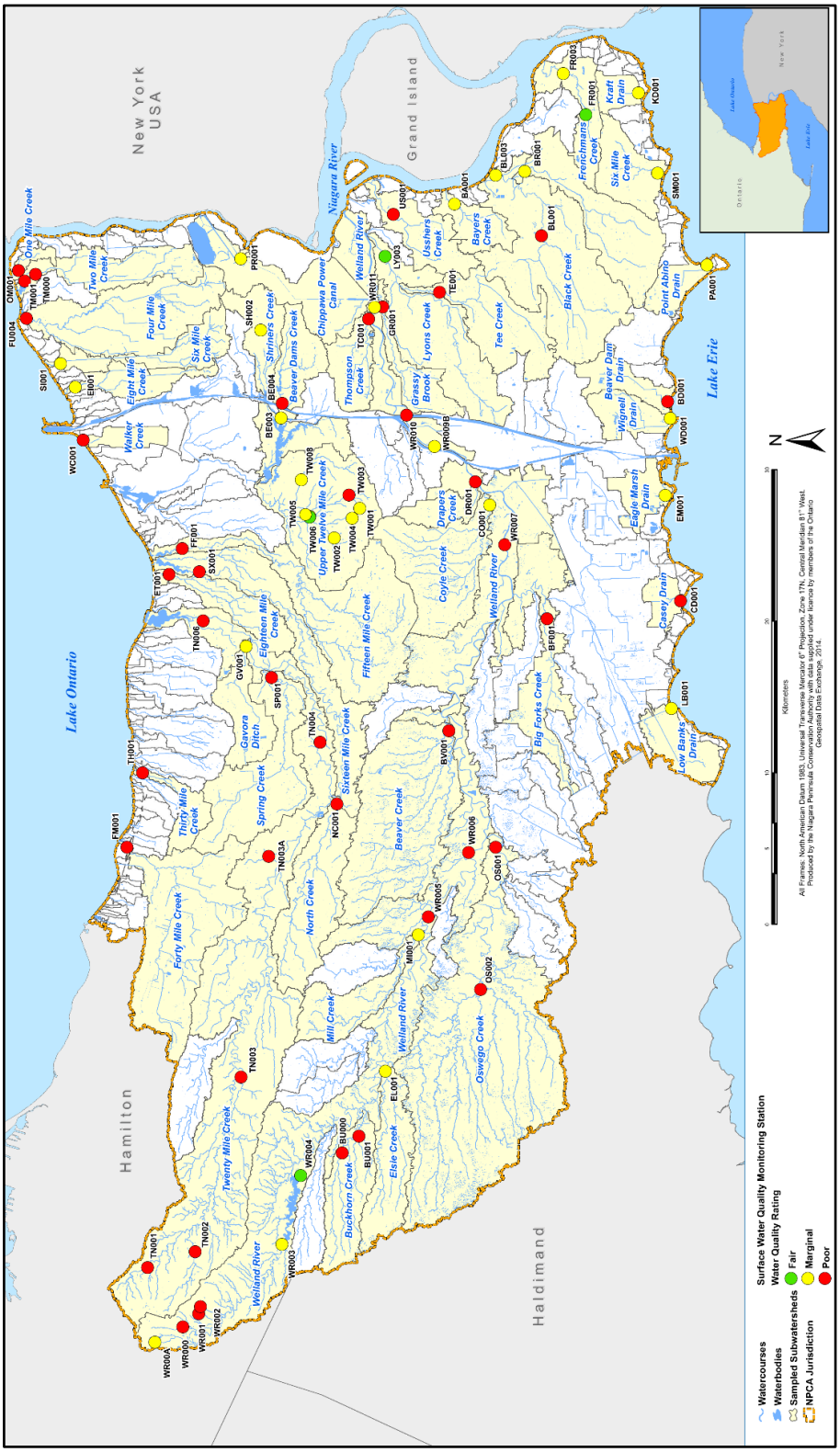

Figure 4S.1 Sampling locations of Niagara Peninsula. Sampling locations are indicated by colored circles. Green, yellow and red colored circles indicate Water Quality Index (fair, marginal and poor, respectively) used by the Canadian Council of Ministers of the Environment (CCME) to summarize data collected from Niagara Peninsula Conservation Authority (NPCA) water quality monitoring stations. Image: NPCA.

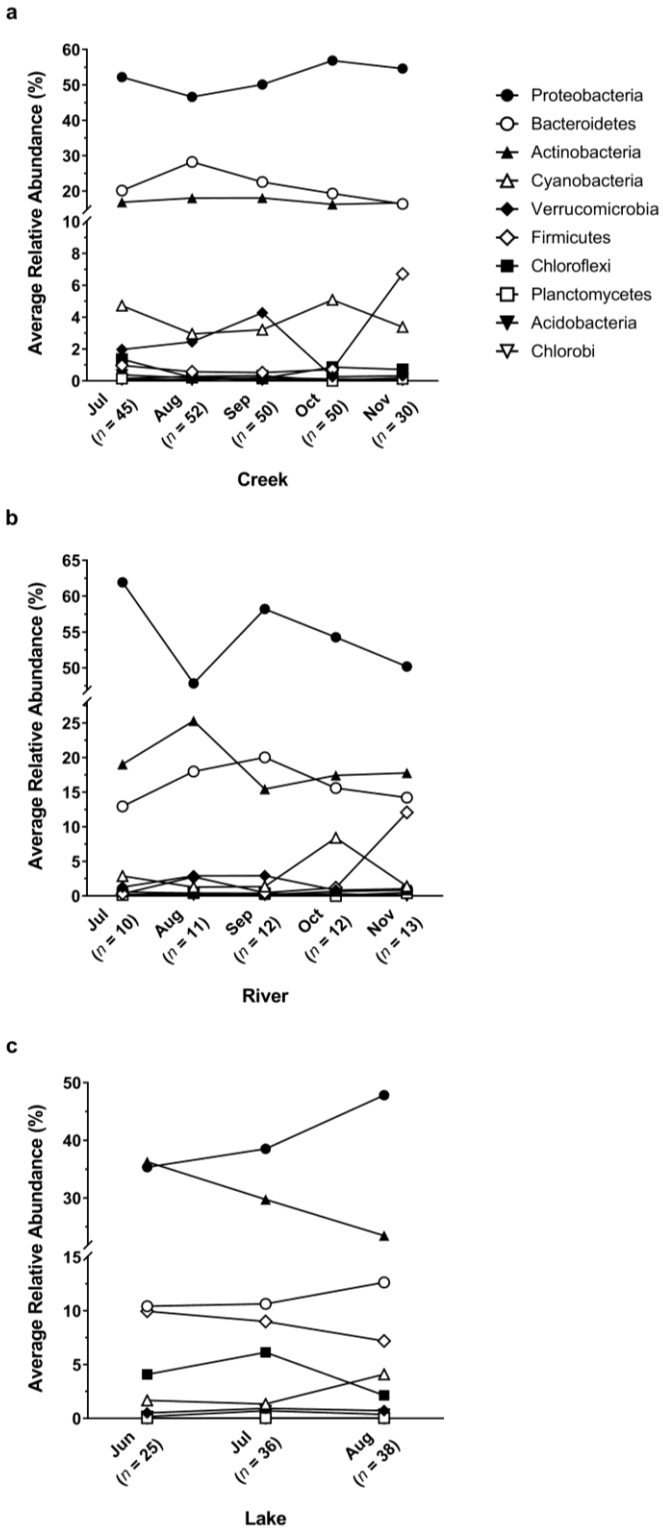

Figure 4S.2 Temporal distribution of major bacterial phyla in creek (a), river (b), and lake habitats (c). Relative abundances of only top 10 phyla are included.

Table 4S.1 Alpha diversity indices for bacterial communities

Habitat	*n*	Indices (± SD)	
		Shannon	ACE
Creek	227	3.58 ± 0.47	821.43 ± 322.40
River	58	3.65 ± 0.45	805.05 ± 252.97
Canal	11	3.72 ± 0.42	790.50 ± 204.80
Stormwater Outfall	19	3.41 ± 0.45	668.40 ± 256.48
Lake	99	3.92 ± 0.42	723.18 ± 181.69

Here, SD indicates standard deviation of the mean

Table 4S.2 Analysis of similarity (ANOSIM) of bacterial communities

Between Habitats	Dissimilarity ranks between and within habitats					
	0%	25%	50%	75%	100%	N
	1	29933.75	51990.5	68697.5	85463.5	53110
Creek	3	20304.50	33503.0	51661.5	85463.5	25651
River	60	22231.00	35329.0	59895.0	84592.0	1653
Canal	391	28416.50	40414.0	54863.5	74849.0	55
Stormwater Outfall	427	10266.00	27997.0	43876.0	72131.0	171
Lake	2	4424.50	9514.0	16645.0	84569.0	4851

Table 4S.3 Association between microbial community diversity with factors of interest

	Degrees of Freedom	Sums of Squares	Mean Squares	F.Model	R^2	Probability (> F)
Month	5	43.21	8.64	46.60	0.26	0.001
Habitat	4	21.16	5.29	28.52	0.12	0.001
Watershed	42	12.83	0.31	1.65	0.08	0.001
Month:Habitat	12	9.29	0.77	4.18	0.05	0.001
Month:Watershed	121	40.96	0.34	1.83	0.24	0.001
Residuals	229	42.47	0.19		0.25	
Total	413	169.93			1	

Table 4S.4 Relative abundance of top twenty families from all habitats

Family[a]	Mean relative abundance ± SD[b] (%)				
	Creek	River	Canal	Stormwater Outfall	Lake
Comamonadaceae	24.24 ± 9.73	21.89 ± 11.03	13.89 ± 3.87	28.69 ± 12.97	14.55 ± 3.82
ACK-M1	7.13 ± 5.79	9.14 ± 7.92	9.36 ± 7.65	6.86 ± 4.07	1.53 ± 0.97
Cryomorphaceae	5.87 ± 6.17	4.41 ± 5.04	3.42 ± 4.08	2.79 ± 5.75	6.66 ± 6.57
Flavobacteriaceae	4.33 ± 7.34	3.23 ± 4.71	3.73 ± 3.78	8.33 ± 13.98	0.29 ± 0.37
Cyclobacteriaceae	3.98 ± 5.89	2.18 ± 2.92	2.55 ± 2.96	1.15 ± 1.18	0.39 ± 0.61
Cytophagaceae	3.71 ± 5.45	1.71 ± 1.35	1.61 ± 1.19	4.95 ± 7.03	0.30 ± 0.32
Cerasicoccaceae	2.47 ± 2.69	2.12 ± 1.81	1.58 ± 1.31	4.26 ± 4.24	0.11 ± 0.30
Chitinophagaceae	2.40 ± 2.49	2.38 ± 2.20	1.00 ± 1.03	0.88 ± 1.61	2.11 ± 1.16
C111	2.23 ± 2.35	1.90 ± 1.91	1.37 ± 0.80	3.62 ± 2.51	1.25 ± 1.04
Xanthomonadaceae	2.01 ± 3.99	2.64 ± 4.16	5.24 ± 13.62	1.49 ± 2.92	0.73 ± 2.41
Oxalobacteraceae	1.40 ± 5.12	2.24 ± 7.63	0.22 ± 0.19	1.10 ± 1.03	2.92 ± 2.55
Microbacteriaceae	1.35 ± 1.88	2.09 ± 4.33	3.65 ± 6.61	0.46 ± 0.46	0.04 ± 0.16
Rhodobacteraceae	1.32 ± 1.34	1.03 ± 0.95	0.43 ± 0.50	0.39 ± 0.45	1.89 ± 2.02
Pelagibacteraceae	1.24 ± 1.66	1.24 ± 1.09	0.80 ± 0.67	2.32 ± 3.79	0.13 ± 0.44
Rhodocyclaceae	1.21 ± 1.34	1.62 ± 1.76	1.50 ± 1.17	3.37 ± 3.70	2.94 ± 1.68
Synechococcaceae	1.13 ± 2.10	1.03 ± 1.50	0.92 ± 1.29	0.09 ± 0.07	0.47 ± 0.46
Verrucomicrobiaceae	1.06 ± 1.80	0.91 ± 2.54	0.58 ± 0.79	0.13 ± 0.17	0.27 ± 0.49
Methylophilaceae	0.77 ± 5.34	0.70 ± 2.62	3.95 ± 8.28	0.22 ± 0.35	0.69 ± 0.74
Sphingomonadaceae	0.76 ± 1.46	0.82 ± 0.90	0.68 ± 0.62	0.66 ± 0.65	3.80 ± 1.94
R4-41B	0.74 ± 0.74	0.64 ± 0.68	0.30 ± 0.20	0.10 ± 0.14	0.37 ± 0.67

[a]Unclassified families were excluded from this analysis
[b]Standard Deviation

Table 4S.5 Differentially abundant bacterial species in habitats

Habitat	OTU ID	IndVal	FDR-corrected *P*-value
River	k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Pseudomonadales.f Pseudomonadaceae.g Azomonas.s agilis	0.6754	0.0028
Canal	k Bacteria.p Cyanobacteria.c Nostocophycideae.o Nostocales.f Nostocaceae.g Anabaena.s spiroides	0.5738	0.0014
	k Bacteria.p Firmicutes.c Bacilli.o Bacillales.f Bacillaceae.g Lentibacillus.s salis	0.5345	0.0014
	k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Corynebacteriaceae.g__Corynebacterium.s__aurimucosum	0.6275	0.0017
	k Bacteria.p Actinobacteria.c Actinobacteria.o Actinomycetales.f Propionibacteriaceae.g Propionibacterium.s acnes	0.5308	0.0017
	k Bacteria.p Cyanobacteria.c Nostocophycideae.o Nostocales.f Nostocaceae.g Nodularia.s harveyana	0.7598	0.0020
	k Bacteria.p Actinobacteria.c Actinobacteria.o Actinomycetales.f Intrasporangiaceae.g Kribbia.s dieselivorans	0.5409	0.0023
	k__Bacteria.p__Cyanobacteria.c__Synechococcophycideae.o__Pseudanabaenales.f__Pseudanabaenaceae.g__Prochlorothrix.s__hollandica	0.5353	0.0161
Stormwater Outfall	k Bacteria.p Actinobacteria.c Actinobacteria.o Actinomycetales.f Micromonosporaceae.g Actinoplanes.s durhamensis	0.7031	0.0006
	k Bacteria.p Firmicutes.c Clostridia.o Thermoanaerobacterales.f Caldicellulosiruptoraceae.g Caldicellulosiruptor.s saccharolyticus	0.7251	0.0006
	k Bacteria.p Proteobacteria.c Betaproteobacteria.o Burkholderiales.f Comamonadaceae.g Acidovorax.s wohlfahrtii	0.6122	0.0006
	k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Bacillaceae.g__Bacillus.s__ginsengihumi	0.5721	0.0026
	k Bacteria.p Firmicutes.c Erysipelotrichi.o Erysipelotrichales.f Erysipelotrichaceae.g [Eubacterium].s cylindroides	0.9011	0.0055
	k Bacteria.p Actinobacteria.c Actinobacteria.o Actinomycetales.f Cellulomonadaceae.g Oerskovia.s jenensis	0.5616	0.0077
	k Bacteria.p Proteobacteria.c Alphaproteobacteria.o Sphingomonadales.f Sphingomonadaceae.g Sphingobium.s xenophagum	0.5943	0.0134
	k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Leuconostocaceae.g__Oenococcus.s__oeni	0.5627	0.0170
Lake	k Bacteria.p Bacteroidetes.c Flavobacteriia.o Flavobacteriales.f Flavobacteriaceae.g Sediminibacter.s furfurosus	1.0149	0.0006
	k Bacteria.p Bacteroidetes.c Sphingobacteriia.o Sphingobacteriales.f Sphingobacteriaceae.g Pedobacter.s cryoconitis	0.9266	0.0006
	k Bacteria.p Chlamydiae.c Chlamydiia.o Chlamydiales.f Parachlamydiaceae.g Parachlamydia.s acanthamoebae	1.0023	0.0006
	k__Bacteria.p__Chlamydiae.c__Chlamydiia.o__Chlamydiales.f__Simkaniaceae.g__Fritschea.s__eriococci	0.7998	0.0006
	k Bacteria.p Fibrobacteres.c Fibrobacteria.o Fibrobacterales.f Fibrobacteraceae.g Fibrobacter.s succinogenes	0.8882	0.0006
	k Bacteria.p Firmicutes.c Bacilli.o Bacillales.f Bacillaceae.g Bacillus.s megaterium	0.8102	0.0006
	k Bacteria.p Firmicutes.c Bacilli.o Bacillales.f Listeriaceae.g Listeria.s grayi	0.9892	0.0006
	k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Paenibacillaceae.g__Ammoniphilus.s__oxalaticus	0.7687	0.0006
	k Bacteria.p Bacteroidetes.c Cytophagia.o Cytophagales.f Cyclobacteriaceae.g Algoriphagus.s hitonicola	0.9181	0.0006
	k Bacteria.p Proteobacteria.c Alphaproteobacteria.o Rhizobiales.f Methylocystaceae.g Pleomorphomonas.s oryzae	0.9976	0.0006
	k Bacteria.p Cyanobacteria.c Nostocophycideae.o Nostocales.f Nostocaceae.g Anabaena.s fertilissima	0.8209	0.0006
	k__Bacteria.p__Cyanobacteria.c__Oscillatoriophycideae.o__Chroococcales.f__Chroococcaceae.g__Limnococcus.s__limneticus	0.8873	0.0006
	k Bacteria.p Proteobacteria.c Betaproteobacteria.o Burkholderiales.f Burkholderiaceae.g Burkholderia.s glathei	0.9860	0.0006
	k Bacteria.p Proteobacteria.c Betaproteobacteria.o Burkholderiales.f Oxalobacteraceae.g Janthinobacterium.s lividum	0.8861	0.0006
	k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Oxalobacteraceae.g__Massilia.s__aerolata	0.9799	0.0006
	k Bacteria.p Proteobacteria.c Betaproteobacteria.o Burkholderiales.f Oxalobacteraceae.g Polynucleobacter.s cosmopolitanus	0.9150	0.0006
	k Bacteria.p Proteobacteria.c Betaproteobacteria.o Burkholderiales.f Oxalobacteraceae.g Telluria.s chitinolytica	0.8929	0.0006
	k Bacteria.p Proteobacteria.c Betaproteobacteria.o Neisseriales.f Neisseriaceae.g Chitinimonas.s taiwanensis	0.7692	0.0006

Table 4S.5 Differentially abundant bacterial species in habitats (continued)

k Bacteria.p Proteobacteria.c Deltaproteobacteria.o Bdellovibrionales.f Bacteriovoracaceae.g Peredibacter.s starrii	0.9353	0.0006
k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria.o__Desulfobacterales.f__Desulfobacteraceae.g__Desulfatibacillum.s__alkenivorans	0.8035	0.0006
k Bacteria.p Proteobacteria.c Deltaproteobacteria.o Desulfovibrionales.f Desulfovibrionaceae.g Desulfovibrio.s sulfodismutans	0.7776	0.0006
k Bacteria.p Proteobacteria.c Deltaproteobacteria.o Myxococcales.f Polyangiaceae.g Polyangium.s fumosum	0.8822	0.0006
k Bacteria.p Proteobacteria.c Epsilonproteobacteria.o Campylobacterales.f Campylobacteraceae.g Arcobacter.s cryaerophilus	0.9373	0.0006
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Alteromonadales.f__[Chromatiaceae].g__Alkalimonas.s__amylolytica	0.9623	0.0006
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Enterobacteriales.f Enterobacteriaceae.g Brenneria.s alni	0.9838	0.0006
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Enterobacteriales.f Enterobacteriaceae.g Enterobacter.s arachidis	0.9402	0.0006
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Enterobacteriales.f Enterobacteriaceae.g Xenorhabdus.s budapestensis	0.7475	0.0006
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae.g__Xenorhabdus.s__japonica	0.7547	0.0006
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Legionellales.f Legionellaceae.g Legionella.s impletisoli	0.8510	0.0006
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Vibrionales.f Vibrionaceae.g Photobacterium.s angustum	0.8417	0.0006
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Xanthomonadales.f Xanthomonadaceae.g Xanthomonas.s axonopodis	0.8363	0.0006
k__Bacteria.p__Tenericutes.c__Mollicutes.o__Acholeplasmatales.f__Acholeplasmataceae.g__Acholeplasma.s__laidlawii	0.9235	0.0006
k Bacteria.p Thermotogae.c Thermotogae.o Thermotogales.f Thermotogaceae.g Kosmotoga.s mrcj	0.8334	0.0006
k Bacteria.p Verrucomicrobia.c Verrucomicrobiae.o Verrucomicrobiales.f Verrucomicrobiaceae.g Prosthecobacter.s debontii 1	0.8044	0.0010
k Bacteria.p Verrucomicrobia.c [Spartobacteria].o [Chthoniobacterales].f [Chthoniobacteraceae].g Candidatus Xiphinematobacter.s 1	0.6606	0.0020
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Ruminococcaceae.g__Sporobacter.s__termitidis	0.5799	0.0020
k Bacteria.p Firmicutes.c Bacilli.o Bacillales.f Bacillaceae.g Bacillus.s thuringiensis	0.5657	0.0026
k Bacteria.p Proteobacteria.c Deltaproteobacteria.o Desulfuromonadales.f Pelobacteraceae.g Desulfuromonas.s michiganensis	0.8214	0.0043
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Enterobacteriales.f Enterobacteriaceae.g Enterobacter.s turicensis	0.5463	0.0050
k__Bacteria.p__Cyanobacteria.c__Nostocophycideae.o__Nostocales.f__Nostocaceae.g__Nodularia.s__sphaerocarpa	0.6102	0.0065
k Bacteria.p Proteobacteria.c Epsilonproteobacteria.o Campylobacterales.f Helicobacteraceae.g Wolinella.s succinogenes	0.6825	0.0065
k Bacteria.p Proteobacteria.c Gammaproteobacteria.o Alteromonadales.f Shewanellaceae.g Shewanella.s amazonensis	0.5809	0.0071
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Thiotrichales.f__Thiotrichaceae.g__Thiothrix.s__eikelboomii	0.5548	0.0141
k Bacteria.p Proteobacteria.c Betaproteobacteria.o Burkholderiales.f Burkholderiaceae.g Burkholderia.s andropogonis	0.6521	0.0039
k Bacteria.p Firmicutes.c Bacilli.o Lactobacillales.f Lactobacillaceae.g Lactobacillus.s salivarius	0.7125	0.0021

Chapter 5: Conclusion and Future Research

Summary

This dissertation sought to investigate the diversity of microbial and viral communities using a metagenomic approach to better understand bacterial and viral community composition in freshwater environments. To aid traditional water quality monitoring programs, the presence of pathogenic bacterial and viral species in recreational freshwaters was also investigated.

Recreational freshwaters are often investigated by MST-based approaches. Such approaches, while useful in identifying the source of fecal pollution, do not necessarily provide information about resident microbial and viral communities. In addition, pathogens that are not included in routine examination of water quality, may also be present in recreational waters. To address this limitation, an easy and routinely applicable methodology was employed for robust characterization of microbial and viral communities of freshwaters. Metagenomic exploration of viral communities provided insights into virus type and spatiotemporal distribution. The majority of the viruses identified are bacteriophages while viruses of plants and animals are also present in freshwaters. Among the major viral families identified, Myoviridae, Podoviridae and Siphoviridae constituted >80% of total viral populations. Although changes in the abundance of major viral families were observed on a temporal scale, the changes were statistically insignificant. Among the bacteriophages, phages of cyanobacteria were

predominant in all study sites. Within the context of the study, geographic location did not appear to have impact on viral community composition and diversity.

Presence of diverse viral communities in freshwaters suggested the need for examining the host microbial populations. Exhaustive analyses of bacterial communities indicated that the bacterial community composition is highly diverse in recreational beaches. Although Proteobacteria, Actinobacteria and Bacteroidetes were the most dominant groups in both recreational water and beach sand, a greater taxonomic diversity was examined in beach sand. Bacterial phyla corresponding to uncultured microorganisms or poorly characterized phyla are present only in beach sand. In addition, the beach sand environment was identified as a major determinant in shaping adjacent recreational water microbial community structure. Sequences of pathogens and fecal indicator bacteria are also present in both recreational water and beach sand.

Microbial community composition is often influenced by environmental and spatiotemporal factors. Investigation of the impact of abiotic factors on microbial communities of inter-connected habitats indicated a strong association between dissolved oxygen levels and bacterial community composition suggesting that bacterial communities are sensitive to changes in abiotic factors. Among the inter-connected habitats, lakes exhibited a greater taxonomic diversity than any other habitats. Using a computational approach, a possible link between inter-connected habitats was examined. Our analysis suggested that lake microbial communities are highly influenced by adjacent

creeks and canals. Examination of habitat-specific bacterial communities indicated that lakes harbor the widest range of habitat-specific bacterial species than any other habitats.

Future Directions

With the advances in sequencing technologies and reduction in costs of sequencing, metagenomic analysis of complex microbial and viral communities have become a standard approach in investigating microbial and viral community diversity. However, due to the lack of standard methodologies and bioinformatic analysis pipelines comparison between metagenomic datasets remains difficult. In addition, the difficulty in linking microbial and viral community composition to metabolic pathways still prevents a robust characterization of bacteria and viruses present in a given environment. However, the field of metagenomics is still evolving and moving towards a level of quantitative rigor. This will facilitate a more reliable interpretation of metagenomic datasets.

During the research completed for this thesis, many new questions developed that would continue our work and provide valuable data towards understanding microbial and viral communities.

1. The high abundance of cyanobacteria and cyanophages in freshwater warrants a closer look into the interactions between cyanobacteria and cyanophage populations. Understanding the role of cyanophages in the ecology of bloom forming cyanobacterial populations and the evolution of cyanophage and cyanobacterial genomes may aid in controlling harmful algal blooms in freshwaters.

2. Sequencing depth is a limiting factor in reconstructing metabolic pathways from shotgun metagenomic datasets. Deep sequencing of a small number of samples may resolve this issue and aid in linking metabolic pathways to microbial and viral communities present in a given environment. This will provide a better understanding of the ecology of microbial and viral communities as well as their function in specific environments.
3. Understanding bacterial host-virus interactions may facilitate a better understanding how microbial and viral communities are shaped in freshwaters. Identification of genes previously unknown and host-associated genes from metagenomic datasets may provide important insight into the horizontal gene transfer events between bacteria and viruses. Identification of CRISPR-Cas systems using genome-resolved metagenomics may also provide a better understanding of virus-host interactions.

Chapter 6 / Appendix A: Application of high-throughput 16S rRNA sequencing to identify fecal contamination sources and to complement the detection of fecal indicator bacteria in rural groundwater

Paul Naphtali, Mahi M. Mohiuddin, Athanasios Paschos, Herb E. Schellhorn*
Department of Biology, McMaster University, Hamilton, ON, Canada

*Correspondence: Herb E. Schellhorn, LSB 433, Department of Biology, McMaster University
1280 Main St W, Hamilton, ON, L8S 4K1, Canada
E-mail: schell@mcmaster.ca

Reproduced with permission from Naphtali, P., M.M. Mohiuddin, A. Paschos, H.E. Schellhorn, (2019) Application of high-throughput 16S rRNA sequencing to identify fecal contamination sources and to complement the detection of fecal indicator bacteria in rural groundwater. *Journal of Water and Health* https://doi.org/10.2166/wh.2019.295. IWA Publishing

Abstract

Residents in rural communities across Canada collect potable water from aquifers. Fecal contaminants from sewage and agricultural runoffs can penetrate aquifers, posing a public health risk. Standard methods for detecting fecal contamination test for fecal indicator bacteria (FIB), but the presence of these do not identify sources of contamination. In contrast, DNA-based diagnostic tools can achieve this important objective. We employed qPCR and high-throughput DNA sequencing to trace fecal contamination sources in Wainfleet, a rural Ontario township that has been under the longest active boil water advisory in Canada due to FIB contamination in groundwater wells. Using traditional methods, we identified FIBs indicating persistent fecal pollution in well waters. We used 16S rRNA sequencing to profile groundwater microbial communities and identified *Campylobacteraceae* as a fecal contamination DNA marker in septic tank effluents (STEs). We also identified *Turicibacter* and *Gallicola* as a potential cow and chicken fecal contamination marker, respectively. Using human specific *Bacteroidales* markers, we identified leaking septic tanks as the likely primary fecal contamination source in some of Wainfleet's groundwater. Overall, the results support the use of sequencing-based methods to augment traditional water quality testing methods and help end-users assess fecal contamination levels and identify point and non-point pollution sources.

1. Introduction

Fecal pollutants from sewage and agricultural runoff can penetrate decaying groundwater wells and render the well water unsafe to drink (USEPA 1993). Fecal contaminants may contain waterborne pathogens that transfer into aquatic environments and cause infectious disease (Harwood *et al.* 2014). Waterborne pathogen detection remains a challenge because of the diversity and low abundance of pathogens in water. *Escherichia coli* and *Enterococcus*, the standard fecal indicator bacteria (FIB), are present in high densities within the intestine of warm-blooded animals. FIB detection act as proxies for high fecal contamination levels (Field & Samadpour 2007). Public officials in rural communities enact boil water advisories upon FIB detection to lower the risk of waterborne disease outbreaks.

While FIBs indicate fecal contamination, the source of contamination and the true abundance of pathogens cannot be identified using FIBs alone. Current practices for monitoring water quality include the use of microbial source tracking (MST) methods that target genetic markers such as the 16S rRNA gene to quantify and source fecal contamination (Field & Samadpour 2007). The HF183 16S rRNA sequence, belonging to a human *Bacteroidales* 16S rRNA gene fragment, was the first genetic DNA marker used to detect human fecal contamination in drinking water (Bernhard & Field 2000). Quantitative PCR (qPCR) assays are now commonly used to quantify the HF183 marker as an indicator of human fecal pollution (Seurinck et al. 2005). Human and animal-associated DNA markers can also be used to measure point and non-point source contamination inputs in freshwater environments (Staley et al. 2013).

In addition to MST-based approaches, next-generation DNA sequencing is now being used to identify fecal contamination sources and to examine the co-occurrence of fecal and source water bacteria in respective environments (Unno et al. 2010). Next-generation sequencing based methods are also used to identify FIBs and pathogens in freshwater reservoirs (Mohiuddin et al. 2017, 2019). Aquatic microbiota containing taxa belonging to fecal bacteria are more likely to be contaminated through fecal sources (Cao et al. 2013). A 16S rRNA gene-based sequencing approach can also identify additional human fecal markers, such as *Lachnospiraceae*, for further characterization with other genetic methods such as oligotyping (McLellan et al. 2013). Despite the decreasing cost and increasing usefulness of next-generation sequencing methods to identify fecal contamination sources, MST protocols have not yet, however, integrated next-generation DNA sequencing as a standard for monitoring Canadian drinking water sources.

Wainfleet, a rural Ontario Township by Lake Erie, is under the longest active boil water advisory in Canada. A previous monitoring study on fecal contamination in 280 private residential groundwater wells determined that >50% contained detectable FIBs (Niagara Region Report, 2007). The town's boil water advisory provides an opportunity to test next-generation sequencing DNA methods in MST monitoring. Combining next-generation DNA sequencing approaches with traditional MST-based methods can augment current water quality monitoring approaches by incorporating new fecal-specific DNA markers to quantify host-specific contamination.

In this proof-of-principle study, we tested whether a next-generation DNA sequencing approach can be used to trace fecal contamination sources in Wainfleet's private well

waters. Using both traditional methods of FIB detection and 16S rRNA amplicon sequencing, we quantified fecal contamination levels and identified likely fecal pollution sources. We also measured the concentration of human *Bacteroidales* gene markers as a proxy of sewage-based contamination. Information obtained from our analyses can augment traditional methods of water quality monitoring by providing additional information on fecal pollution markers in potable waters.

2. Methods

2.1 Study site description

Wainfleet is situated in the southwest portion of the Niagara Region (42.92ºN, 79.38ºW). Residents obtain potable water using on-site groundwater wells. Many residences are built in low-lying areas close to Lake Erie. Of the 107 residences surveyed on March 2005 that use on-site groundwater wells, 44 have septic tanks that are >20 years old, and 49% of the residences do not comply with current provincial building codes (Niagara Region Report, 2007). Most homeowners install septic tanks to discharge septic tank effluents (STE) into the underlying soils through tile beds. Many of the plots have an area too small to install functioning septic tanks to current building standards. Besides the potential for leaking, concentrated raw sewage seeps through the underlying bed and into aquifers that supply wells (Niagara Region Report, 2007).

2.2 Groundwater tap sample collection

We identified nine test sampling sites (Sites A–K, Figure 6.1) and 21 wells based on FIB detection in the previous independent study (Niagara Region Report, 2007). We received

written consent from the township and identified volunteers based on the sampling site selection process. We collected tap water samples every month from April to November 2015 and grouped the samples based on season (spring, summer, and fall). A total of 48 samples were collected from nine test sites. For each sampling event, town technical staff collected groundwater tap samples from homeowners by filling 500 mL autoclaved plastic bottles (Nalgene). The water samples were then kept on ice, transported to the lab, and processed within 6 h of collection.

2.3 Septic tank effluent (STE) and manure sample collection

STE were collected from two septic tanks owned by two homeowners that participated in this study (fall 2015). Three biological replicates for each sewage sample were collected on three consecutive days. Manure samples were collected from manure mounds stored by a concentrated animal feeding operation for chickens, a cow farm, a horse hobby farm, and a pig farm. Similar to septic samples, three biological replicates for each manure type were collected on three consecutive days. All samples were transferred into 500 mL autoclaved bottles, kept in ice and transported to the lab. The samples were then stored at –80 ºC until further analysis.

2.4 FIB detection assay

FIBs in water samples were detected using standard methods (APHA 2012). To enumerate *E. coli* and *Enterococcus* spp., 100-fold serial dilutions were prepared by transferring 1 mL of well water into 100 mL of 1× PBS solution (pH 7.0), to a 10^4-fold dilution. One hundred mL of the well water sample and PBS diluted samples were passed

through 0.45 µm pore-size 47-mm-diameter sterile mixed cellulose ester membrane filters (Thermo Fisher Scientific, Burlington, ON, Canada). Each membrane filter was placed on differential coliform (DC, Oxoid) and mEI agar (BD Difco) plate and incubated at 42 ºC for 24 h. After incubation, *E. coli* and *Enterococcus* spp. colony forming units (CFUs) for filters containing stock and diluted well water samples were enumerated. To determine the concentration of FIBs (CFU/mL) in water samples, we multiplied the CFU count by its associated dilution factor. Detection of one (or more) CFU per 100 mL of drinking water samples were considered positive for FIBs (deemed unsafe for drinking (Health-Canada 2013)).

2.5 DNA extraction and library generation for sequencing

DNA was extracted from water samples as described earlier (Mohiuddin *et al.* 2019). Briefly, 500-mL of each groundwater sample was passed through 0.45-m pore-size 47-mm-diameter sterile mixed cellulose ester membrane filters. The filters were then cut into fragments (1 cm^2 size) with sterile scissors and the cut fragments were aseptically transferred with sterile forceps into 1.7-mL microfuge tubes for DNA extraction. DNA was extracted from the filters using a Soil DNA Isolation Kit (Norgen Biotek, Thorold, ON, Canada) according to manufacturers' instruction and isolated DNA was stored at –20 ºC for further analysis. DNA quantity and purity were measured using the Nanodrop 2000 UV-Vis Spectrophotometer (Thermo Fisher Scientific, Burlington, ON, Canada). Manure samples stored at –80 ºC were taken out overnight to thaw the samples and 200 mg of thawed samples were used for DNA extraction using a Stool Nucleic Acid

Isolation Kit (Norgen Biotek, Thorold, ON, Canada). Similar to manure samples, septic tank samples were taken out overnight to thaw the samples and 10 mL of the thawed samples were centrifuged at 10,000× *g* for 10 min at 4 ºC. After centrifugation, the pellet was resuspended with resuspension buffer provided with Stool Nucleic Acid Isolation Kit (Norgen Biotek, Thorold, ON, Canada) and DNA was extracted according to manufacturers' instruction.

After DNA extraction, PCR assays were conducted to amplify the V3–V4 region of the 16S rRNA gene using an optimized primer pair (S-D-Bact-0341-b-102 S-17/S-D-Bact-0785-a-A-21) evaluated elsewhere (Klindworth et al. 2013). Reaction mixes of 25.0 µL were prepared as follows: 2.5 µL 10× PCR buffer minus Mg^{2+} (Invitrogen, Burlington, ON, Canada), 0.5 µL of 100 mM dNTP solution, 1.0 µL each of the forward and reverse primer (1.0 µM each) with unique Illumina adapter sequences attached to them, 1.0 µL 10 mg/mL UV-treated BSA solution, 0.75 µL 25 mM $MgCl_2$ solution (Invitrogen, Burlington, ON, Canada), 0.25 µL Taq DNA Polymerase (Invitrogen), 2.0 µL of DNA template, and 16.0 µL ddH_2O. The target DNA was amplified using the T100 Thermal Cycler (Bio-Rad, Mississauga, ON, Canada) and conducted as follows: Initial denaturation at 95 ºC for 3 min, 40 cycles of denaturation at 94 ºC for 30 s, annealing at 50 ºC for 30 s, and elongation at 72 ºC for 1 min, and a final extension step at 72 ºC for 10 min.

All PCR products were examined in 0.9% agarose gels, the desired band was cut from the gel and DNA was extracted from the cut band with the Nucleospin PCR and Gel Cleanup Kit (Macherey-Nagel, Bethelhem, PA, USA). All amplicon extracts were pooled in equal

mass amounts and sent to the Farncombe Metagenomics Facility at McMaster University for further processing and sequencing. Product size and the presence of primer dimers in each library were checked using BioAnalyzer and High Sensitivity DNA Kit (Agilent, Mississauga, ON, Canada). Each library was then quantified by qPCR and sequencing was performed on the Illumina MiSeq platform which generated 2×300 bp paired-end sequences.

2.6 Processing and analysis of 16S rRNA gene sequences

Paired-end sequences generated by Illumina MiSeq were de-multiplexed and quality-filtered using methods described earlier (Bokulich et al. 2013; Mohiuddin et al. 2019). Only full amplicons with 464 bp sequences were considered for further analysis. Before downstream analysis, all sequences were trimmed using a quality score threshold (Q score) of 25 over at least 75% of the sequence read. Sequences that contained ambiguous bases, errors in barcode sequences, >10 consecutive low-quality base pairs and >2 nt mismatches from the primer sequences were also removed (Bokulich et al. 2013). Processed sequence reads were then analyzed using software package QIIME v.1.9.0 (Caporaso et al. 2010). Sequences were clustered into Operational Taxonomic Units (OTUs) sharing 97% nucleotide sequence identity and a minimum query alignment length of 50% to the Greengenes 2013 reference database (McDonald *et al.* 2012) with a sequence similarity threshold of 97% using UCLUST (Edgar 2010). The resulting OTU table was then rarified to 5000 OTU counts/sample before statistical analyses were performed. A potential DNA marker of a human or animal-specific fecal or manure contaminant was identified if it was detected in ≥10% abundance in a fecal sample but

≤1% abundance in all other fecal samples. The relative abundance of the potential markers was determined in the groundwater samples across the sampling months.

2.7 Preparation of DNA standards for quantification of human fecal pollution

The HF183 primer pair (HF183F and Bac708R) (Bernhard & Field 2000) that flanks the fragments in the 16S rRNA gene of human *Bacteroidales* populations was used to prepare standards for quantification of human-based fecal pollution. First, a conventional PCR assay was performed to generate a 500 bp product. The amplified product was then cloned into a vector which served as standard in qPCR assay. For conventional PCR, 50.0 µL reaction volumes were prepared. The reaction mix contained 5.0 µL of Thermopol Buffer (10×, NEB), 1.0 µL (10.0 µM each) each of the forward (HF183F: 5′-ATCATGAGTTCACATGTCCG-3′) and reverse primer (Bac708R: 5′-CAATCGGAGTTCTTCGTG-3′), 1.0 µL of dNTP mix (NEB, 10 mM), 1.0 µL of 5 U *Taq* Polymerase (NEB), and 2.0 µL of DNA template. All PCR assays were run with the conditions as follows: initial denaturation for 5 min at 95 ºC, 35 cycles of denaturation at 95 ºC for 30 s, annealing at 52 ºC, and extension at 72 ºC for 1 min, and a final extension step at 72 ºC for 6 min. PCR products were loaded onto 1% agarose gels and amplicons were extracted from the gel using the Nucleospin PCR and Gel Cleanup Kit (Macherey-Nagel, Bethelhem, PA, USA). To generate plasmid DNA standards, the TOPO TA Cloning Reaction (Invitrogen, Burlington, ON, Canada) was used. Briefly, 2.0 µL of the amplicons were ligated into the TOPO vector, according to the manufacturer's instructions. Transformants were isolated using blue-white screening technique and

subsequently purified. Purified colonies were then inoculated into 3.0 mL of LB broth supplemented with 50.0 µg/mL kanamycin for overnight incubation at 37 ºC with shaking at 220 rpm. DNA was extracted from 1.5 mL aliquots of the cultures with a Plasmid DNA Miniprep Kit (Norgen Biotek, Thorold, ON, Canada). DNA concentrations were quantified using the Qubit Fluorometer 2.0 (Invitrogen, Burlington, ON, Canada). DNA extracts were diluted such that a range of 10^1–10^7 plasmid copies of the marker was present in the standards.

2.8 Quantification of human *Bacteroidales* marker in sewage and groundwater samples

Quantification of human specific *Bacteroidales* marker was performed using the same forward primer used for the conventional PCR (HF183F) and a different reverse primer (5′-TACCCCGCCTACTATCTAATG-3′) (Seurinck et al. 2005). The newly amplified product had a length of 82 bp. qPCR assays were conducted using 2.0 µL of groundwater DNA extracts and plasmid standard in 10.0 µL reaction mix volumes. All DNA templates were diluted 10-fold to minimize qPCR inhibition. Each reaction mix contained 1.0 µL of DNA template, 5.0 µL of the SsoFast EvaGreen SuperMix (Bio-Rad), and 0.5 µL each of the forward and reverse primer (10 µM). Thermal cycling conditions included the following steps: Initial denaturation at 95 ºC for 30 s, 40 cycles of denaturation at 95 ºC for 30 s and 60 ºC for 10 s, and then a melt curve analysis from 65 to 95 ºC at increments of 0.5 ºC for 5 s per increment. The CFX Analyzer (Bio-Rad, Mississauga, ON, Canada) was used to generate a standard curve and a melt curve and quantify qPCR copy numbers.

3. Results

3.1 Detection of FIB in groundwater wells

As a preliminary assessment of fecal contamination levels in Wainfleet's well waters, we examined FIB detection frequency in tap water collected from private wells using Health Canada recommended guidelines (see above under "Methods"). Sixty-one per cent of the groundwater samples collected from 21 wells within the boil water advisory in 2015 contained either *E. coli* or *Enterococcus* (Figure 6.2). Positive *E. coli* detection ranged from 60.0 to 70.0%. *Enterococcus* detection rates had a large seasonal range from 10.0 to 81.0%, with the detection rate lowest in the spring and highest in the summer (Figure 6.2). Groundwater samples collected from two residential wells, B and K, also contained much higher mean *E. coli* and *Enterococcus* than the other groundwater wells (Figure 6.3).

3.2 Basic sequencing data

As a first step towards identifying potential fecal contamination sources in the FIB-positive well waters, we profiled microbial communities from 21 groundwater samples, six samples from two STEs, and 12 manure samples from four different animal farm using 16S rRNA sequencing. Sequence reads obtained from septic tank samples were classified into 972 (±89) OTUs. The number of reads obtained from animal manure and groundwater samples were higher than septic tank samples and reads from animal manure and groundwater wells were classified into 2987 (±315) and 3101 (±287) OTUs respectively.

3.3 Identification of potential STE and animal manure contamination markers

To identify potential fecal contamination markers in groundwater, we parsed for sewage and manure-associated bacteria in the 16S rRNA sequencing data. We defined a host-specific fecal marker as a microbial group detected at ≥10.0% relative abundance in one type of fecal matter but detected at ≤1.0% abundance in the other fecal sources. In the two STE samples, the mean abundance of *Campylobacteraceae* was 32.5% (Figure 6.4). At the genus level, sequences annotated to *Sulphospirillum* and *Arcobacter* were the most abundant members of *Campylobacteraceae.* In contrast, the average abundance of *Campylobacteraceae* was ≤1% for all four animal manure samples. *Gallicola* and *Turicibacter* were the most abundant genera in chicken and cow manure, comprising 42.2 and 9.4% of 16S rRNA sequences respectively. In contrast, these markers were detected at ≤1.0% abundance among the pig and horse manure and STE samples (Figure 6.4). No genetic marker was identified in pig and horse manure based on the classification criterion.

3.4 Abundance of STE and manure-based OTUs in groundwater samples

To identify possible fecal contaminants in well water, we determined the relative abundance of the human and animal-associated markers we identified in the reference fecal samples. Groundwater wells sampled in July, September, and November had the highest relative abundance of STE-based *Campylobacteraceae* sequences (Figure 6.5). None of the chicken and cow markers had a relative abundance above 2.0%, with the

relative abundance below 1.0% in wells collected in July, August, and November (Figure 6.5).

3.5 Identification of human fecal contamination

To determine whether the human-specific fecal contamination was present in the groundwater samples, we conducted qPCR assays of the HF183 human *Bacteroidales* marker in selected wells with or without FIB detection. We prepared a qPCR standard curve that quantifies the human *Bacteroidales* marker in each run (Figure 6S.1A). Each standard curve was robust for HF183 quantification (Figure 6S.1). Melt curve analysis of the standard curve and all groundwater samples that were examined after the qPCR assay yielded a single peak at 84 ºC (Figure 6S.1B).

The HF183 marker genome was more abundant in STE samples than in groundwater wells (Figure 6.6). qPCR of the human *Bacteroidales* marker in groundwater wells B, F and K resulted in positive detection, containing 30–50 genome copies/100 mL (Figure 6.6). A subset of groundwater wells without *E. coli* detection also tested positive for the HF183 marker. Groundwater well A, treated with UV light before sampling, contained more HF183 copies than other groundwater wells (Figure 6.6). Groundwater well I had a slightly lower HF183 marker levels B, F, and K.

4. Discussion

The presence of FIBs in drinking waters is a major health concern. Using national guidelines for drinking water and the results of a previous study as a reference, we sought

to determine whether high levels of FIBs are still present in the Wainfleet well waters. Our analyses suggest that FIBs are still present in over half of the tested well water sites and, therefore, the quality of drinking water may still be a concern in majority of the households within the township. Many of the sampled groundwater wells are located near the shores of Lake Erie, a low-lying region where most of the *E. coli* and total coliform exceedances were previously observed (Niagara Region Report, 2007). Some well water samples also contained far higher FIB counts than others. Mean *E. coli* and *Enterococcus* levels in some groundwater wells were three and two orders of magnitude higher than the average FIB counts for the groundwater wells collected across all the other locations. These wells with higher than average FIB counts are located near septic tanks that may leak raw sewage into the aquifer (Niagara Region Report, 2007). Poor well maintenance also facilitates sewage leaching into the groundwater (Howard et al. 2003). Altogether, chronic FIB contamination in individual groundwater wells like B and K require further investigation to confirm potential fecal contamination sources.

4.1 Characterization of fecal microbiota and identification of potential fecal markers

The presence of host-associated fecal contamination may help to trace fecal contamination sources in complex environmental samples. As a proof-of-principle experiment, we examined whether we could complement FIB detection with 16S rRNA sequencing methods by identifying sewage and manure-associated markers. We selected a maximum detection rate of 1.0% limit to ensure the specificity of the marker to its host. In cow manure, we identified *Turicibacter* spp., a member of the Firmicutes phylum

present in high abundance at the genus level, agreeing with previous cow microbiome profiles (Kim et al. 2011). In chicken manure, *Gallicola* – also a member of the Firmicutes phylum – was the most abundant genus. Both markers were also detected at ≤0.1% abundance in other manure types and sewage samples. These genera may act as DNA markers of host-associated fecal contamination in the town's groundwater wells.

In the two STE sites, we identified the *Campylobacteraceae* family (member of the Proteobacteria phylum) as fecal pollution marker, comprising 32.5% of 16S rRNA gene sequences at the family level. The predominance of *Campylobacteraceae* 16S rRNA gene sequences in STEs is identified in other septic tank systems (Tomaras *et al.* 2009). Although we only collected STEs from two septic tanks, both septic tank microbiomes contained a similar abundance of *Campylobacteraceae* sequences. *Campylobacteraceae* was also absent in the animal manure samples, suggesting that the *Campylobacteraceae* microbes are likely exclusive to STEs in the township.

While we identified host-associated markers for chicken manure, cow manure, and STEs, we could not establish the presence of host-associated markers for horse and pig manure. Although *Clostridium* and *Turicibacter* were more abundant in horse and pig manure samples, these sequences were also present in other manure types. However, analyses of additional samples that include both biological and technical replicates are required to confirm the findings of this proof- of-principle study. A higher sequencing depth, in conjunction with shotgun metagenomic sequencing, may facilitate the identification of novel or rare (but nonetheless important) taxa that can be used to identify contamination at the species level. DNA markers unique to horse or pig manure are yet to be established.

The robust resolution of horse and pig-associated markers with next-generation sequencing represents opportunities for future investigation.

4.2 Fecal source identification in residential groundwater sites using 16S rRNA sequencing

The co-detection of animal and/or human associated markers in drinking water provides evidence of possible fecal contamination by that source. After identifying human and animal-associated markers in our reference fecal samples, we screened for possible fecal contamination in private well waters by profiling the well water microbial communities. The mean *Campylobacteraceae* relative abundance was highest in November and September, but lowest in August. The large standard deviations in *Campylobacteraceae* abundances within the sampling months suggest that some groundwater sites contain more fecal contamination than others. The relative abundance of *Campylobacteraceae* sequences still far exceeded the relative abundance of chicken and cow fecal markers, *Gallicola* and *Turicibacter*. While we cannot exclude the possibility of cow or chicken farm contamination in September and October, the higher mean relative abundance of *Campylobacteraceae* across the groundwater wells in every month except August raises the likelihood of STE contamination in the residential well water.

Interestingly, well waters collected in August contained the lowest abundance of the three host-associated markers. Except for well water site K, *E. coli* counts did not exceed 100 CFUs/100 mL (raw data), suggesting diffuse fecal contamination among the well water samples collected in August. In other sampling months, there was a far higher

abundance of *Campylobacteraceae* sequences in the autumn. The detection of *E. coli* counts could be due to leaking septic tanks from individual sites.

4.3 Quantification of human-based fecal contamination

To confirm the presence of STE contamination in selected well sites, we used the HF183 marker to quantify human-based contamination in DNA extracted from STE and groundwater samples. We first validated the use of the HF18 marker by preparing standard and melt curves of the HF183 qPCR assay. Although qPCR assays were conducted with 100 bp fragments to minimize spurious fluorescence, we observed high R^2 value and robust E-value, suggesting the quality of the amplification reactions (Figure 6S.1A). Furthermore, the melt curve analysis shows that a single product was detected by the qPCR assay (Figure 6S.1B). Furthermore, we detected the HF183 marker in the two STE sites at 10-fold higher concentration than groundwater samples. These results validate quantifying the entire HF183 amplicon in DNA extracted from the town's well water as a proxy of STE contamination.

We also detected the HF183 DNA marker in B and K, groundwater wells containing the highest *E. coli* CFUs. Curiously, groundwater wells I and F, which did not contain *E. coli*, were also positive for the HF183 marker. A weak correlation between *E. coli* counts and HF183 marker concentrations was established in residential areas (Nshimyimana et al., 2014) and the Great Lakes beach sands (Staley et al., 2015). The weak correlation (Pearson correlation; r = 0.33) between *E. coli* counts and HF183 marker concentrations indicates that *E. coli* levels is not a reliable indicator of human fecal contamination in well waters. This is primarily due to the factors that affect viability

of *E. coli* in well water and contamination of well waters through sources other than humans. Within Wainfleet, groundwater wells B and K may receive *E. coli* from Lake Erie where surface water flows into the aquifer. However, these groundwater wells also receive human contamination loads that may come from leaking septic tanks, evidenced by the positive HF183 detection in septic tanks and the wells.

Interestingly, the groundwater well sample collected from well A contained almost five times the HF183 marker concentration as groundwater wells B and K. Well A's homeowner installed a UV treatment system in residence to inactivate fecal microbes. While *E. coli* and *Enterococcus* were absent in the UV treated well water, the HF183 marker was still present in UV treated well water. This could be due to the presence of DNA from dead (or inactivated) *Bacteroidales* (inactivated through UV treatment) which was amplified during the qPCR assay. Other groundwater wells, like G and J, did not contain a detectable HF183 marker, removing the possibility of STE contamination in those wells. Culturable FIBs were also absent in G and J, corroborating the absence of STE-based contamination.

5. Conclusion

A prior study found extensive FIB contamination throughout a rural community (Niagara Region Report, 2007). In this study, we complemented cultured-based methods with 16S sequencing and qPCR to re-assess fecal contamination levels and, further, to identify potential contamination sources in Wainfleet. We found FIB contamination in the well waters we tested, with *E. coli* counts as high as 10^6 CFU/mL. In addition, the

identification of additional STE DNA markers to *Campylobacteraceae* and the human *Bacteroidales* is consistent with the idea that human waste may impact residential groundwater wells. The low abundance (and absence) of animal manure-associated DNA markers – *Turicibacter* in cow manure and *Gallicola* in chicken manure – reduces the possibility that observed contamination is due to animal sources. These results indicate that profiling microbial communities using 16S rRNA sequencing can augment culture-based methods in contamination analysis studies. The use of next-generation sequencing methods can specifically facilitate the assessment of groundwater quality by detecting host-associated markers and quantifying relative contributions of likely fecal sources to groundwater contamination.

To trace fecal contamination sources in water sources more robustly, amplicon sequencing at higher depths (more reads per sample) may be useful in identification of novel or rare taxa. Shotgun metagenomic sequencing may also be used which, at a higher depth, may facilitate the identification of fecal indicator bacteria at the species level. While deeper sequencing depth may facilitate species identification, such as the *Campylobacter* spp. that comprise STEs and possible pig and horse-associated markers within *Clostridium* and other genera, the cost of such in depth analyses are fairly high. Highly-focused DNA sampling programs that target well water sites having high FIB levels may allow detailed identification contamination inputs that include potential pathogens that, unlike *E. coli*, are difficult to culture.

Acknowledgement

We thank members of the Schellhorn Lab for their help and comments on the manuscript. We thank Trevor Imhoff (Wainfleet Township) for advice on study methodology. We gratefully acknowledge the financial support through the Niagara Region WaterSmart Program, the Natural Sciences and Engineering Research Council of Canada's (NSERC) Collaborative Research and Development Program (CRDP), and NSERC operating grants to HES. The authors declare no conflict of interest.

References

APHA, (2012) *Standard methods for the examination of water and wastewater*. American Public Health Association, Washington D.C.

Bernhard, A.E. & K.G. Field, (2000) A PCR assay to discriminate human and ruminant feces on the basis of host differences in *Bacteroides-Prevotella* genes encoding 16S rRNA. *Applied and Environmental Microbiology* 66: 4571-4574.

Bokulich, N.A., S. Subramanian, J.J. Faith, D. Gevers, J.I. Gordon, R. Knight, *et al.*, (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Methods* 10: 57-59.

Cao, Y., L.C. Van De Werfhorst, E.A. Dubinsky, B.D. Badgley, M.J. Sadowsky, G.L. Andersen, *et al.*, (2013) Evaluation of molecular community analysis methods for discerning fecal sources and human waste. *Water Res* 47: 6862-6872.

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

Edgar, R.C., (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460-2461.

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

Harwood, V.J., C. Staley, B.D. Badgley, K. Borges & A. Korajkic, (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.

Health-Canada, (2013) Guidance on the use of the microbiological drinking water quality guidelines. In. W.a.A.Q. Bureau (ed). Ottawa, Ontario, pp.

Howard, G., S. Pedley, M. Barrett, M. Nalubega & K. Johal, (2003) Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda. *Water Res* 37: 3421-3429.

Kim, M., M. Morrison & Z. Yu, (2011) Status of the phylogenetic diversity census of ruminal microbiomes. *FEMS Microbiol Ecol* 76: 49-63.

Klindworth, A., E. Pruesse, T. Schweer, J. Peplies, C. Quast, M. Horn, *et al.*, (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 41: e1.

Niagara Region Report, 2007 Wainfleet Township – Lakeshore Area: Water and Wastewater Servicing; Final Report on Alternatives. Wainfleet, Ontario, Canada.

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

McLellan, S.L., R.J. Newton, J.L. Vandewalle, O.C. Shanks, S.M. Huse, A.M. Eren, *et al.*, (2013) Sewage reflects the distribution of human faecal *Lachnospiraceae*. *Environmental Microbiology* 15: 2213-2227.

Mohiuddin, M.M., S.R. Botts, A. Paschos & H.E. Schellhorn, (2018) Temporal and spatial changes in bacterial diversity in mixed use watersheds of the Great Lakes region. *Journal of Great Lakes Research*.

Mohiuddin, M.M., Y. Salama, H.E. Schellhorn & G.B. Golding, (2017) Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Res* 115: 360-369.

Nshimyimana, J.P., E. Ekklesia, P. Shanahan, L.H. Chua & J.R. Thompson, (2014) Distribution and abundance of human-specific *Bacteroides* and relation to traditional indicators in an urban tropical catchment. *J Appl Microbiol* 116: 1369-1383.

Seurinck, S., T. Defoirdt, W. Verstraete & S.D. Siciliano, (2005) Detection and quantification of the human-specific HF183 Bacteroides 16S rRNA genetic marker with real-time PCR for assessment of human faecal pollution in freshwater. *Environ Microbiol* 7: 249-259.

Staley, Z.R., E. Chase, C. Mitraki, T.L. Crisman & V.J. Harwood, (2013) Microbial water quality in freshwater lakes with different land use. *J Appl Microbiol* 115: 1240-1250.

Staley, Z.R., L. Vogel, C. Robinson & T.A. Edge, (2015) Differential occurrence of *Escherichia coli* and human *Bacteroidales* at two great lakes beaches. *Journal of Great Lakes Research* 41: 530-535.

Tomaras, J., J.W. Sahl, R.L. Siegrist & J.R. Spear, (2009) Microbial diversity of septic tank effluent and a soil biomat. *Applied and Environmental Microbiology* 75: 3348-3351.

Unno, T., J. Jang, D. Han, J.H. Kim, M.J. Sadowsky, O.-S. Kim, *et al.*, (2010) Use of barcoded pyrosequencing and shared OTUs to determine sources of fecal bacteria in watersheds. *Environmental science & technology* 44: 7777-7782.

USEPA, (1993) Wellhead protection: a guide for small communities. In.: USEPA, Office of Research and Development Office for Water Washington, DC, pp.

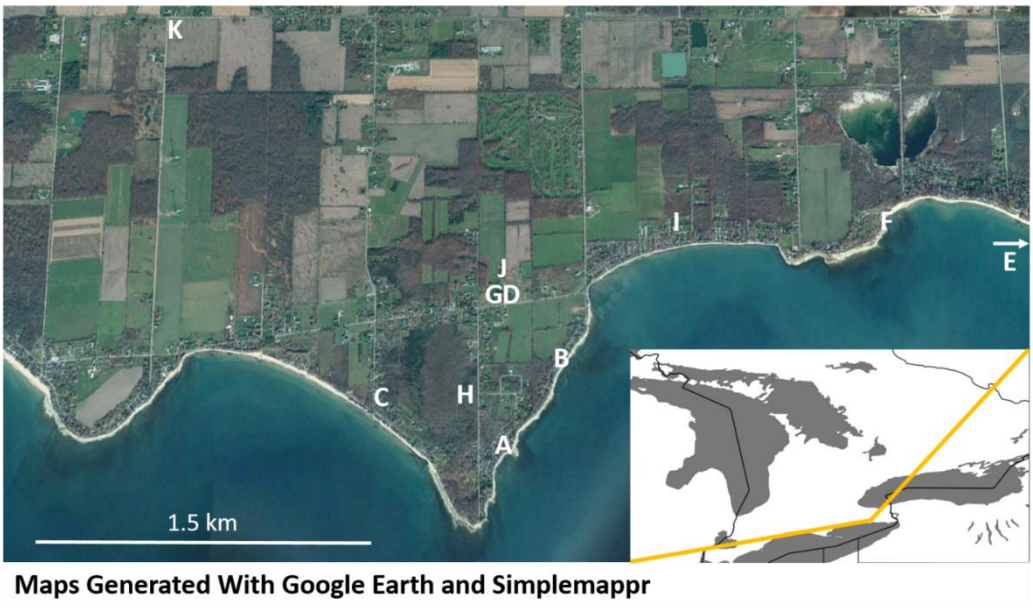

Figure 6.1 Location of sampling sites.

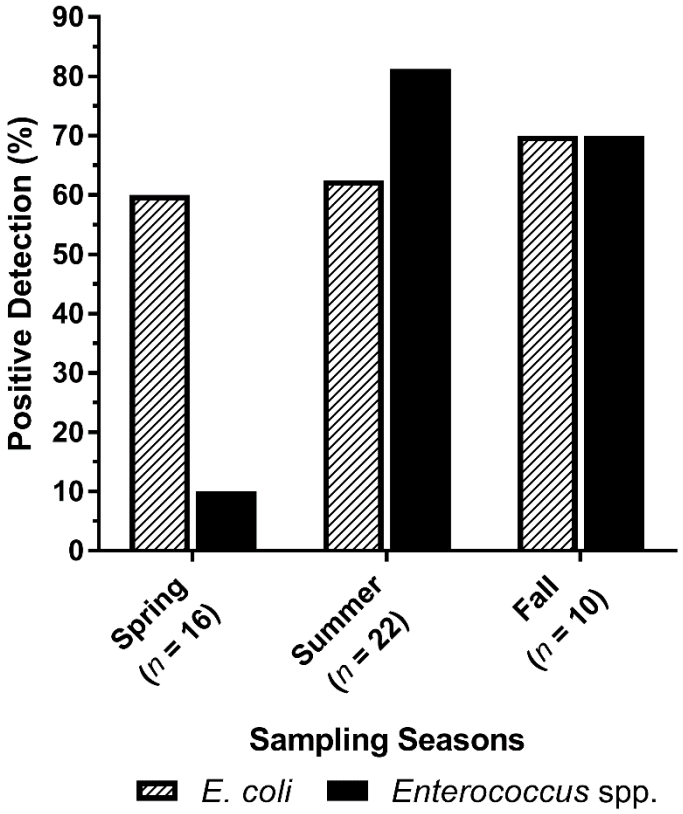

Figure 6. 2 Detection of *E. coli* and *Enterococcus* in well water samples (*n* = 48) at locations in the boil water advisory zone in 2015.

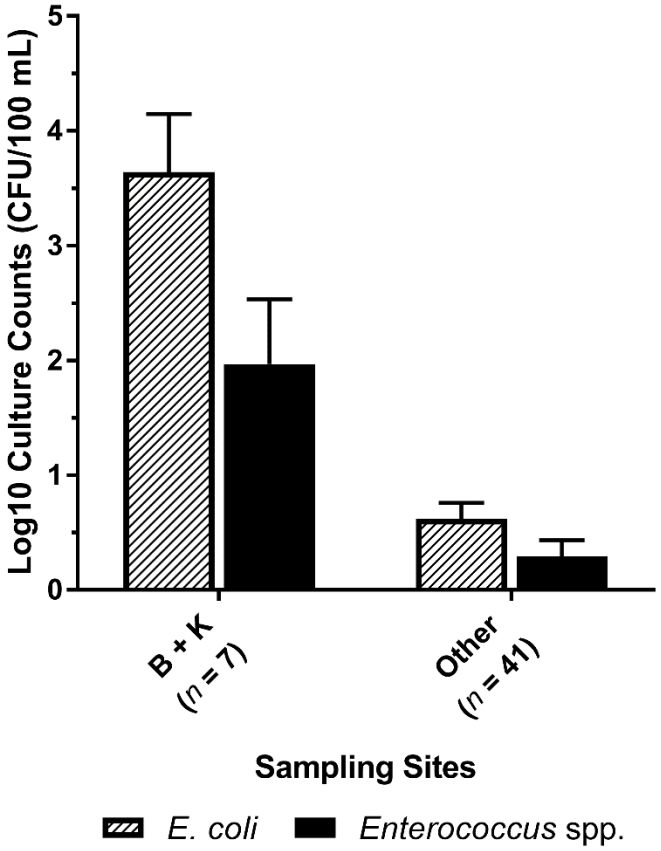

Figure 6.3 *E. coli* and *Enterococcus* levels in wells waters B and K compared with the other well water sites. Error bars represent standard deviation of the mean.

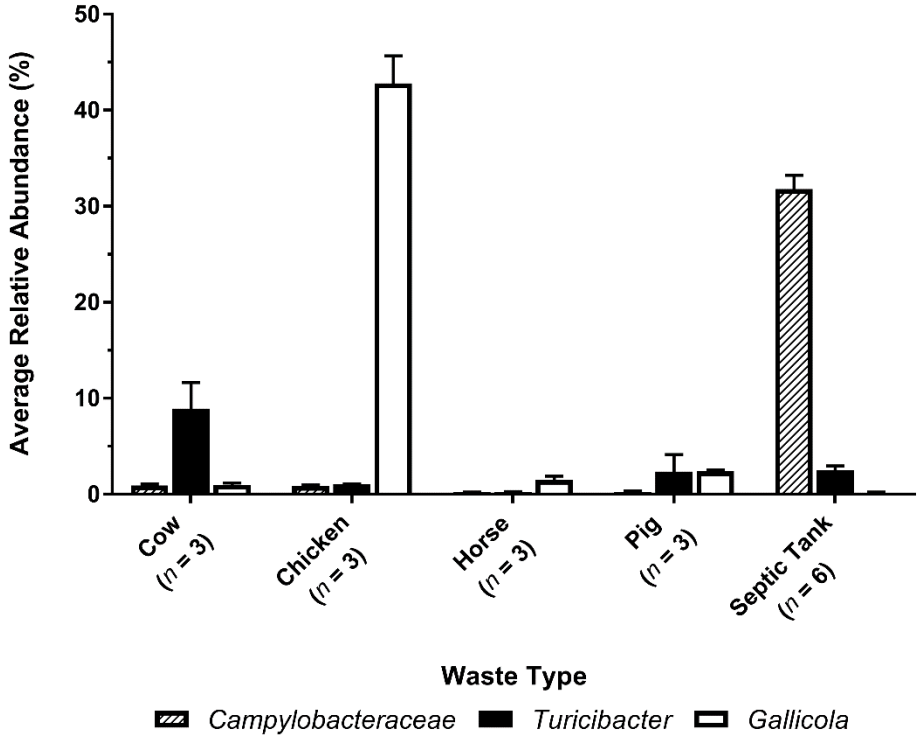

Figure 6.4 Relative abundance of potential manure and sewage markers in waste samples collected in Wainfleet. Error bars represent standard deviation of the mean. Three biological replicates for each of the sample was used to calculate mean and standard deviation.

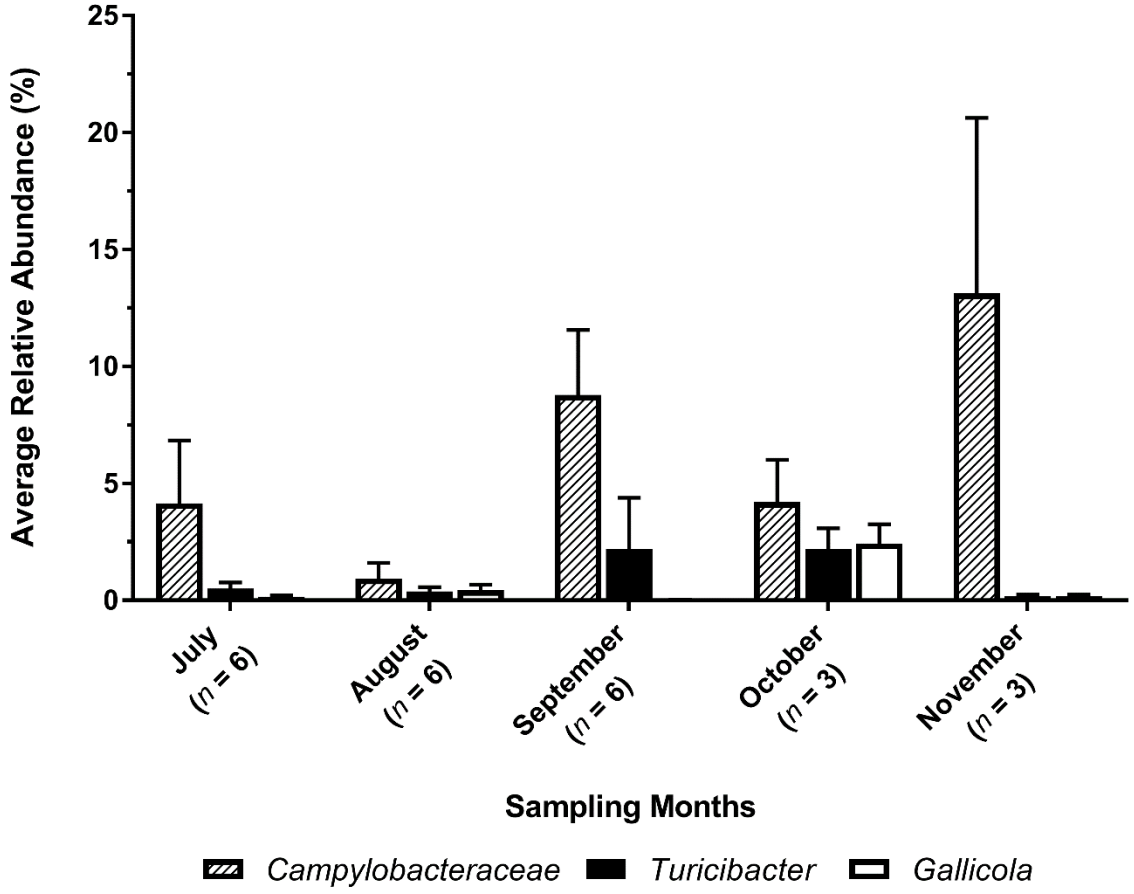

Figure 6.5 Relative abundance of potential STE and animal-specific contamination markers in well water sites within the active boil water advisory zone. Error bars represent standard deviation of the mean.

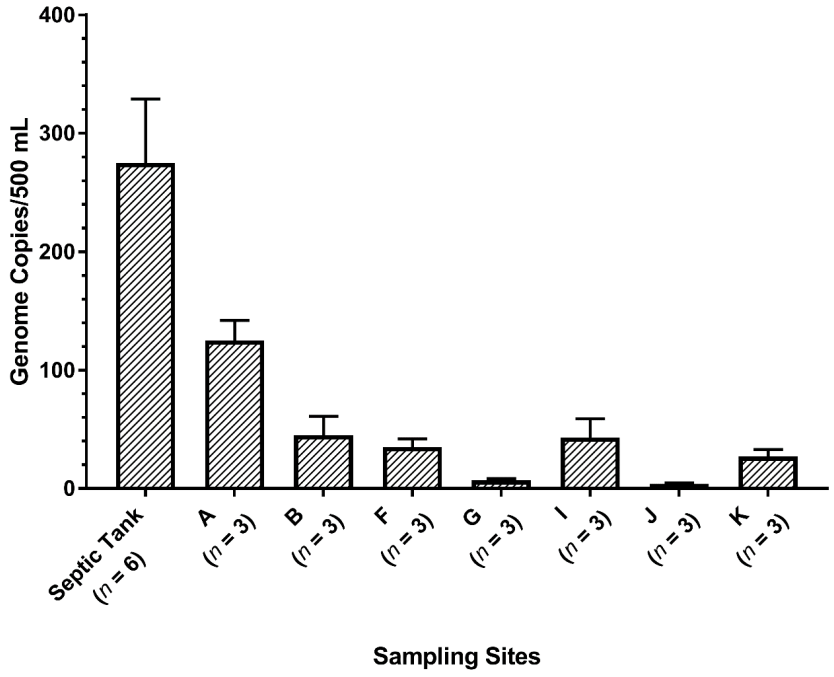

Figure 6.6 Presence of human *Bacteroidales* marker in selected septic tank and groundwater wells within the boil water advisory zone in Wainfleet. Error bars represent standard deviation of the mean.

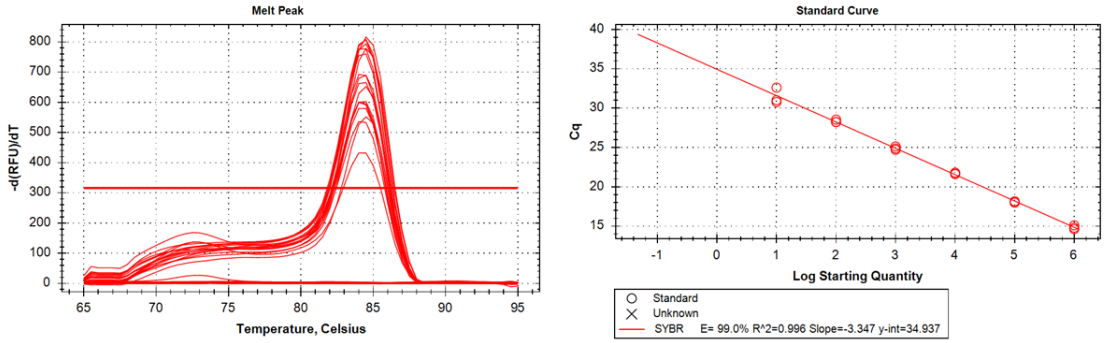

Figure 6S.1 Standard curve (A) and melt curve (B) from qPCR of HF183 gene. (A) High R^2 value and robust E value indicate the high quality of the amplification reaction and (B) melt curve shows the specificity of the amplification reaction.

Chapter 7 / Appendix B: Epidemiology and ecology of emerging viruses in two freshwater lakes of the Northern Hemisphere

Mahi M. Mohiuddin, Herb E. Schellhorn*
Department of Biology, McMaster University, Hamilton, ON, Canada

***Correspondence:** Herb E. Schellhorn, LSB 433, Department of Biology, McMaster University, 1280 Main St W, Hamilton, ON, L8S 4K1, Canada.
E-mail: schell@mcmaster.ca

Reproduced with permission from Mohiuddin, and H.E. Schellhorn, (2019) Epidemiology and ecology of emerging viruses in two freshwater lakes of the Northern Hemisphere. *In* Emerging and Re-emerging Viral Pathogens (ed. M. M. Ennaji). Elsevier B.V.

Abstract

Changing patterns of waterborne viral diseases across the globe have been the focus of many studies in recent years. Despite the recent increase in recreational water-related diseases, studies performed in the Great Lakes region have primarily focused on the identification of fecal indicator microorganisms or select pathogenic viruses. These studies, although provide useful information in determining sources of fecal contamination of aquatic environments, do not provide a comprehensive profile of viral species inhabiting this region. Understanding the dynamics and the diversity of total viral communities is essential in assessing the impact of these communities on water quality and the risk they pose to public health. Therefore, in this chapter, we have summarized the current understanding of viral community structures of the lower Great Lakes area, particularly of Lake Ontario and Lake Erie, the epidemiology and ecology of pathogens, their sources, and the public health implications of these findings. The information obtained from this review may aid in our understanding of recreational water quality and may provide useful information to water quality decision makers for the optimization of municipal sampling programs currently employed in the Great Lakes area.

1. Introduction

Viruses are the most abundant microorganisms on earth (Bergh et al., 1989) and play important roles in global biogeochemical and ecological cycles (Roux et al., 2016). Because of their ubiquity and environmental complexity, the role of viruses in aquatic environments has been the focus of many investigations in recent years (Brum et al., 2015;Dell'Anno et al., 2015;Sweet and Bythell, 2017). However, the majority of such investigations have focused mostly on marine environments, while freshwater habitats have remained largely unexplored. Studies performed in freshwater habitats indicate that, freshwater viral communities, although exhibit less diversity compared to marine viruses, likely contain diverse and novel viruses (Djikeng et al., 2009;Hewson et al., 2012;Roux et al., 2012).

The majority of the freshwater viruses identified are bacteriophages, human and animal viruses and, to a lesser extent, some plant viruses (Djikeng et al., 2009;Fancello et al., 2013;Tseng et al., 2013). Identification of viruses from the environment poses a challenge, as most viruses are not culturable through conventional techniques. Unlike 16S rRNA of prokaryotes and 18S rRNA of eukaryotes, viruses lack universal conserved sequences. This lack of a universal phylogenetic marker limits their direct detection from the environment (Rohwer and Edwards, 2002). Metagenomic approach, an alternative to culture-dependent techniques, does not rely on the presence of any shared gene between viruses and since its first application for the identification of viruses at two marine locations (Breitbart et al., 2002), metagenomic approach has been the primary technique used to identify viruses. This approach relies on the extraction of DNA from mixed

microbial and viral communities and has expanded our knowledge of viral and bacterial diversity in many environments including soil, freshwater sources, marine sediments and the human gut (Qin et al., 2010;Reyes et al., 2010;Rodriguez-Brito et al., 2010).

The Great Lakes make up the largest freshwater reservoir in North America and the second largest in the world (Waples et al., 2008). These lakes are of longstanding interest, as they provide essential services including drinking water, agricultural and industrial use, recreational activities and routes for transportation and, therefore, are closely monitored by public health and municipal authorities. Failure to properly monitor water quality and water-dependent agricultural activities may lead to serious economic and human health consequences as suggested in Walkerton tragedy (Brown and Hussain, 2003) and the *E. coli* outbreak in Germany (Bielaszewska et al., 2011).

The Great Lakes basin harbors many beaches, approximately 822 of which are monitored by local municipal agencies because of their frequent closure due to high *E. coli* counts , yet high recreational usage during the summer months (Ulrich, 2009). Examination of freshwater beaches of the Great Lakes, particularly of the lower Great Lakes region, provides evidence for the presence of pathogenic viruses (Fong et al., 2005;Fong et al., 2007;Mohiuddin and Schellhorn, 2015). Despite this evidence, very little is known about the viral communities inhabiting this region. Therefore, a comprehensive knowledge of bacterial and viral community composition, structure and ecology is required to assess the presumed risk that these microorganisms pose to public health. As a result, we have compiled the information available from the lower Great

Lakes region, specifically Lake Ontario and Lake Erie, with the aim of providing a better understanding of viral communities inhabiting this area and demonstrating the diverse nature of these freshwater microorganisms.

2. Viral community structure

Viruses, the most abundant microorganisms in aquatic environments, infect cellular organisms and viral infections kill approximately 10-20% of cellular organisms each day (Suttle, 2007;Evans and Brussaard, 2012). Viruses have also been associated with a number of waterborne disease outbreaks in the Great Lakes region (Fong et al., 2007;Hlavsa et al., 2011). In addition, viruses have low infectious dose (Yezli and Otter, 2011;Atmar et al., 2014) and can persist in the environment for months (Kotwal and Cannon, 2014;Prevost et al., 2016). Because of their low infectious dose, persistence in water and association with recreational water related disease outbreaks, viruses pose a serious threat to public health. Therefore, continuous research on viral epidemiology and ecology is required for better understanding the dynamics of viruses in aquatic environments and associated viral risk. However, unlike bacterial communities, characteristics of viral communities as well as their impact on the environment did not receive much attention in Lake Ontario and Lake Erie. The majority of the studies performed in this region have focused primarily on the identification of either coliphages (somatic and F+-specific) (Fong et al., 2007), the viral indicator of fecal contamination, or select pathogenic viruses (Payment and Locas, 2011;Wu et al., 2011). More recently,

shotgun metagenomic sequencing approach was used to identify a wide range of viruses from both Lake Ontario and Lake Erie (Mohiuddin and Schellhorn, 2015).

2.1 Viral pathogens

While viruses are abundant in aquatic environments, not all viruses are pathogenic to humans. Viruses that cause gastrointestinal infections are frequently found in fecal contaminated water. Although viruses cannot replicate outside their host, they can persist in the environment for a long period without significant loss of infectivity and their high rate and ease of transmissibility and capacity to cause infection at a very low concentration pose a significant threat to public health. Viruses that are mainly responsible for waterborne diseases, includes Enteroviruses, Adenoviruses and Caliciviruses (Noroviruses) (Abbaszadegan et al., 1999;Rutjes et al., 2006;Xagoraraki et al., 2007). These viruses are transmitted via fecal-oral route and present a great risk of infection to both humans and animals. Both human and animal feces excrete a large number of pathogenic viruses and these viruses are transported to the environment through malfunctioning sewage treatment plants, leakage of sewage systems, untreated waste water (combined sewer overflow), river water and ground water.

The majority of the viruses identified in the lower Great Lakes region, particularly in Lake Ontario and Lake Erie, are bacteriophages. These viruses (phages) kill bacterial cells and play important roles in global biogeochemical and ecological cycles (Fuhrman, 1999;Suttle, 2007). In addition to bacteriophages, viruses including phycodnaviruses, cyanophages, noroviruses and other human human enteroviruses have also been identified

in this region (Greer et al., 2009;Short and Short, 2009;Short et al., 2011;Edge et al., 2013). More recently, using shotgun metagenomic sequencing, sequences originating from an array of human and animal viruses including adenoviruses, poxviruses and herpesviruses have been identified in Lake Ontario and Lake Erie (Mohiuddin and Schellhorn, 2015). The presence of these viruses in both Lake Ontario and Lake Erie suggests that both these lakes serve as a natural reservoir for many pathogenic viruses and therefore, requires constant monitoring. A list of potential pathogenic viruses identified in the lower Great Lake area are included in Table 7.1.

2.1.1 Adenoviruses

Adenoviruses are mainly responsible for mild infections including gastroenteritis, respiratory diseases (tonsillitis, pharyngitis, otitis media and bronchiolitis/bronchitis) involving upper or lower respiratory tract, conjunctivitis, encephalitis and urinary tract diseases. Although to date 52 different serotypes and 7 different subgroups (A through G) have been described, the majority of the waterborne gastroenteritis diseases is caused by serotype 40 and 41 under subgroup F (van Heerden et al., 2005). Adenoviruses are transmitted through the fecal-oral route, and the viral load in the feces of infected individuals are very high (10^6 viral particles/gm of fecal matter) (Jiang, 2006). Transmission occurs through direct contact with contaminated objects, particularly recreational waters, drinking water and with contaminated food. Using shotgun metagenomic sequencing approach, sequences originating from both human and animal

adenoviruses were identified in Lake Ontario and Lake Erie (Mohiuddin and Schellhorn, 2015).

2.1.2 Noroviruses

Noroviruses, previously known as Norwalk viruses, are the primary cause of non-bacterial gastroenteritis. Norovirus infection occurs throughout the year, though the incidence of infection increases during the winter (Mounts et al., 2000). Contaminated water is the primary source of infection (Maunula et al., 2005) and these viruses are transmitted mainly through the fecal-oral route. Infection can occur through the intake of contaminated water, consumption of contaminated food or contact with contaminated environmental surfaces and infected individuals. Norovirus infection occurs in people of all ages and infection is characterized by the onset of vomiting or diarrhea or both. Symptoms may also include nausea, mild fever, abdominal pain, muscle aches and mild fever. Norovirus infections are self-limiting, and symptoms develop within 24-48 hours of infection and typically last from 2-3 days (Graham et al., 1994). However, infected individuals excrete a large number of viral particles ranging from 10^7 to 10^8 copies per gram of stool (Chan et al., 2006) and shedding can continue for more than one month for patients with acute gastroenteritis (Murata et al., 2007). Noroviruses are present in watersheds of the Lake Ontario region (Greer et al., 2009).

2.1.3 Other human enteroviruses

Association between recreational water related outbreaks and other human enteroviruses, including Coxsackieviruses and Echoviruses have also been identified

(Sinclair et al., 2009). These viruses are present in recreational beaches of Lake Michigan and many other freshwater beaches across the globe (Xagoraraki et al., 2007;Sinclair et al., 2009;Aslan et al., 2011). While human enteric viruses were identified at the offshore intakes of Lake Ontario, the types of viruses identified were not specified (Edge et al., 2013). In a separate experiment, using coliphages as an indicator, the presence of human enteroviruses was also confirmed in Lake Erie (Fong et al., 2007).

2.2 Cyanophages

Cyanophages, although not directly pathogenic to humans, are present in Lake Ontario and Lake Erie (Matteson et al., 2011;Mohiuddin and Schellhorn, 2015). The majority of the cyanophages identified are Synechococcus phages, followed by Prochlorococcus phages. Cyanophages prey on Cyanobacteria, which are responsible for harmful algal blooms (HABs) (Cheung et al., 2013). Therefore, monitoring their abundance in freshwater lakes may provide important insight into the cyanobacteria-cyanophage relationship. This, in turn, may provide useful information for controlling HABs in the lower Great Lakes watershed.

2.3 Phycodnaviruses

Phycodnaviruses are also present in both Lake Ontario and Lake Erie (Short and Short, 2009;Mirza et al., 2015;Mohiuddin and Schellhorn, 2015). Although Phycodnaviruses are generally considered to exclusively infect algal species, recent findings suggest that Chlorovirus Acanthocystis turfacea virus 1 (ATCV-1) belonging to

the Phycodnaviridae family, can infect humans and is associated with diminished cognitive function (Yolken et al., 2014).

2.4 Viral Hemorrhagic Septicemia virus

Viral Hemorrhagic Septicemia virus (VHSv), one of the most serious fish pathogens that kills over 80 species of marine and freshwater finfish (Faisal et al., 2012), have also been identified from the Great Lakes region including Lake Ontario and Lake Erie (Thompson et al., 2011;Cornwell et al., 2012). VHSv is widely abundant in the Laurentian Great Lakes basin and understanding the genetics, epidemiology, and ecology of this virus may provide useful information to the surveillance programs employed in this region.

2.5 Giant viruses

Using shotgun metagenomic sequencing approach, sequences originating from giant viruses (also known as nucleocytoplasmic large DNA viruses or NCLDVs) belonging to the families *Marseilleviridae* and *Mimiviridae* have also been identified in the lower Great Lakes region (Mohiuddin and Schellhorn, 2015). Compared to other viruses, giant viruses possess extremely large genomes and often carries genes that ae common in bacteria and eukaryotes (Arslan et al., 2011;Aherfi et al., 2014). While the giant viruses mostly infect Amoeba (Philippe et al., 2013;Aherfi et al., 2016), more recently, these viruses have been identified in patients with unexplained pneumonia (Colson et al., 2016) and lymph node adenitis (Popgeorgiev et al., 2013).

2.6 Other viruses

Similar to other aquatic environments, the majority of the viruses identified in the lower Great Lakes region are bacteriophages belonging to the families *Myoviridae*, *Podoviridae* and *Siphoviridae* (Mohiuddin and Schellhorn, 2015). In addition to bacteriophages, viruses infecting animals and insects have also been identified in Lake Ontario and Lake Erie. These include iridoviruses, poxviruses, alloherpesviruses, and baculoviruses (Table 7.1) (Mohiuddin and Schellhorn, 2015).

3. Sources of pathogenic viruses

Studies investigating potential pathogens in recreational beaches have indicated that pathogenic viral species can be introduced into aquatic environments through several sources. These include point-source (wastewater) (Aslan et al., 2011;Tamaki et al., 2012) and non-point-source pollution (Bae and Wuertz, 2012), direct fecal discharge from humans and other animals, (Wright et al., 2009;Ahmed et al., 2010). Multiple microbial source tracking methods are used to investigate the source of fecal indicator organisms as well as some pathogens in the Great Lakes area (Fong et al., 2007).

Malfunctioning wastewater treatment facilities and leaking septic tanks were associated with the presence of noroviruses in Lake Erie (Fong et al., 2007). In addition, runoffs from agricultural farms, wastewater effluents and combined sewer overflow (CSO) were responsible for adenovirus and other enteric virus occurrence in rivers (Hundesa et al., 2006). Ballast water of commercial vessels has also been well-established as another vehicle for and /or source of pathogens in aquatic environments including the

Great Lakes (Grigorovich et al., 2003;Drake et al., 2007;Laboratory, 2010). VHSv, a fish pathogen responsible for killing many finfish in North America, is a recent example of pathogenic microorganism introduced into the Great Lakes through ballast water (Elsayed et al., 2006;Bain et al., 2010).

4. Current approaches for the identification of pathogens, their limitations and scope

Pathogenic bacteria and viruses are not directly measured in recreational water since there is a wide range of possible agents and detection methods vary from pathogen to pathogen. In addition, simultaneous identification of a large number of pathogens would be time consuming and expensive. Therefore, fecal indicator microorganisms including *E. coli*, *Enterococcus* spp., *Bacteroide*s spp. and F-specific RNA coliphages are used as proxy for pathogen occurrence in freshwater environments and identification of fecal indicators is the primary choice of method for pathogen identification in the lower Great Lakes area. Using culture-based approaches, *E. coli*, *Enterococcus* spp. and coliphages were identified in both Lake Ontario and Lake Erie (Edge and Hill, 2007;Fong et al., 2007;Edge et al., 2013). While fecal indicators are useful for predicting pathogen occurrence, their concentration is rarely predictive of individual pathogens (Payment and Locas, 2011;Wu et al., 2011). In addition, indicator microorganisms are not always reliable because of the differences in properties between indicator microorganisms and target pathogens. For example, indicator microorganisms are inactivated effectively through the water treatment processes (such as UV irradiation, chlorination treatment etc.) whereas human viral and protozoan parasites are not inactivated efficiently through these processes (40, 41) indicating that some pathogens are more persistent in the environment than indicator organisms. To circumvent the limitations of these culture-based methods, molecular techniques such as PCR and qPCR, have been developed in recent years and are being used regularly to detect individual pathogens. More recently,

because of the rapid development and cost reduction of sequencing techniques, metagenomic studies are being used regularly to identify microorganisms from the environment.

Traditional fecal indicators including *E. coli*, *Enterococcus* spp. and coliphages, are commonly found in the intestine of birds, humans and other animals. Therefore, the presence of these indicators does not provide information regarding the source of fecal contamination and emphasize the need for developing alternative indicators which can distinguish human fecal contamination (or wastewater contamination) from fecal contamination by birds or other animals. One such indicator organism is *Bacteroidales* and use of human specific *Bacteroidales* makers such as HF183 can identify human fecal contamination in aquatic environments (Seurinck et al., 2005). *Bacteroidales* DNA markers have also been used to identify the source of fecal contamination in Lake Ontario (Edge et al., 2010;Edge et al., 2013).

Another approach to identify the source of fecal contamination as well as the type of pathogen present is the sequence-based approach. Both shotgun metagenomic sequencing and amplicon sequencing are used to identify the characteristics of bacterial and viral populations inhabiting the Great Lakes area (Rinta-Kanto and Wilhelm, 2006;Hotto et al., 2007;Bouzat et al., 2013;Mohiuddin and Schellhorn, 2015;Mohiuddin et al., 2017b). However, due to the absence of any signature sequence in viruses (Rohwer and Edwards, 2002), amplicon sequencing which relies on the presence of signature sequences, cannot be applied for the detection of viruses.

The majority of the metagenomic studies use relative abundance of bacterial and viral taxa as an estimate of microbial community composition. Relative abundance is defined as the relative amounts of different taxa present within an environment. Relative abundance estimates are useful for identifying the characteristics of microbial and viral communities but do not provide any information regarding the true abundance of pathogens. An alternative parameter to consider is absolute abundance which quantifies the absolute quantities of taxa present in a sample and can be useful for identifying the true abundance of pathogens. However, factors including the sequencing technology used, and the primer pairs used to target the hypervariable region of the 16S rRNA gene may introduce biases in the measurement of microbial community abundances (Tremblay et al., 2015;Tessler et al., 2017) and therefore, sequencing data alone cannot be used to estimate absolute abundance (Nayfach and Pollard, 2016). Sequence based approaches, if complemented with culture-based and/or qPCR-based approaches, may provide an accurate estimation of the true abundance of pathogens.

While coliphages are used as a viral indicator of fecal contamination and are routinely used in the lower Great Lake area for monitoring water quality (Fong et al., 2007), reverse transcriptase-polymerase chain reactions (RT-PCR) (Greer et al., 2009) and immunoperoxidase methods (Edge et al., 2013) can also be used to identify noroviruses and cultivable human enteric viruses, respectively. More recently, metagenomic approaches are being used to identify viruses in both Lake Ontario and Lake Erie (Matteson et al., 2011;Short et al., 2011;Mohiuddin and Schellhorn, 2015).

Compared to bacteria, detection of viruses is difficult since the majority of the viruses are not culturable and lack universal genetic marker. Therefore, targeted amplification and shotgun metagenomic sequencing remain the only choice of identification for these pathogens.

5. Summary and Conclusion

Here we have reviewed the diverse nature of viral communities of the lower Great Lakes region. We reviewed studies showing that freshwater environments serve as a reservoir of many viral pathogens, which pose potential threat to public health, may lead to the impairment of recreational opportunities and could be harmful to the economy. Although wastewater contamination is primarily responsible for the presence of many of these pathogens in recreational water, ballast water and beachgoers also contribute to the contamination of these aquatic environments. Currently there are many surveillance programs being employed in North America to monitor recreational water quality. However, the majority use source-tracking methods, which are unable to identify the types of pathogens present in water. While sequencing-based approaches provide information of total microbial and viral communities, routine use of such approaches is not feasible due to their high cost and the time required for analysis. However, information obtained from sequencing-based approaches can be used to complement these traditional monitoring programs. Developing sequence-based identification methods for pathogens and using adequate quantification standards will improve the sensitivity of pathogen detection and therefore, will augment the existing monitoring programs.

References

Abbaszadegan, M., P. Stewart & M. LeChevallier, (1999) A strategy for detection of viruses in groundwater by PCR. *Appl Environ Microbiol* **65**: 444-449.

Aherfi, S., P. Colson, B. La Scola & D. Raoult, (2016) Giant Viruses of Amoebas: An Update. *Frontiers in Microbiology* **7**.

Aherfi, S., B. La Scola, I. Pagnier, D. Raoult & P. Colson, (2014) The expanding family Marseilleviridae. *Virology* **466-467**: 27-37.

Ahmed, W., A. Goonetilleke & T. Gardner, (2010) Human and bovine adenoviruses for the detection of source-specific fecal pollution in coastal waters in Australia. *Water Res* **44**: 4662-4673.

Arslan, D., M. Legendre, V. Seltzer, C. Abergel & J.M. Claverie, (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences* **108**: 17486-17491.

Aslan, A., I. Xagoraraki, F.J. Simmons, J.B. Rose & S. Dorevitch, (2011) Occurrence of adenovirus and other enteric viruses in limited-contact freshwater recreational areas and bathing waters. *Journal of Applied Microbiology* **111**: 1250-1261.

Atmar, R.L., A.R. Opekun, M.A. Gilger, M.K. Estes, S.E. Crawford, F.H. Neill, *et al.*, (2014) Determination of the 50% Human Infectious Dose for Norwalk Virus. *The Journal of Infectious Diseases* **209**: 1016-1022.

Bae, S. & S. Wuertz, (2012) Survival of Host-Associated Bacteroidales Cells and Their Relationship with Enterococcus spp., Campylobacter jejuni, Salmonella enterica Serovar Typhimurium, and Adenovirus in Freshwater Microcosms as Measured by Propidium Monoazide-Quantitative PCR. *Applied and Environmental Microbiology* **78**: 922-932.

Bain, M.B., E.R. Cornwell, K.M. Hope, G.E. Eckerlin, R.N. Casey, G.H. Groocock, *et al.*, (2010) Distribution of an invasive aquatic pathogen (viral hemorrhagic septicemia virus) in the Great Lakes and its relationship to shipping. *PLoS ONE* **5**.

Bergh, O., K.Y. Borsheim, G. Bratbak & M. Heldal, (1989) High abundance of viruses found in aquatic environments. *Nature* **340**: 467-468.

Bielaszewska, M., A. Mellmann, W. Zhang, R. Kock, A. Fruth, A. Bauwens, *et al.*, (2011) Characterisation of the Escherichia coli strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect Dis* **11**: 671-676.

Bouzat, J.L., M.J. Hoostal & T. Looft, (2013) Spatial patterns of bacterial community composition within Lake Erie sediments. *Journal of Great Lakes Research* **39**: 344-351.

Breitbart, M., P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, *et al.*, (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**: 14250-14255.

Brown, R.S. & M. Hussain, (2003) The Walkerton tragedy—issues for water quality monitoring. *The Analyst* **128**: 320-322.

Brum, J.R., J.C. Ignacio-Espinoza, S. Roux, G. Doulcier, S.G. Acinas, A. Alberti, *et al.*, (2015) Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**: 1261498.

Chan, M.C., J.J. Sung, R.K. Lam, P.K. Chan, N.L. Lee, R.W. Lai, *et al.*, (2006) Fecal viral load and norovirus-associated gastroenteritis. *Emerg Infect Dis* **12**: 1278-1280.

Cheung, M.Y., S. Liang & J. Lee, (2013) Toxin-producing cyanobacteria in freshwater: A review of the problems, impact on drinking water safety, and efforts for protecting public health. In: Journal of Microbiology. pp. 1-10.

Colson, P., S. Aherfi, B. La Scola & D. Raoult, (2016) The role of giant viruses of amoebas in humans. *Curr Opin Microbiol* **31**: 199-208.

Cornwell, E.R., G.E. Eckerlin, T.M. Thompson, W.N. Batts, R.G. Getchell, G.H. Groocock, *et al.*, (2012) Predictive factors and viral genetic diversity for viral hemorrhagic septicemia virus infection in Lake Ontario and the St. Lawrence River. *Journal of Great Lakes Research* **38**: 278-288.

Dell'Anno, A., C. Corinaldesi & R. Danovaro, (2015) Virus decomposition provides an important contribution to benthic deep-sea ecosystem functioning. *Proc Natl Acad Sci U S A* **112**: E2014-2019.

Djikeng, A., R. Kuzmickas, N.G. Anderson & D.J. Spiro, (2009) Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**: e7264.

Drake, L.A., M.A. Doblin & F.C. Dobbs, (2007) Potential microbial bioinvasions via ships' ballast water, sediment, and biofilm. *Marine Pollution Bulletin* **55**: 333-341.

Edge, T.A. & S. Hill, (2007) Multiple lines of evidence to identify the sources of fecal pollution at a freshwater beach in Hamilton Harbour, Lake Ontario. *Water Research* **41**: 3585-3594.

Edge, T.A., S. Hill, P. Seto & J. Marsalek, (2010) Library-dependent and library-independent microbial source tracking to identify spatial variation in faecal contamination sources along a Lake Ontario beach (Ontario, Canada). *Water Science and Technology* **62**: 719-727.

Edge, T.A., I.U.H. Khan, R. Bouchard, J. Guo, S. Hill, A. Locas, *et al.*, (2013) Occurrence of waterborne pathogens and escherichia coli at offshore drinking water intakes in lake Ontario. *Applied and Environmental Microbiology* **79**: 5799-5813.

Elsayed, E., M. Faisal, M. Thomas, G. Whelan, W. Batts & J. Winton, (2006) Isolation of viral haemorrhagic septicaemia virus from muskellunge, Esox masquinongy (Mitchill), in Lake St Clair, Michigan, USA reveals a new sublineage of the North American genotype. *Journal of Fish Diseases* **29**: 611-619.

Evans, C. & C.P. Brussaard, (2012) Regional variation in lytic and lysogenic viral infection in the Southern Ocean and its contribution to biogeochemical cycling. *Applied and Environmental Microbiology* **78**: 6741-6748.

Faisal, M., M. Shavalier, R.K. Kim, E.V. Millard, M.R. Gunn, A.D. Winters, *et al.*, (2012) Spread of the emerging viral hemorrhagic septicemia virus strain, genotype IVb, in Michigan, USA. *Viruses* **4**: 734-760.

Fancello, L., S. Trape, C. Robert, M. Boyer, N. Popgeorgiev, D. Raoult, *et al.*, (2013) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *ISME J* **7**: 359-369.

Fong, T.T., D.W. Griffin & E.K. Lipp, (2005) Molecular Assays for Targeting Human and Bovine Enteric Viruses in Coastal Waters and Their Application for Library-Independent Source Tracking. *Applied and Environmental Microbiology* **71**: 2070-2078.

Fong, T.T., L.S. Mansfield, D.L. Wilson, D.J. Schwab, S.L. Molloy & J.B. Rose, (2007) Massive microbiological groundwater contamination associated with a waterborne outbreak in Lake Erie, South Bass Island, Ohio. *Environmental Health Perspectives* **115**: 856-864.

Fuhrman, J.A., (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541-548.

Graham, D.Y., X. Jiang, T. Tanaka, A.R. Opekun, H.P. Madore & M.K. Estes, (1994) Norwalk virus infection of volunteers: new insights based on improved assays. *J Infect Dis* **170**: 34-43.

Greer, A.L., S.J. Drews & D.N. Fisman, (2009) Why "Winter" vomiting disease? seasonality, hydrology, and norovirus epidemiology in Toronto, Canada. *EcoHealth* **6**: 192-199.

Grigorovich, I.A., R.I. Colautti, E.L. Mills, K. Holeck, A.G. Ballert & H.J. MacIsaac, (2003) Ballast-mediated animal introductions in the Laurentian Great Lakes: retrospective and prospective analyses. *Canadian Journal of Fisheries and Aquatic Sciences* **60**: 740-756.

Hewson, I., J.G. Barbosa, J.M. Brown, R.P. Donelan, J.B. Eaglesham, E.M. Eggleston, *et al.*, (2012) Temporal dynamics and decay of putatively allochthonous and autochthonous viral genotypes in contrasting freshwater lakes. *Applied and Environmental Microbiology* **78**: 6583-6591.

Hlavsa, M.C., V.A. Roberts, A.R. Anderson, V.R. Hill, A.M. Kahler, M. Orr, *et al.*, (2011) Surveillance for waterborne disease outbreaks and other health events associated with recreational water --- United States, 2007--2008. *Morbidity and mortality weekly report. Surveillance summaries (Washington, D.C. : 2002)* **60**: 1-32.

Hotto, A.M., M.F. Satchwell & G.L. Boyer, (2007) Molecular characterization of potential microcystin-producing cyanobacteria in Lake Ontario embayments and nearshore waters. *Applied and Environmental Microbiology* **73**: 4570-4578.

Hundesa, A., C. Maluquer De Motes, S. Bofill-Mas, N. Albinana-Gimenez & R. Girones, (2006) Identification of human and animal adenoviruses and polyomaviruses for determination of sources of fecal contamination in the environment. *Applied and Environmental Microbiology* **72**: 7886-7893.

Jiang, S.C., (2006) Human Adenoviruses in Water: Occurrence and Health Implications: A Critical Review†. *Environmental Science & Technology* **40**: 7132-7140.

Kotwal, G. & J.L. Cannon, (2014) Environmental persistence and transfer of enteric viruses. *Curr Opin Virol* **4**: 37-43.

Laboratory, G.L.E.R., (2010) Great Lakes Aquatic Nonindigenous Species Information System (GLANSIS). Ann Arbor: Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration. In., pp.

Matteson, A.R., S.N. Loar, R.A. Bourbonniere & S.W. Wilhelm, (2011) Molecular enumeration of an ecologically important cyanophage in a laurentian Great Lake. *Applied and Environmental Microbiology* **77**: 6772-6779.

Maunula, L., I.T. Miettinen & C.H. von Bonsdorff, (2005) Norovirus outbreaks from drinking water. *Emerg Infect Dis* **11**: 1716-1721.

Mirza, S.F., M.A. Staniewski, C.M. Short, A.M. Long, Y.V. Chaban & S.M. Short, (2015) Isolation and characterization of a virus infecting the freshwater algae Chrysochromulina parva. *Virology* **486**: 105-115.

Mohiuddin, M. & H.E. Schellhorn, (2015) Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* **6**: 960.

Mohiuddin, M.M., Y. Salama, H.E. Schellhorn & G.B. Golding, (2017) Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Research* **115**: 360-369.

Mounts, A.W., T. Ando, M. Koopmans, J.S. Bresee, J. Noel & R.I. Glass, (2000) Cold weather seasonality of gastroenteritis associated with Norwalk-like viruses. *J Infect Dis* **181 Suppl 2**: S284-287.

Murata, T., N. Katsushima, K. Mizuta, Y. Muraki, S. Hongo & Y. Matsuzaki, (2007) Prolonged norovirus shedding in infants <or=6 months of age with gastroenteritis. *Pediatr Infect Dis J* **26**: 46-49.

Nayfach, S. & K.S. Pollard, (2016) Toward accurate and quantitative comparative metagenomics. *Cell* **166**: 1103-1116.

Payment, P. & A. Locas, (2011) Pathogens in Water: Value and Limits of Correlation with Microbial Indicators. *Ground Water* **49**: 4-11.

Philippe, N., M. Legendre, G. Doutre, Y. Coute, O. Poirot, M. Lescot, *et al.*, (2013) Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341**: 281-286.

Popgeorgiev, N., G. Michel, H. Lepidi, D. Raoult & C. Desnues, (2013) Marseillevirus Adenitis in an 11-Month-Old Child. *Journal of Clinical Microbiology* **51**: 4102-4105.

Prevost, B., M. Goulet, F.S. Lucas, M. Joyeux, L. Moulin & S. Wurtzer, (2016) Viral persistence in surface and drinking water: Suitability of PCR pre-treatment with intercalating dyes. *Water Res* **91**: 68-76.

Qin, J., R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, *et al.*, (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**: 59-65.

Reyes, A., M. Haynes, N. Hanson, F.E. Angly, A.C. Heath, F. Rohwer, *et al.*, (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**: 334-U381.

Rinta-Kanto, J.M. & S.W. Wilhelm, (2006) Diversity of microcystin-producing cyanobacteria in spatially isolated regions of Lake Erie. *Applied and Environmental Microbiology* **72**: 5083-5085.

Rodriguez-Brito, B., L. Li, L. Wegley, M. Furlan, F. Angly, M. Breitbart, *et al.*, (2010) Viral and microbial community dynamics in four aquatic environments. *ISME J* **4**: 739-751.

Rohwer, F. & R. Edwards, (2002) The Phage Proteomic Tree: a genome-based taxonomy for phage. *Journal of Bacteriology* **184**: 4529-4535.

Roux, S., J.R. Brum, B.E. Dutilh, S. Sunagawa, M.B. Duhaime, A. Loy, *et al.*, (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**: 689-693.

Roux, S., F. Enault, A. Robin, V. Ravet, S. Personnic, S. Theil, *et al.*, (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**: e33641.

Rutjes, S.A., H.H.J.L. van den Berg, W.J. Lodder & A.M. de Roda Husman, (2006) Real-Time Detection of Noroviruses in Surface Water by Use of a Broadly Reactive Nucleic Acid Sequence-Based Amplification Assay. *Applied and Environmental Microbiology* **72**: 5349-5358.

Seurinck, S., T. Defoirdt, W. Verstraete & S.D. Siciliano, (2005) Detection and quantification of the human-specific HF183 Bacteroides 16S rRNA genetic marker with real-time PCR for assessment of human faecal pollution in freshwater. *Environ Microbiol* **7**: 249-259.

Short, C.M., O. Rusanova & S.M. Short, (2011) Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada. *Isme Journal* **5**: 810-821.

Short, S.M. & C.M. Short, (2009) Quantitative PCR reveals transient and persistent algal viruses in Lake Ontario, Canada. *Environmental Microbiology* **11**: 2639-2648.

Sinclair, R.G., E.L. Jones & C.P. Gerba, (2009) Viruses in recreational water-borne disease outbreaks: A review. In: Journal of Applied Microbiology. pp. 1769-1780.

Suttle, C.A., (2007) Marine viruses - major players in the global ecosystem. *Nature Reviews Microbiology* **5**: 801-812.

Sweet, M. & J. Bythell, (2017) The role of viruses in coral health and disease. *Journal of Invertebrate Pathology* **147**: 136-144.

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

Tessler, M., J.S. Neumann, E. Afshinnekoo, M. Pineda, R. Hersch, L.F.M. Velho, *et al.*, (2017) Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports* **7**.

Thompson, T.M., W.N. Batts, M. Faisal, P. Bowser, J.W. Casey, K. Phillips, *et al.*, (2011) Emergence of viral hemorrhagic septicemia virus in the North American Great Lakes region is associated with low viral genetic diversity. *Diseases of Aquatic Organisms* **96**: 29-43.

Tremblay, J., K. Singh, A. Fern, E.S. Kirton, S. He, T. Woyke, *et al.*, (2015) Primer and platform effects on 16S rRNA tag sequencing. *Front Microbiol* **6**: 771.

Tseng, C.H., P.W. Chiang, F.K. Shiah, Y.L. Chen, J.R. Liou, T.C. Hsu, *et al.*, (2013) Microbial and viral metagenomes of a subtropical freshwater reservoir subject to climatic disturbances. *ISME J* **7**: 2374-2386.

Ulrich, D., Whitman R. and D. Alley., (2009) Work group report on beaches and recreational water quality. In: Great Lakes water quality agreement priorities. pp. 20.

van Heerden, J., M.M. Ehlers, A. Heim & W.O. Grabow, (2005) Prevalence, quantification and typing of adenoviruses detected in river and treated drinking water in South Africa. *J Appl Microbiol* **99**: 234-242.

Waples, J.T., B. Eadie, J.V. Klump, M. Squires, A. J. Cotner & G. Mckinley, (2008) The Laurential Great Lakes. *Continental Margins: A synthesis and planning workshop*.

Wright, M.E., H.M. Solo-Gabriele, S. Elmir & L.E. Fleming, (2009) Microbial load from animal feces at a recreational beach. *Marine Pollution Bulletin* **58**: 1649-1656.

Wu, J., S.C. Long, D. Das & S.M. Dorner, (2011) Are microbial indicators and pathogens correlated? A statistical analysis of 40 years of research. *Journal of Water and Health* **9**: 265-278.

Xagoraraki, I., D.H.W. Kuo, K. Wong, M. Wong & J.B. Rose, (2007a) Occurrence of human adenoviruses at two recreational beaches of the great lakes. *Appl. Environ. Microbiol.* **73**: 7874-7881.

Xagoraraki, I., D.H.W. Kuo, K. Wong, M. Wong & J.B. Rose, (2007b) Occurrence of Human Adenoviruses at Two Recreational Beaches of the Great Lakes. *Applied and Environmental Microbiology* **73**: 7874-7881.

Yezli, S. & J.A. Otter, (2011) Minimum Infective Dose of the Major Human Respiratory and Enteric Viruses Transmitted Through Food and the Environment. *Food Environ Virol* **3**: 1-30.

Yolken, R.H., L. Jones-Brando, D.D. Dunigan, G. Kannan, F. Dickerson, E. Severance, *et al.*, (2014) Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. *Proc Natl Acad Sci U S A* **111**: 16106-16111.

Table 7.1 List of pathogenic viruses identified in the lower Great Lakes region

Potential Pathogenic Viruses	Virus Family	Primary Host	Identification Method[1]	Location	References[2]
Adenoviruses	*Adenoviridae*	Human and other vertebrates	Shotgun Metagenomic Sequencing, Immunoperoxidase assay	Lake Ontario, Lake Erie	(Edge *et al.*, 2013, Mohiuddin & Schellhorn, 2015)
Noroviruses	*Caliciviridae*	Human	RT-PCR, Immunoperoxidase assay	Lake Ontario	(Edge *et al.*, 2013, Greer *et al.*, 2009)
Iridoviruses	*Iridoviridae*	Insects, Amphibians, Fish, Invertebrates	Shotgun Metagenomic Sequencing	Lake Ontario, Lake Erie	(Mohiuddin & Schellhorn, 2015)
Cyanophages	*Myoviridae, Podoviridae, Siphoviridae*	Bacteria	Shotgun Metagenomic Sequencing, qPCR	Lake Ontario, Lake Erie	(Matteson *et al.*, 2011, Mohiuddin & Schellhorn, 2015)
Phycodnaviruses	*Phycodnaviridae*	Algae, Human (rare)	PCR, Shotgun Metagenomic Sequencing, qPCR,	Lake Ontario, Lake Erie	(Mirza *et al.*, 2015, Mohiuddin & Schellhorn, 2015, Short & Short, 2009)
Viral Hemorrhagic Septicemia viruses (VHSv)	*Rhabdoviridae*	Fish	PCR, qPCR	Lake Ontario, Lake Erie	(Cornwell *et al.*, 2012, Thompson *et al.*, 2011)
Giant viruses	*Mimiviridae, Marseilleviridae*	Amoeba, Human (still unclear)	Shotgun Metagenomic Sequencing	Lake Ontario, Lake Erie	(Mohiuddin & Schellhorn, 2015)
Poxviruses	*Poxviridae*	Human and other vertebrates, Arthropods	Shotgun Metagenomic Sequencing	Lake Ontario, Lake Erie	(Mohiuddin & Schellhorn, 2015)

[1] RT-PCR (Reverse transcriptase-polymerase chain reaction), qPCR (Quantitative polymerase chain reaction); [2] Edge *et al.*, 2013, used immunoperoxidase method to identify enteroviruses which include both adenoviruses and noroviruses. However, the type of viruses present were not specified in the study.