Investigating freshwater bacterial diversity, community composition and function in hundreds of Canadian Lakes

Vera Eyenkpobomini Onana

A Thesis
In the Department of
Biology

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Biology)
at Concordia University
Montreal, Quebec, Canada

June 2024

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Vera E. Onana**

Entitled: **Investigating freshwater bacterial diversity, community composition and function in hundreds of Canadian Lakes**

and submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy (Biology)

Complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____Chair
Dr. Malcolm Whiteway

_____External Examiner
Dr. Cassandre Lazar

_____External to Program
Dr. Selvadurai Dayanandan

_____Examiner
Dr. Irene Gregory-Eaves

_____Examiner
Dr. John Philippe Lessard

_____Thesis CoSupervisor
Dr. Beatrix Beisner

_____Thesis Supervisor
Dr. David Walsh

Approved by _____
Dr. Robert Weladji, Graduate Program Director

July 16, 2024_____
Dr. Pascale Sicotte, Dean Faculty of Arts and Science

Abstract

Investigating freshwater bacterial community composition and function in hundreds of Canadian Lakes

Vera Onana, PhD
Concordia University, 2024

Freshwater bacterial communities play important roles in global biogeochemical cycling and aquatic food webs, yet bacterial diversity, community composition, and community metabolism in freshwater ecosystems remain less explored compared to terrestrial and marine ecosystems. In this thesis work, I investigated bacterial communities in hundreds of lakes located across Canada, a country that contains millions of lakes. Utilizing 16S rRNA and metagenomic techniques, this research explores diversity patterns, community composition and functional capabilities of lake bacterial communities and links variation in these three components to human-mediated alterations, specifically watershed land use types within lake watersheds. In the first research chapter, I performed an investigation of communities in 403 lakes from seven ecozones. I identified distinct bacterial diversity patterns between western (Semi-Arid Plateaux, Prairies, and Boreal Plains ecozones) and eastern (Boreal Shield, Mixedwood Plains, Atlantic Maritimes, and Atlantic Highlands ecozones) Canada. The identified pattern was primarily influenced by lake physicochemistry including productivity, ion concentration, and lake depth. Bacterial community structure was influenced particularly by lake pH and trophic state. In the next research chapter, I expanded the study to 621 lakes across 12 ecozones and explored variation in diversity and community composition patterns in relation to water quality and land use. Total phosphorus (TP) was identified as a key variable shaping community composition, with notable shifts occurring at 110 µg/L TP. Variation in bacterial communities within the Prairies ecozone were driven by agriculture while urbanisation played a role in structuring community composition within the Pacific Maritimes ecozone. In the final research chapter, I investigated bacterial functional capabilities using gene-centric metagenomics. Physicochemical parameters emerged as top predictors of variation in functional gene composition, with xenobiotics biodegradation and metabolism notably influenced. Overall, the research presented in this thesis demonstrates that bacterial diversity, community composition, and community function exhibit variations across

continental and regional scales that can be attributed to within-lake conditions and watershed land use types.

Acknowledgements

I would like to express my sincere gratitude to all those who have supported me throughout the journey of completing this thesis. I am deeply thankful to my supervisors, Drs. David Walsh and Beatrix Beisner, for their invaluable guidance, encouragement, and knowledge impartation. Their expertise, patience, and insightful feedback have been instrumental in shaping this research and guiding me through its various stages. I express immense gratitude to the Director of the NSERC Canadian LakePulse Network, Dr. Yannick Huot, without which this thesis would be impossible. I also acknowledge research grants by ECOLAC and the GRIL.

I am also grateful to the members of my thesis committee, Drs. Irene Gregory-eaves and Jean Philippe-Lessard for their valuable input, constructive criticism, and scholarly advice, which have enriched the quality of this work. I extend my appreciation to my colleagues and peers (at the Walsh and Beisner labs respectively) for their camaraderie and intellectual exchange.

Special thanks are due to my family (the Onanas and Adebonas) and friends (Ayooluwa Bolaji, PhD and Naila Barbosa de Costa, PhD) for their love, encouragement, and understanding during this demanding period. Their belief in my abilities and their steadfast support have been the cornerstone of my academic pursuits.

In conclusion, I am deeply grateful to everyone who has contributed to this thesis in various ways. Your support, guidance, and encouragement have been indispensable, and I am sincerely thankful for your role in this academic endeavor.

Dedication

Dedicated to every girl child defying the odds to pursue quality education. Your only limitations are those you set for yourself. If you can envision it, you can achieve it.

To *Eni Ebitobouh*, for revealing a new dimension of resilience to me. I didn't know my own strength, until you helped me discover it.

Finally, *To Eni Ebizhei*, for your unwavering love, support, and being a consistent pillar of comfort. You are my rock. I couldn't finish this marathon without you.

Contribution of Authors

Chapter 2: Susanne Kraemer and Rebecca Garner processed the sequence data. Rebecca Garner performed the Generalized Dissimilarity Modelling analyses. Amplicon sequencing was performed by Susanne Kraemer and Naila Barbosa de Costa for the 2017 samples. Vera Onana carried out DNA extractions and PCR amplifications on 2018 samples. 2018 samples were sequenced by Vera Onana in collaboration with Naila Barbosa de Costa and Genevieve Bourret. Amplicon data and other statistical analysis was performed by Vera Onana. Vera Onana and David A. Walsh, with contributions from Beatrix Beisner, drafted the manuscript. This chapter has not been submitted for publication.

Chapter 3: Rebecca Garner processed the sequence data. Amplicon sequencing was performed by Susanne Kraemer and Naila Barbosa de Costa for the 2017 subset of the dataset. 2018 subset was sequenced by Vera Onana in collaboration with Naila Barbosa de Costa and Genevieve Bourret. 2019 subset was prepared by Vera Onana and sequenced at Genome Quebec (GQ). Amplicon data and all other statistical analysis was performed by Vera Onana. Vera Onana and David A. Walsh, with contributions from Beatrix Beisner, drafted the manuscript. This chapter has been submitted for publication. All authors reviewed the final manuscript and approved of the contents.

Chapter 4: Vera Onana extracted the environmental genomic DNA, carried out PCR analysis and prepared data for shotgun sequencing at GQ. Susanne Kraemer processed sequence data. Thomas Grevesse contributed to the bioinformatic analyses and performed the non-negative matrix factorization analysis. Vera Onana performed statistical analysis and drafted the manuscript together with David A. Walsh and Beatrix Beisner. This chapter has not been submitted for publication.

Table of Contents

List of Figures

List of Supplementary Figures

List of Tables

List of Supplementary Tables

List of Abbreviations

ANOVA	Analysis of Variance
ARGs	Antibiotic resistant genes
ASVs	Amplicon sequence variants
BDOD	Bulk density of the fine earth fraction
CEC	Cation exchange capacity
CFVO	Volumetric fraction of coarse fragments
db-RDA	Distance-based Redundancy Analysis
DIC	Dissolved inorganic carbon
DNA	Deoxyribonucleic Acid
DOC	Dissolved organic carbon
DON	Dissolved organic nitrogen
GDMs	Generalized dissimilarity models
ISA	Indicator species analysis
JGI/IMG	Joint genome Institute/Integrated Microbial Genomes
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG orthology
LCBD	Local contribution to β-diversity
LCLU	Land use land cover
MAG	Metagenome-assembled genomes
MEM	Moran's eigenvector maps
MRGs	Metals resistant genes
N	Nitrogen
NGS	Next-generation sequencing
NMDS	Non-metric multi-dimensional scaling
NMF	Non-negative matrix factorization
NSERC	Natural Science and Engineering Research Council
OC	Organic carbon
OCD	Organic Carbon Density
OOB	Out of bag error
P	Phosphorus
PAHs	Polycyclic aromatic hydrocarbons

PCA	Principal Component Analysis
PCoA	Principal Coordinate Analysis
PCPs	Personal care products
PCR	Polymerase chain reaction
PDP	Partial dependency plots
pH	Potential of hydrogen
PhACs	Pharmaceutical active compounds
pHH_20	Soil pH
PON	Particulate organic nitrogen
RDA	Redundancy Analysis
RF	Random Forest
rRNA	ribosomal ribonucleic acid
SMGs	Sub-metagenomes
SOC	Soil organic carbon content in the fine earth fraction
TITAN	Threshold indicator taxa analysis
TN	Total nitrogen
TP	Total phosphorus

Chapter 1: Introduction

1.1 Freshwater microbial diversity

Microorganisms, constituting an incredible 60% of Earth's total biomass, represent a massive reservoir of living diversity on Earth (Whitman et al., 1998; Wooley et al., 2010). Despite their ubiquity, the microbial world remains vastly unexplored, with an estimated 5×10^{30} prokaryotic cells globally, out of which a mere 0.01% have been identified (Jurasinski & Koch, 2011). Microbial diversity encompasses the variety of microorganisms at the genetic, species, and ecosystem levels. It encapsulates the ecological complexity in which microbial organisms occur and the ecological processes they contribute to. Thus, microbial diversity comprises aspects of richness, evenness, composition, and function (Achtman & Wagner, 2008). Lakes, specifically, emerge as hotspots teeming with microbial life, housing an extraordinary vastness of bacteria (Newton et al., 2011; Strayer & Dudgeon, 2010) In freshwater ecosystems, microbial communities exhibit dense assemblages, showcasing a diverse range of prokaryotic organisms with varied morphology, physiology, and ecological preferences (Cotner & Biddanda, 2002).

Bacteria in particular, are abundant in freshwater habitats, establishing extensive populations in both pelagic and benthic regions of lakes and rivers. These microorganisms actively participate in crucial biogeochemical cycles, influencing various aspects of ecosystem dynamics. Their roles encompass nutrient cycling dynamics (Arora-Williams et al., 2018; Butman et al., 2016), decomposition of organic matter, and nutrient release, which are pivotal for sustaining ecosystem health (Gayer et al., 2021; Stadler et al., 2020; S. Wang et al., 2019). Additionally, they play a central role in carbon cycling and greenhouse gas dynamics (Bastviken et al., 2011; DelSontro et al., 2018; Li et al., 2024; Reis et al., 2022). Furthermore, these microorganisms contribute significantly to supporting primary production (Straškrábová et al., 2005) and influence food web dynamics (Berman, 1990; Ives et al., 2019; Newton et al., 2011)

Freshwater bacteria play a crucial role as primary sources of food and nutrients for other organisms within aquatic ecosystem from protists to animals, thereby forming the base of the aquatic food chain (Burns & Galbraith, 2007). Moreover, aquatic bacteria contribute to maintaining water quality through processes like denitrification and contaminant degradation (Castellano-Hinojosa et al., 2017; Wu et al., 2019), thereby enhancing the purification of freshwater resources. The intricate biological interactions of these microorganisms also contribute to ecosystem resilience (Peter et al., 2011; Shade et al., 2012), highlighting their indispensable role in the intricate balance of both biotic

and abiotic systems within lakes. The diversity of microbial life in aquatic ecosystems is closely linked to various physical and chemical factors, each exerting its influence in complex and often opposing ways. These factors significantly shape the composition and metabolic functions of bacterial communities. The field of bacterial ecology delves into this intricate interplay, encompassing studies at various levels – from individual organisms to entire ecosystems, and employing diverse tools in molecular biology such as amplicon gene sequencing and shotgun metagenomics.

Understanding how these abiotic factors sculpt bacterial communities in aquatic environments holds profound implications for assessing ecosystem health. Microbes demonstrate remarkable adaptability, swiftly responding to environmental fluctuations with discernible physiological and metabolic adjustments (Nguyen et al., 2021; Sadeghi et al., 2021). Yet, despite significant progress in microbiological research, the precise balance between stochastic and deterministic processes in shaping microbial communities, and their connections to local and broader environmental parameters, like physicochemical conditions and land use, remain elusive. While studies have hinted at the influence of ecological factors such as water chemistry (Lindström et al., 2005; Methé & Zehr, 1999; Zwart et al., 2002) water temperature (Pearce, 2008), organic matter availability (Crump et al., 2003), pH, and water retention time (Lindström et al., 2005, 2006) on aquatic bacterial community composition in lakes, the impact of environmental changes and anthropogenic activities (e.g. watershed human land use) on bacterial diversity, community composition and function remain poorly understood.

1.2 Global significance of lakes

Freshwater lakes have significant impacts on global carbon and nitrogen cycling. Lakes are active sites for the transport, transformation, and storage of considerable amounts of carbon received from their surrounding terrestrial environment (Toming et al., 2020). In addition, they act as collectors of terrestrial carbon, with dissolved inorganic carbon (DIC) and dissolved organic carbon (DOC) being the predominant carbon inputs, whose variability is generally influenced by lake location and hydrology (Toming et al., 2020). For instance, half of the carbon received by freshwater ecosystems from the terrestrial landscape is emitted as carbon dioxide (0.2 Pg C/year) or buried in sediments (0.8 Pg C/year) (Lindström et al., 2005; Methé & Zehr, 1999; Zwart et al., 2002). Collectively, nearly half as much organic carbon (OC) as in the world's oceans is buried in lakes globally (42 vs. ~100 Tg C yr^{-1}). Though small lakes (<500 km^2) may account for 60–70% of this total OC burial, large

freshwater lakes still sequester an estimated 6–13% as much OC annually as the world oceans (Alin & Johnson, 2007).

Similarly, nitrogen cycling in aquatic ecosystems follows a complex series of transformations that involves a variety of nitrogen (N) forms and oxidation states. It is estimated that 20% of global denitrification occurs in freshwater, roughly equivalent to the amount of denitrification taking place in soils (22%) (Seitzinger et al., 2006). Both the oxidized and reduced inorganic N species (NO_2^-, NO_3^-, and NH_3) and organic N fractions (DON and PON) are commonly found in freshwater and are introduced into water bodies either directly, via point sources (wastewater influx, sewage effluent), or indirectly by non-point sources (agricultural runoffs, road salt). Such nutrient loads are often delivered to lakes from catchment areas depending on the hydrological processes, particularly from intensively farmed agricultural watersheds (Robertson et al., 2019).

From time immemorial, humanity has depended on freshwater ecosystems for several survival necessities including food, energy, water supply, transportation, as well as recreational and cultural needs (Sterner et al., 2020). Lakes have provided small communities as well as massive cities with vital ecosystem services such as hydroelectric power generation, and often serve as a source of water for domestic purposes such as drinking, cooking, or cleaning. Some of the most intriguing investigations on the ecosystem services provided by lakes have been carried out within the Laurentian Great Lakes (Steinman et al., 2017), some of which are located in Canada. In North America, large lakes have been found to be significant in the development of some of the biggest and most advanced industrial regional economies on the planet (Sterner et al., 2020).

In addition to the key roles played in biogeochemical cycling, ample amount of research has demonstrated the sensitivity of lakes to climate (Adrian et al., 2009; Butcher et al., 2015; Crossman et al., 2016; Woolway et al., 2022) and emphasized the fact that lake properties (physical, chemical and biological) respond in a rapid manner to climate-mediated changes (Adrian et al., 2009). This suggests that lakes are “sentinels” of climate change because they are sensitive to environmental changes and can depict changes in the surrounding landscape due to alterations of lake properties (Carpenter et al., 2007; Pham et al., 2008; Williamson et al., 2009).

1.2.1 The Canadian lake landscape and ecozones

Canada is the country with the most lakes in the world, containing more than 2 million lakes of all sizes (Minns et al., 2008). Canada hosts about 20% of the world's freshwater stock and 90% of its municipal drinking water comes from lakes (Huot et al., 2019). These lakes are scattered across the longitudinal and latitudinal spread of the country, nestled within unique ecological and environmental conditions. Canada's territory is divided into 18 terrestrial ecozones, based on landforms, soils, water features, vegetation and climatic conditions (Figure 1.1 CCEA, 2016; Wiken, 1986). These ecozones depict unique geologic, climate, terrestrial and environmental conditions across provinces. Spread across Alberta, Saskatchewan and Manitoba are the Prairies and Boreal Plains ecozones. Both ecozones are characterised by intensive crop and agricultural farmland. Lakes within the Prairies are generally shallow and nutrient-rich, with low N:P ratios and toxic cyanobacteria blooms (Nanayakkara, 2018; Quinlan et al., 2002; Taranu et al., 2010). These lakes possess a high concentration of ions and generally range from eutrophic to hypereutrophic. For instance, Lake Winnipeg receives high nutrient loads and algal blooms occur annually in the lake (Schindler et al., 2012).

In the Boreal Shield ecozone, forestry is prevalent and acidification within lakes in this ecozone could be related to TP and calcium declines in lakes (Jeziorski et al., 2008; Pinder et al., 2014; Yan et al., 2008). Generally, regions in Eastern Canada are highly affected by urbanization, including parts of the Boreal Shield, Mixedwood Plains and Atlantic Maritime ecozones. For instance, the Mixedwood Plains ecozones, spanning the provinces of Ontario and Quebec is recognised as the industrial and urbanisation heartland of the country. Lakes within this ecozone are known to be deeper, range from oligotrophic to intermediate trophic levels and are less nutrient-rich when compared to the Prairies (Figure 1.2). A description of predominant land cover and land use (LCLU) in Canada's various ecozones is highlighted in **Table 1**.

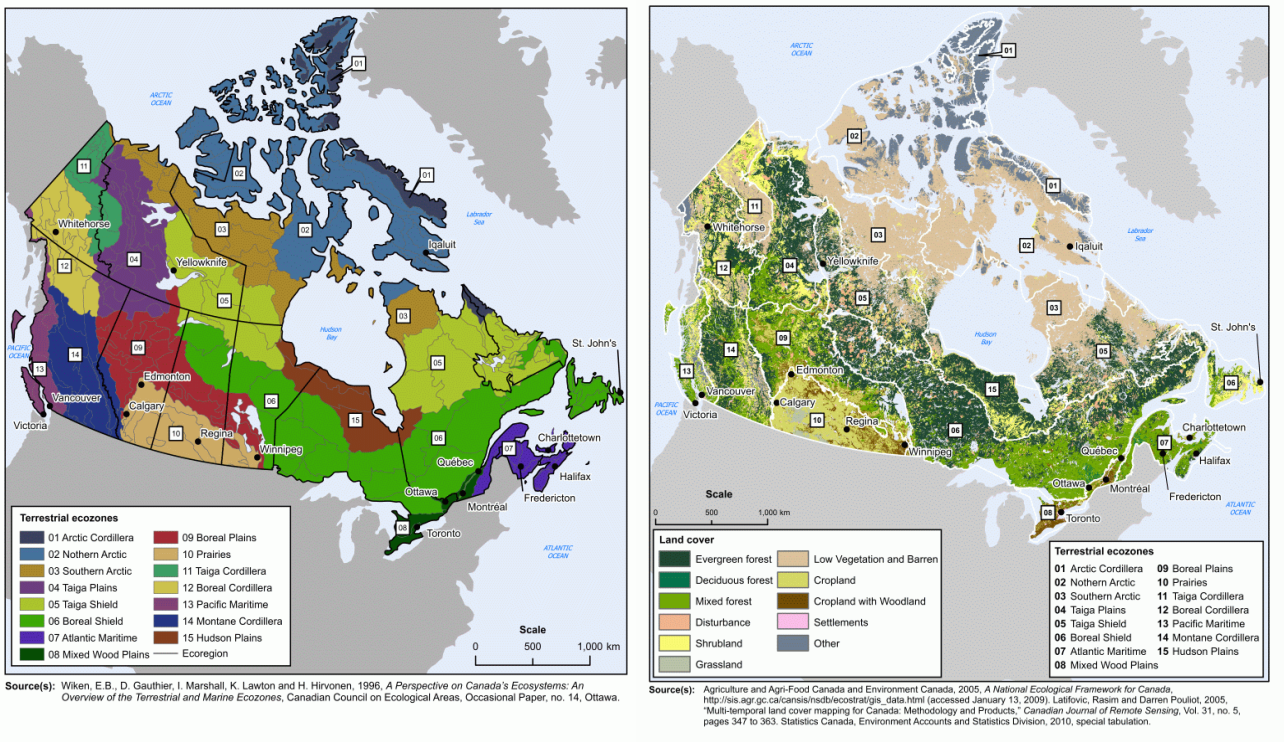


Figure. 1.1 *Map of Canada showing terrestrial ecozones and major cities (left) and land use land cover predominant within ecozones (right) (Figures from Wiken et al., 1996; Agriculture and Agri-Food Canada and Environment Canada, 2005)*

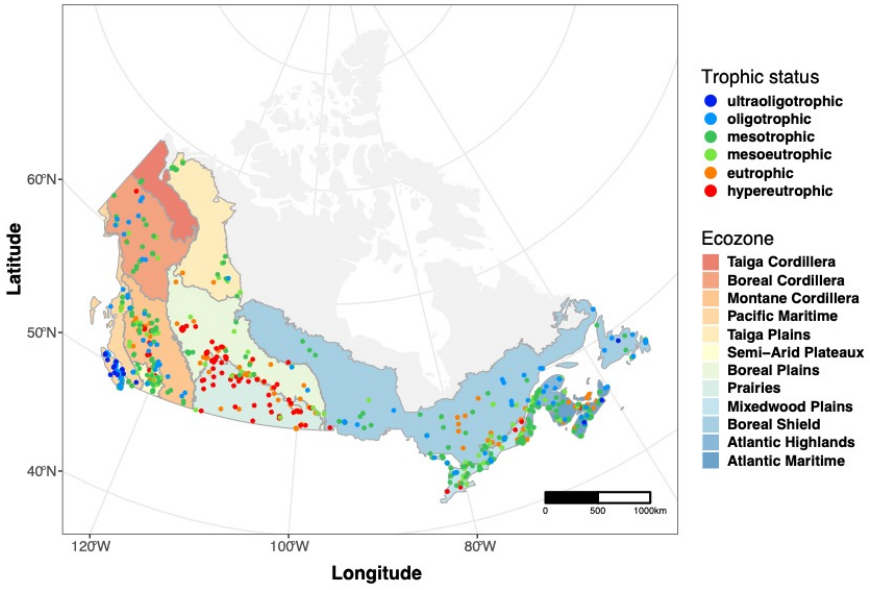


Figure 1.2 *Map of the 621 freshwater lakes sampled across Canada over three summers (2017-2019) as part of the NSERC Canadian LakePulse field campaign. Coloured dots depict the lake trophic state from ultraoligotrophic to hypereutrophic. Map creation: Atlas Lambert projection of Canada (NAD83 CSIS).*

Table 1 Human Activities within Canada's Terrestrial Ecozones based on Wiken (1986), Brown et Lomolino (1998) and (Gibbs et al., 2009).

	Agriculture	Urbanisation	Forestry	Pasture
Taiga Cordillera			x	
Boreal Cordillera			x	
Montane Cordillera			x	x
Pacific Maritime		x		
Taiga Plains			x	
Semi-Arid Plateaux	x	x		
Boreal Plains	x		x	x
Prairies	x			x
Mixedwood Plains		x		x
Boreal Shield			x	
Atlantic Highlands	x	x	x	
Atlantic Maritimes	x		x	

1.3. Major anthropogenic pressures on lake ecosystems

Changes in lake ecosystems could be natural or mediated by anthropogenic activities. Almost two decades ago, Dudgeon et al., (2006) identified overexploitation, water pollution, flow modification, destruction or degradation of habitat and invasion by exotic species as five leading causes of population declines and range reductions of freshwater organisms worldwide. However, the current geological era has seen more destruction bringing about multiple new, intensified, and varied threats that impact freshwater systems. Reid et al. (2019) identified a dozen emerging threats to freshwater biodiversity that could be categorized as naturally occurring or induced by human activities. Also, over the last two decades, a considerable growth in interest in potential multiple

stressor problems has been observed (Craig et al., 2017; Ormerod et al., 2010; Vörösmarty et al., 2010). First is the increasing allocation of freshwater resources for human use in addition with escalating impacts from human activities (Strayer & Dudgeon, 2010). Second is that human effects on fresh waters often occur in combination, either because different activities coincide (e.g. urbanisation with industry, agriculture with water extraction, or biomass exploitation with invasive species release) or because they affect freshwater ecosystems through multiple pathways. Lastly, climate change is expected to have widespread direct and indirect effects on fresh waters.

In this thesis, we investigated the impact of such anthropogenic pressures on lake ecosystems, focusing on bacterial diversity, community composition, and function. Our focus has primarily revolved around understanding the repercussions of watershed land use types, with particular emphasis on agriculture and urbanization. These investigations were conducted in regions where these land use types are prevalent, spanning the longitudinal and latitudinal scale of the Canadian lake landscape. This is because we have identified an expansive ecoregional versus continental scale gap in studying watershed human impact influence on bacterial diversity, community composition, and function across the nation. These gaps may be attributable to challenges such as the fact that most aquatic research focuses on macro-organisms such as zooplankton and fish, the financial cost of conducting a continental-scale study in a country with a massive landmass and millions of lakes, the prevalence of localised freshwater studies that primarily focus on one ecoregion and the paucity of water column bacterial studies. Consequently, bridging these existing knowledge gaps requires carrying out a novel ecoregional versus continental scale analysis of human influence on bacterial diversity, community composition, and function across Canadian lakes.

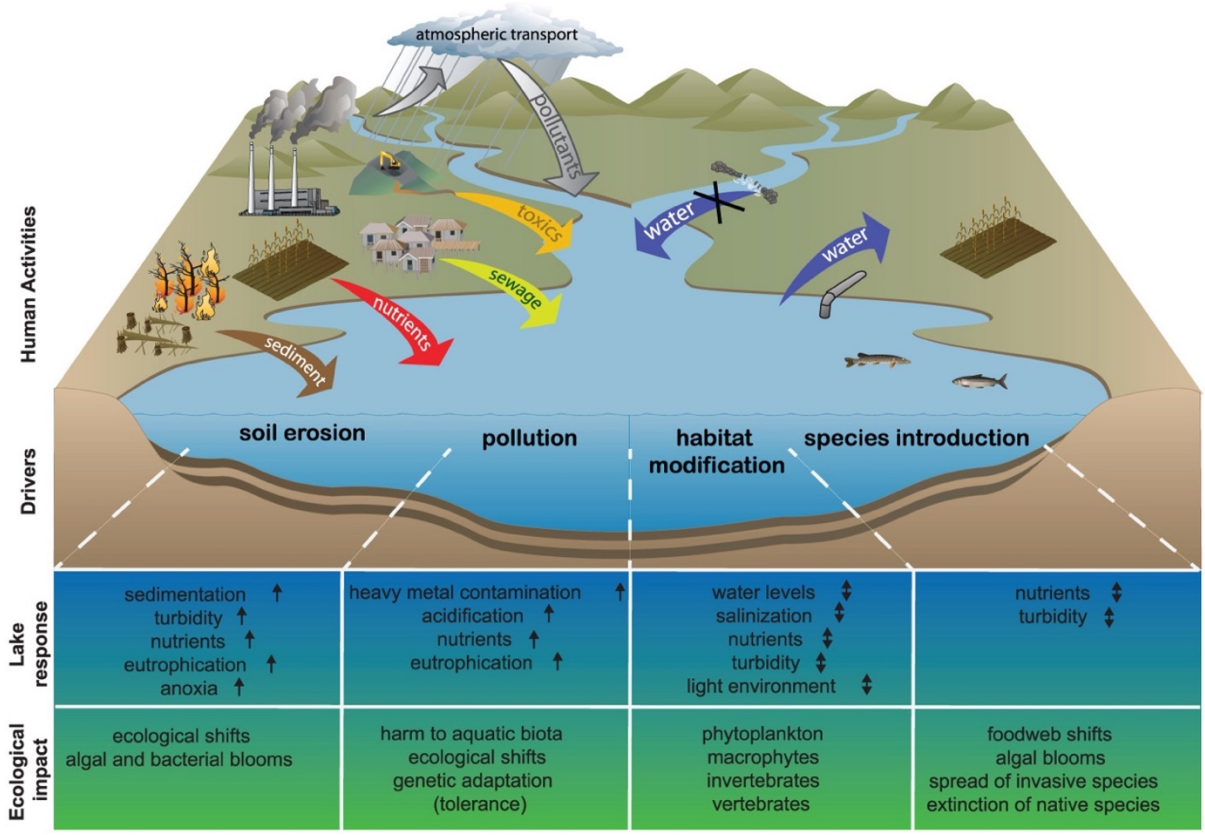


Figure 1.3 *Overview of various human-induced alterations of aquatic ecosystems, lake responses to such stressors, and the ecological impact of responses. Dubois et al, (2018)*

1.4 Associating changing environmental conditions in lakes to bacterial diversity, composition, and function

Changes in environmental conditions influence lakes in diverse ways (Figure 1.3). Since lake ecosystems are made up of physical, chemical, and biological properties, the prevailing lake conditions are an interplay of these properties. Lakes may be shallow or deep, of different ages, nutrient-rich or nutrient poor, depending on the unique environmental forcings in play. Changes in physical conditions within lakes could affect temperature, pH, concentrations of nutrients, ions, chlorophyll-*a*, turbidity, specific conductance, dissolved oxygen, dissolved organic and inorganic carbon. However, lakes do not exist in isolation, but are interconnected with the surrounding terrestrial ecosystem. Therefore, watershed soil variables like soil pH, location of lakes, and climatic influences also influence prevalent lake conditions. Since bacterial communities in lakes respond to within lake environmental changes, they are often used as a biological indicator of ecological changes in the lake environment. Previous studies have showed that environmental

variables, such as organic matter and nutrient concentration (Comte & del Giorgio, 2010; Lindström et al., 2010; Ylla et al., 2013) are significant factors influencing the bacterial diversity, community structure and function in lakes.

1.4.1 Physicochemistry

Lakes, even within the same geographical region, exhibit considerable variation in their physicochemical properties, both spatially and temporally, influenced by factors such as their origin, age, and trophic state (Clement et al., 2015). These distinctive attributes profoundly influence microbial communities and the biogeochemical processes they govern, as microbial compositions are intricately intertwined with local environmental conditions. There is a growing body of evidence suggesting a connection between changing environmental conditions in lakes and bacterial communities within lakes (Schindler et al., 2012). For example, in the freshwater lakes surrounding Beaver Island, Michigan, situated within the Laurentian Great Lakes region, a sampling campaign was conducted to explore the correlation between microbial communities and local physicochemical parameters in surface water and bottom water habitats (specifically, the epilimnion and hypolimnion during stratification). These selected lakes were characterized by distinct and divergent physicochemical attributes (Clement et al., 2015), making them ideal candidates for the investigation. The investigation revealed significant relationships between environmental factors and microbial community composition, highlighting correlations between community structure and various parameters such as dissolved oxygen, dissolved organic carbon, and temperature (Hengy et al., 2017). Despite the proximity of lakes to Beaver Island, the exhibition of diverse physicochemical characteristics suggests that environmental factors may exert stronger constraints on microbial communities than geographic distance. These findings by Hengy et al. (2017) thus aligns with previous existing theories that emphasize the predominant influence of environmental conditions on microbial community structure within lakes, with geographic proximity playing a secondary role (Van der Gucht et al., 2007; Yannarell & Triplett, 2005).

Similarly, physicochemical factors such as temperature, light availability, and total oxidized nitrogen were found to correlate with variations in microbial community composition within the Laurentian Great Lakes (Paver et al., 2020). This trend was also observed in Lake Erie, where previous studies had revealed the impact of phosphorus loading on biotic carbon flow within the ecosystem (DeBruyn et al., 2004; Jankowiak et al., 2019). Also, microbial metabolic activities related to carbon, nitrogen, and sulfur cycles in 51 Qinghai-Tibet Plateau lakes were linked to

variations in microbial community composition. These variations were found to be correlated with environmental factors, particularly elevation and salinity (Zhao et al., 2023). Similar observations were recently made in Canadian lakes, where lake trophic state emerged as a prominent driver of both taxonomic and functional diversity among metagenome-assembled genome (MAG) assemblages (Garner et al., 2023).

1.4.2 Lake Morphology

Lakes come in various shapes and sizes, with differences in area and depth. This diversity in lake morphology is often indicative of factors such as the lake's age, origin, chemical composition, and the organisms inhabiting it. The shape and volume of a lake play crucial roles in its physical and biological dynamics (Wetzel., 2001; Copetti et al., 2020). For instance, the lake's shape is influenced by the surrounding landscape and significantly impacts the underwater environment. Lakes with numerous small inlets and bays, for instance, tend to warm up rapidly and are less affected by wind compared to larger, more open bodies of water (Karamigolbaghi et al., 2019; Pilla et al., 2020). In lakes, a phenomenon known as stratification, or lake zonation, is the formation of distinct depth layers that play a vital role in shaping the entire aquatic ecosystem, including microbial communities. Studies have underscored a notable correlation between the depth profile and microbial community structure in lakes, revealing distinct bacterial communities within various layers of the aquatic environment. These layers include the epilimnion, characterized as a well-mixed layer functioning as a bioreactor for primary production; the metalimnion, situated between the epilimnion and hypolimnion layers; and the hypolimnion, serving as a sink for biomass accumulation and remineralization (Clement et al., 2015). Therefore, deep lakes and shallow lakes support different microbial communities as deep lakes have well stratified water sections (De Wever et al., 2005).

Such physical conditions within a lake often govern the depth-dependent distribution of bacteria in the lakes. While pioneering work on the depth-dependent community structure of bacterioplankton was performed in oceans (De Wever et al., 2005; Fuhrman et al., 1992; Giovannoni et al., 1990), some studies have also been carried out on the depth distribution of microorganisms in freshwater lakes (Keshri et al., 2018; Salcher et al., 2010). In a deep meromictic lake in high Arctic Canada, bacterial and archaeal community composition was determined by Comeau et al., (2012)

using high-throughput 16S rRNA gene amplicon techniques. The researchers reported that both prokaryote communities were stratified by depth. In addition, when taxa were matched to known taxon-specific biogeochemical functions, a close correspondence was found between the depth of functional specialists and chemical gradients. They concluded that a pronounced vertical structure in taxonomic and potential functional composition existed within the lake. Similarly, in the Laurentian lakes, Paver et al., 2020 demonstrated bacterial communities were distinct across depth profiles across seasons. Bacterial diversity patterns in hypereutrophic shallow lakes versus mesotrophic deep lakes of Turkey were also reported to be distinct by Ozbayram et al., (2021) wherein higher bacterial diversity and abundance was observed in shallow lake Manyas.

1.4.3 Geographic conditions

Geography can significantly influence the structure of ecological parameters and a geographical perspective is often valuable in understanding lake ecosystems. For instance, a study on epilimnetic bacterial community composition in Wisconsin lakes by Yannarell & Triplett, (2005) suggested that differences attributed to lake productivity may also be related to regional differences between northern (oligotrophic) and southern (eutrophic) Wisconsin lakes. This highlights the potential confounding effects of geographic details in study designs. Limnologists argue that a lake's position in the landscape can impact various aspects of its ecology. Additionally, research indicates that a wide range of biological and environmental variables tend to correlate with a lake's landscape position (Quinlan et al., 2003; Riera et al., 2000). Typically, there is an increase in bacterial community dissimilarity (or a decrease in similarity) with increasing geographic distance, a phenomenon known as the distance-decay relationship (Nekola & White, 1999). This relationship may be attributed to variations in environmental factors, including temperature, which tend to be correlated with geographic distance.

In 2020, Kraemer et al. employed advanced spatial modeling techniques, specifically Moran's eigenvector maps (MEMs), to analyze the geographic patterns of over 200 eastern Canadian lakes spanning four ecozones – Mixedwood Plains, Atlantic Highlands, Atlantic Maritime, and Boreal Shield. Unlike traditional distance-decay relationships, MEMs allow for the modeling of complex spatial effects, capturing both similarities among nearby lakes and spatial decay over varying scales. Through this investigation, Kraemer et al. (2020) were able to disentangle the effects of geography

from environmental factors in eastern Canadian lakes and demonstrated that while geographic variation, did not significantly impact bacterial diversity metrics such as the Shannon–Weaver index (which takes into account the diverse number of species and their relative abundance), it was associated with Chao1 richness (total number of species) in lakes across the region. In the same vein, Obieze et al. (2022) demonstrated that species distribution within Osisko Lake, in the center of Rouyn-Noranda in Quebec, Lake Winnipeg in Manitoba, and McClelland Lake in Alberta were marginally influenced by geographic distances across the Canadian land mass.

1.5 Influence of watershed land use types on bacteria

1.5.1 Influence of agriculture within lake watershed on bacteria

Chemicals originating from agricultural landscapes, including fertilizers and pesticides, represent prevalent sources of disturbance for freshwater ecosystems (Vörösmarty et al., 2010), often resulting in eutrophication (Keatley et al., 2011; Taranu & Gregory-Eaves, 2008) and biodiversity decline (Stehle & Schulz, 2015). Agricultural runoffs introduce limiting nutrients, herbicides, and insecticides to water bodies, which potentially interact to influence aquatic microbial taxa because some taxa may be less tolerant to agricultural contamination than others (Allen et al., 2021; Bani et al., 2022; Stehle & Schulz, 2015). Studies in aquatic ecosystems have shown that land use and other human activities in varying thresholds can influence the microbial content of these systems (Chen et al., 2018; Kraemer et al., 2020). Lakes are recipients of external materials from their encompassing watershed areas, which may be of point or non-point sources, thus shaping bacterial community composition (Niño-García et al., 2016). While contaminants introduced from within watersheds can influence patterns of bacterial diversity, they may also alter interaction within bacterial community and their metabolic capacities (Kiersztyn et al., 2019). This is because bacterial communities contribute substantially to ecological functioning of lakes (such as nutrient cycling). In a study conducted in German lakes, Marmen et al. (2020) found that land use within the drainage basin of 21 interconnected lakes could partially predict nitrite and nitrate concentrations in the water. These nutrient concentrations, along with temperature, chlorophyll-*a*, and total phosphorus, showed some correlation (both positive and negative) with bacterial community structure.

1.5.2 Influence of urbanization within lake watershed on bacteria

Global urbanization has rapidly increased throughout the last six decades, and by 2050, it is estimated that two-thirds of the world population will live in urban settlements (Ritchie et al., 2018). Urbanization severely impacts Earth's ecology in diverse ways. Effects of urbanisation range from alteration of natural habitats (Marzluff, 2001) and of species composition, to disruption of hydrological systems (Arnold Jr. & Gibbons, 1996; Booth & Jackson, 1997), as well as distortion of energy flow and modification of lake biogeochemistry (Grimm et al., 2000). Surface water quality is most commonly impacted by urbanization activities. These effects include biodiversity reduction and alteration due to the significant loading of pollutants from point and non-point sources as well as impervious surfaces (Glińska-Lewczuk et al., 2016).

In aquatic ecosystems, urbanization alters ecosystem functioning through the movement, magnitude, and content of surface water runoff (Alberti et al., 2007; Allan, 2004; Hale et al., 2015). Changes in microbial diversity and composition in aquatic ecosystems have been linked to urbanization (Belt et al., 2007; S.-Y. Wang et al., 2011). In freshwater tidal wetlands near Washington DC and Buenos Aires, for instance, alteration in prokaryotic community composition along urban gradient was recorded (Gonzalez Mateu et al., 2019). Scientific investigations have uncovered impacts of urbanization on microbial community diversity in other aquatic ecosystems such as streams and rivers (Hosen et al., 2017; Medeiros et al., 2016; L. Wang et al., 2018; Yu et al., 2020).

In lakes, the relationship between urban development and biodiversity patterns has only been minimally explored. Despite the recognized effects of nutrient loading on aquatic systems, the influence of urbanization on the bacterial community composition of these systems is not fully understood (Newton & McLellan, 2015) but some work has been done in Canada. Kraemer et al. (2020) in a regional study using a subset of dataset used in this thesis showed that urbanization explained variation in bacterial community composition within the Boreal Shield and Atlantic Maritime ecozones, two of the four ecozones that make up the Eastern Canada lakes. Similarly, Garner et al., (2023) revealed the influence of urbanisation on MAG assemblages at the continental scale, encompassing 12 Canadian ecozones.

1.6 Assessing microbial diversity in the environment

The identification and characterization of bacterial populations in freshwater lakes provide valuable insights into the ecological niches inhabited by bacteria across diverse freshwater ecosystems. There are multi-faceted methods for assessing and characterising microbial community diversity, composition, and metabolic capacities in aquatic ecosystems. The traditional method of studying the physiology of microbial cells was to isolate and cultivate pure cultures from a community to identify and characterise a specific organism (Bussmann et al., 2001). Traditional methods for bacterial identification typically involve phenotypic characterization of the target organism through techniques such as Gram staining, culture-based methods, and biochemical assays, as well as the utilization of various carbon, nitrogen, and phosphorus sources (Shafi et al., 2017). Microbiologists commonly employ a range of selective, non-selective, and differential media for the enrichment, isolation and identification of bacterial strains (Dyall-Smith & Oren, 2006) but the assessment of bacterial diversity using cultivation-dependent methods has been reported to generate erroneous information owing to the existence of many unculturable bacterial species (Pearce et al., 2003). While certain clades of aquatic microbes may now have cultured representatives (A. C. Martiny, 2019), several more remain uncultivated making this method capture far less diversity and richness than exists in the environment (Vaz-Moreira et al., 2011). Hence, methods of analysing aquatic microbial diversity have shifted from cultivation-dependent approaches to molecular gene-based cultivation-independent approaches.

1.6.1 16S rRNA gene amplicon technique

Challenges associated with cultivation-dependent strategies were solved using nucleic acid-based sequencing of universal phylogenetic markers like the small-subunit ribosomal RNA (rRNA) genes from microbial communities. This method was based on work pioneered by Carl Woese and others who constructed the universal tree of life in 1987 (Woese 1987; Woese & Fox, 1977). The 16S rRNA gene plays a pivotal role in the study of bacterial evolution and ecology due to its ubiquity across all cellular life, high sequence conservation, and a domain structure with variable regions. These properties have led to two significant revolutions and have transformed our understanding of evolution, shifting from a five Kingdom to a three Domain paradigm. It provides an objective phylogenetic framework for classifying cellular life, enabling a more accurate depiction of evolutionary relationships (Woese 1987). Through the cloning and sequencing of 16S rRNA genes directly from the environment using conserved broad-specificity

PCR primers, a vast extent of microbial diversity that far surpasses what was previously known from culture-based studies has been revealed, highlighting the importance of non-culturable microorganisms in ecosystems (Pace, 1997).

Ever since, rRNA gene sequence analysis has empowered microbial ecologists to discern phylogenetic identity and relative abundance of microbial community, advancing a remarkable comprehension of the intricacies within the aquatic microbial realm. For instance, in 18 freshwater lakes located in North America, Newton et al., (2007) investigated the prevalence and abundance of members of the *acI* lineage of Actinobacteria and demonstrated that both phylogeographic patterns in the landscape and environmental filtering by lake pH contributed to the *acI* community structure. Such findings were made possible by 16S rRNA gene technique.

Intriguingly, most 16S rRNA surveys have been performed in marine and soil habitats. A phenomenal example in the Sargasso Sea is the phylogenetic analysis of sequences from marine environments that revealed habitat-specific phylogenetic clusters. The most prominent are the SAR clusters, monophyletic lineages of solely marine 16S rRNA sequences (Giovannoni et al., 1990; Mullins et al., 1995) but some freshwater-specific clusters of the SAR11 have been recovered in lakes from North America and Europe using 16S rRNA methods (Zwart et al., 1998). In comparative studies investigating cosmopolitan phylogenetic clusters of freshwater bacteria, Glöckner et al. (2000) recovered a total of 190 full and partial 16S rRNA sequences from three different lakes – Lake Gossenköllesee, Austria; Lake Fuchskuhle, Germany; and Lake Baikal, Russia. The authors' combination of phylogenetic analysis and Fluorescent in Situ Hybridization was used to reveal 16 globally distributed sequence clusters, confirming a broad distribution, abundance, and high biomass of members of the class *Actinobacteria* in freshwater ecosystems. More recently, 16S rRNA gene analysis methods have been used, for example, to assess seasonal dynamics of lotic bacterial communities in a Norwegian rural creek named Grytelandsbekken (Paruch et al., 2020), to elucidate spatio-temporal dynamics of bacterial communities in the great lakes (Shahraki et al., 2021) and to reveal the influence of land use on lake bacterial communities in eastern Canadian lakes (Kraemer et al., 2020).

While the 16S rRNA gene technique boasts numerous strengths, recent scientific advancements have brought to light some limitations. These include potential inaccuracies in representing microbial communities due to biases introduced during molecular community analysis. Various methodological factors, such as sample handling, DNA extraction, and PCR can introduce biases (Case et al., 2007; Egert & Friedrich, 2003; Polz & Cavanaugh, 1998). Additionally, the existence of multiple heterogeneous copies of the 16S rRNA gene within a genome can further contribute to inaccuracies (Crosby & Criddle, 2003). Moreover, 16S rRNA-based techniques have other known challenges, including short read lengths, sequencing errors, variations arising from the selection of different gene regions, and challenges in assessing operational taxonomic units (OTUs) (Poretsky et al., 2014). Therefore, using a single marker gene to assess diversity is challenging due to the prevalence of horizontal gene transfer and the difficulty in defining bacterial species (Konstantinidis et al., 2006), as well as the limited resolution of the 16S rRNA gene among closely related species. Indeed, while small subunit rRNA genes serve as valuable phylogenetic markers, they do not provide insights into the metabolic capabilities of microbial communities, rendering them unsuitable for functional studies.

1.6.2 Metagenomics

The term metagenomics was first defined by Handelsman in 1998 as estimating the total genetic material of any microbial communities, providing microbial and genetic diversity, and metabolic processes in a confined environment (Handelsman et al., 1998). In contrast to amplicon sequencing and its limitations, whole-genome shotgun metagenomics provides a comprehensive view of microbial communities, surpassing taxonomic composition and bypassing primer biases introduced during PCR amplification. This approach allows for the sequencing of genomes within an environmental sample, enabling the exploration of *who is present in a community* (taxonomic structure), *what they are doing* (functional structure), and how these microorganisms interact to maintain ecological balance. This capability marks a significant advancement over amplicon-based methods (Oulas et al., 2015; Quince et al., 2017).

Recently, low-cost next-generation sequencing (NGS) technologies and advanced bioinformatics techniques for developing metagenomic libraries have become important tools in metagenomics (Slatko et al., 2018). Due to such advances, the lack of need for the construction of clone library has enabled massive parallelization of NGS techniques and has brough about greater

yield of DNA sequence data, providing remarkable insight into the genetic potentials of microbial communities (Sunagawa et al., 2015). This has enhanced the elucidation of metabolic properties of microbial communities, enabling the identification of novel pathways with significant functionalities and applications.

Function-driven metagenomics using the presence of protein coding genes in aquatic samples can help to not only identify what microbial groups are changing within lakes, but also to determine if variations in taxa translate into changes in prevalent protein coding genes that infer microbial functional capabilities with respect to changes in lake environmental conditions. This can help analyse the impact of environmental alterations and water quality on microbial diversity and functions. Therefore, metagenomics is important in conducting systematic analysis of changes occurring in diverse microbial communities within lakes. Metagenomic studies in aquatic ecosystems have revealed profound findings, such as stronger influence of human activities on antibiotic resistant genes (ARGs) and metals resistant genes (MRGs) within coastal areas than those in deep ocean and Antarctic seawater (Y. Yang et al., 2019), and exacerbation of ARGs by veterinary and human antibiotics use in Canadian lakes (Kraemer et al., 2022). Several other novel insights of metagenomics in aquatic microbial ecology have been described in a review by Grossart et al. (2020).

While metagenomics offers valuable insights, it is not without challenges. Generally, reference databases used for classifying microorganisms are limited, leading to unresolved sequence reads. Also, metagenomics only provides information on the potential functional properties of microbial communities based on gene presence, without indicating gene expression levels. To address these limitations, and identify which metabolic genes are actively expressed in a given environment, post-genomic analyses such as metatranscriptomics may be necessary (Aguiar-Pulido et al., 2016).

1.7 Thesis objectives, research questions, and expected contributions to knowledge

Until 2017, there was no standardized nationwide water quality assessment across Canada. Scientists from provincial and federal governments as well as universities across Canada had identified this gap. This led to the formation of an academic-government coalition aimed at filling in the information gap to effectively understand freshwater resources in Canada and thus provide evidence-based insights to their management and protection. The *NSERC* Canadian Lake Pulse

Network was developed in July 2016 and brought together a wide array of experts to investigate the response of lakes to anthropogenic stressors, and to assess how these responses can in turn be used as predictors of ecosystem health. Since LakePulse defines "health status" as the departure ("sickness") of a lake from its natural ("healthy") state, altering its ability to provide the ecosystem services (Huot et al., 2019), an ecosystem transmogrified by human activities is often times unhealthy. Unfortunately, the reality of this century is that freshwater resources are under mounting pressure from accelerated lake eutrophication in agricultural regions due to increased contamination from diverse sources and cyanobacterial blooms. Land use has reportedly caused shifts in landscape properties, affecting how natural lakes function. However, the extent of these changes across Canada is unclear and establishing the link between human activities, lake bacterial diversity, community composition, and function is one of LakePulse's most pressing environmental questions.

The LakePulse field campaign ran over three years (2017 to 2019) and led to the sampling of 664 lakes across 12 Canadian ecozones wherein over 100 variables (describing biological, chemical, optics and lake characteristics) were collected and analysed per lake. This sampling effort provided a massive resource for interdisciplinary lake studies including aquatic microbial ecology. This thesis contributes to the second research theme of the LakePulse network which asks the question "*How are microscopic species affected by lake changes? How can they be used as indicators of lake health*?" By deploying tools in aquatic microbial ecology, specifically metagenomics, this thesis gleans into the epilimnion of over 600 Canadian lakes to decipher the responses of lake bacterial communities to alterations in environmental conditions induced by human activities at the continental scale and regionally.

This thesis is based on the following broad research questions:

1. What are the impacts of watershed environmental conditions on bacterial diversity and community composition in lakes across Canada?
2. How do drivers of bacterial community composition, specifically water quality and land use, vary at the continental scale and within different regions of Canada?
3. How is bacterial community function (i.e. metabolic capacities) impacted by watershed environmental conditions and human land use?

In this thesis I am investigating the following hypotheses:

1. At the continental scale, local environmental and limnological conditions within lakes (e.g. trophic state, nutrient concentrations) will bring about alterations (decline or increase) in bacterial diversity and a shift in taxonomic composition within lakes and across the continental scale.

2. Environmental drivers of bacterial diversity and community composition vary in different regions based on environmental heterogeneity. Microbial groups will respond differently to trophic state and land use types across different Canadian ecozones.

3. Metabolic capabilities of lake bacteria will be influenced by physicochemical properties within lakes and prevalent human land use type with the surrounding lake watershed.

This thesis is presented in three research chapters that explore each of the overarching research questions in detail. This thesis is expected to have resolved changes in bacterial diversity, community composition and functional capabilities consequent on changing environmental conditions and prevalent land use types in 621 Canadian lakes and for the first time and itemize specific bacterial clades as positive or negative indicators of environmental changes. In addition, it is expected to have resolved what metabolic pathways of lake bacteria are most susceptible to either physicochemical conditions or watershed land use type across the continental scale.

Bibliography

Achtman, M., & Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*, *6*(6), Article 6. https://doi.org/10.1038/nrmicro1872

Adrian, R., O'Reilly, C. M., Zagarese, H., Baines, S. B., Hessen, D. O., Keller, W., Livingstone, D. M., Sommaruga, R., Straile, D., Van Donk, E., Weyhenmeyer, G. A., & Winder, M. (2009). Lakes as sentinels of climate change. *Limnology and Oceanography*, *54*(6), 2283–2297.

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., & Narasimhan, G. (2016). Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis: Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data. *Evolutionary Bioinformatics*, *12s1*, EBO.S36436. https://doi.org/10.4137/EBO.S36436

Alberti, M., Booth, D., Hill, K., Coburn, B., Avolio, C., Coe, S., & Spirandelli, D. (2007). The impact of urban patterns on aquatic ecosystems: An empirical analysis in Puget lowland sub-basins. *Landscape and Urban Planning*, *4*(80), 345–361. https://doi.org/10.1016/j.landurbplan.2006.08.001

Alin, S. R., & Johnson, T. C. (2007). Carbon cycling in large lakes of the world: A synthesis of production, burial, and lake-atmosphere exchange estimates. *Global Biogeochemical Cycles*, *21*(3). https://doi.org/10.1029/2006GB002881

Allan, J. D. (2004). Landscapes and Riverscapes: The Influence of Land Use on Stream Ecosystems. *Annual Review of Ecology, Evolution and Systematics*, *35*(Volume 35, 2004), 257–284. https://doi.org/10.1146/annurev.ecolsys.35.120202.110122

Allen, J., Gross, E. M., Courcoul, C., Bouletreau, S., Compin, A., Elger, A., Ferriol, J., Hilt, S., Jassey, V. E. J., Laviale, M., Polst, B. H., Schmitt-Jansen, M., Stibor, H., Vijayaraj, V., & Leflaive, J. (2021). Disentangling the direct and indirect effects of agricultural runoff on freshwater ecosystems subject to global warming: A microcosm study. *Water Research*, *190*, 116713. https://doi.org/10.1016/j.watres.2020.116713

Arnold Jr., C. L., & Gibbons, C. J. (1996). Impervious Surface Coverage: The Emergence of a Key Environmental Indicator. *Journal of the American Planning Association*, *62*(2), 243–258. https://doi.org/10.1080/01944369608975688

Arora-Williams, K., Olesen, S. W., Scandella, B. P., Delwiche, K., Spencer, S. J., Myers, E. M., Abraham, S., Sooklal, A., & Preheim, S. P. (2018). Dynamics of microbial populations mediating biogeochemical cycling in a freshwater lake. *Microbiome*, *6*(1), 165. https://doi.org/10.1186/s40168-018-0556-7

Bani, A., Randall, K. C., Clark, D. R., Gregson, B. H., Henderson, D. K., Losty, E. C., & Ferguson, R. M. W. (2022). Chapter Six - Mind the gaps: What do we know about how multiple chemical stressors impact freshwater aquatic microbiomes? In D. A. Bohan & A. Dumbrell (Eds.), *Advances in Ecological Research* (Vol. 67, pp. 331–377). Academic Press. https://doi.org/10.1016/bs.aecr.2022.09.003

Bastviken, D., Tranvik, L. J., Downing, J. A., Crill, P. M., & Enrich-Prast, A. (2011). Freshwater Methane Emissions Offset the Continental Carbon Sink. *Science*, *331*(6013), 50–50. https://doi.org/10.1126/science.1196808

Belt, K. T., Hohn, C., Gbakima, A., & Higgins, J. A. (2007). Identification of culturable stream water bacteria from urban, agricultural, and forested watersheds using 16S rRNA gene sequencing. *Journal of Water and Health*, *5*(3), 395–406. https://doi.org/10.2166/wh.2007.035

Berman, T. (1990). Microbial Food-Webs and Nutrient Cycling in Lakes: Changing Perspectives. In M. M. Tilzer & C. Serruya (Eds.), *Large Lakes: Ecological Structure and Function* (pp. 511–525). Springer. https://doi.org/10.1007/978-3-642-84077-7_27

Booth, D. B., & Jackson, C. R. (1997). Urbanization of Aquatic Systems: Degradation Thresholds, Stormwater Detection, and the Limits of Mitigation1. *JAWRA Journal of the American Water Resources Association*, *33*(5), 1077–1090. https://doi.org/10.1111/j.1752-1688.1997.tb04126.x

Brown, J. H., & Lomolino, M. (1998). Biogeography. 2nd Ed. Sunderland, Massachusetts (Sinauer Associates, Inc. Publishers). 691 S. ISBN 0-87893-073-6

Burns, C. W., & Galbraith, L. M. (2007). Relating planktonic microbial food web structure in lentic freshwater ecosystems to water quality and land use. *Journal of Plankton Research*, *29*(2), 127–139. https://doi.org/10.1093/plankt/fbm001

Bussmann, I., Philipp, B., & Schink, B. (2001). Factors influencing the cultivability of lake water bacteria. *Journal of Microbiological Methods*, *47*(1), 41–50. https://doi.org/10.1016/S0167-7012(01)00289-5

Butcher, J. B., Nover, D., Johnson, T. E., & Clark, C. M. (2015). Sensitivity of lake thermal and mixing dynamics to climate change. *Climatic Change*, *129*(1), 295–305. https://doi.org/10.1007/s10584-015-1326-1

Butman, D., Stackpoole, S., Stets, E., McDonald, C. P., Clow, D. W., & Striegl, R. G. (2016). Aquatic carbon cycling in the conterminous United States and implications for terrestrial carbon accounting. *Proceedings of the National Academy of Sciences*, *113*(1), 58–63. https://doi.org/10.1073/pnas.1512651112

Carpenter, S. R., Benson, B. J., Biggs, R., Chipman, J. W., Foley, J. A., Golding, S. A., Hammer, R. B., Hanson, P. C., Johnson, P. T. J., Kamarainen, A. M., Kratz, T. K., Lathrop, R. C., McMahon, K. D., Provencher, B., Rusak, J. A., Solomon, C. T., Stanley, E. H., Turner, M. G., Vander Zanden, M. J., … Yuan, H. (2007). Understanding Regional Change: A Comparison of Two Lake Districts. *BioScience*, *57*(4), 323–335. https://doi.org/10.1641/B570407

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Applied and Environmental Microbiology*, *73*(1), 278–288. https://doi.org/10.1128/AEM.01177-06

Castellano-Hinojosa, A., Correa-Galeote, D., Carrillo, P., Bedmar, E. J., & Medina-Sánchez, J. M. (2017). Denitrification and Biodiversity of Denitrifiers in a High-Mountain Mediterranean Lake. *Frontiers in Microbiology*, *8*, 1911. https://doi.org/10.3389/fmicb.2017.01911

CCEA. (2016). Ecozone introduction. de http://www.ccea.org/ecozones-introduction/

Chen, W., Wilkes, G., Khan, I. U. H., Pintar, K. D. M., Thomas, J. L., Lévesque, C. A., Chapados, J. T., Topp, E., & Lapen, D. R. (2018). Aquatic Bacterial Communities Associated With Land Use and Environmental Factors in Agricultural Landscapes Using a Metabarcoding Approach. *Frontiers in Microbiology*, *9*. https://www.frontiersin.org/articles/10.3389/fmicb.2018.02301

Clement, T. A., Murry, B. A., & Uzarski, D. G. (2015). Fish community size structure of small lakes: The role of lake size, biodiversity and disturbance. *Journal of Freshwater Ecology*, *30*(4), 557–568. https://doi.org/10.1080/02705060.2015.1030787

Comeau, A. M., Harding, T., Galand, P. E., Vincent, W. F., & Lovejoy, C. (2012). Vertical distribution of microbial communities in a perennially stratified Arctic lake with saline, anoxic bottom waters. *Scientific Reports*, *2*(1), Article 1. https://doi.org/10.1038/srep00604

Comte, J., & del Giorgio, P. A. (2010). Linking the patterns of change in composition and function in bacterioplankton successions along environmental gradients. *Ecology*, *91*(5), 1466–1476. https://doi.org/10.1890/09-0848.1

Copetti, D., Guyennon, N., & Buzzi, F. (2020). Generation and dispersion of chemical and biological gradients in a large-deep multi-basin lake (Lake Como, north Italy): The joint effect of external drivers and internal wave motions. *Science of The Total Environment*, *749*, 141587. https://doi.org/10.1016/j.scitotenv.2020.141587

Cotner, J. B., & Biddanda, B. A. (2002). Small Players, Large Role: Microbial Influence on Biogeochemical Processes in Pelagic Aquatic Ecosystems. *Ecosystems*, *5*(2), 105–121. https://doi.org/10.1007/s10021-001-0059-3

Craig, L. S., Olden, J. D., Arthington, A. H., Entrekin, S., Hawkins, C. P., Kelly, J. J., Kennedy, T. A., Maitland, B. M., Rosi, E. J., Roy, A. H., Strayer, D. L., Tank, J. L., West, A. O., & Wooten, M. S. (2017). Meeting the challenge of interacting threats in freshwater ecosystems: A call to scientists and managers. *Elementa: Science of the Anthropocene*, *5*, 72. https://doi.org/10.1525/elementa.256

Crosby, L. D., & Criddle, C. S. (2003). Understanding Bias in Microbial Community Analysis Techniques due to rrn Operon Copy Number Heterogeneity. *BioTechniques*, *34*(4), 790–802. https://doi.org/10.2144/03344rr01

Crossman, J., Eimers, M. C., Kerr, J. G., & Yao, H. (2016). Sensitivity of physical lake processes to climate change within a large Precambrian Shield catchment. *Hydrological Processes*, *30*(23), 4353–4366. https://doi.org/10.1002/hyp.10915

Crump, B. C., Kling, G. W., Bahr, M., & Hobbie, J. E. (2003). Bacterioplankton community shifts in an Arctic lake correlate with seasonal changes in organic matter source. *Applied and Environmental Microbiology*, *69*(4), 2253–2268. Scopus. https://doi.org/10.1128/AEM.69.4.2253-2268.2003

De Wever, A., Muylaert, K., Van der Gucht, K., Pirlot, S., Cocquyt, C., Descy, J.-P., Plisnier, P.-D., & Vyverman, W. (2005). Bacterial Community Composition in Lake Tanganyika: Vertical and Horizontal Heterogeneity. *Applied and Environmental Microbiology*, *71*(9), 5029–5037. https://doi.org/10.1128/AEM.71.9.5029-5037.2005

DeBruyn, J. M., Leigh-Bell, J. A., McKay, R. M. L., Bourbonniere, R. A., & Wilhelm, S. W. (2004). Microbial Distributions and the Impact of Phosphorus on Bacterial Activity in Lake Erie. *Journal of Great Lakes Research*, *30*(1), 166–183. https://doi.org/10.1016/S0380-1330(04)70338-X

DelSontro, T., Beaulieu, J. J., & Downing, J. A. (2018). Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change. *Limnology and Oceanography Letters*, *3*(3), 64–75. https://doi.org/10.1002/lol2.10073

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z.-I., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A.-H., Soto, D., Stiassny, M. L. J., & Sullivan, C. A. (2006). Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biological Reviews*, *81*(2), 163–182. https://doi.org/10.1017/S1464793105006950

Dyall-Smith, M., & Oren, A. (2006). Culture-Dependent Study of Microbial Diversity of Lake Chaka. *Applied and Environmental Microbiology*, *72*(11), 7427–7427. https://doi.org/10.1128/AEM.01401-06

Egert, M., & Friedrich, M. W. (2003). Formation of Pseudo-Terminal Restriction Fragments, a PCR-Related Bias Affecting Terminal Restriction Fragment Length Polymorphism Analysis of Microbial Community Structure. *Applied and Environmental Microbiology*, *69*(5), 2555–2562. https://doi.org/10.1128/AEM.69.5.2555-2562.2003

Fuhrman, J. A., McCallum, K., & Davis, A. A. (1992). Novel major archaebacterial group from marine plankton. *Nature*, *356*(6365), Article 6365. https://doi.org/10.1038/356148a0

Garner, R. E., Kraemer, S. A., Onana, V. E., Fradette, M., Varin, M.-P., Huot, Y., & Walsh, D. A. (2023). A genome catalogue of lake bacterial diversity and its drivers at continental scale. *Nature Microbiology*, *8*(10), Article 10. https://doi.org/10.1038/s41564-023-01435-6

Garner, R. E., Kraemer, S. A., Onana, V. E., Huot, Y., Gregory-Eaves, I., & Walsh, D. A. (2022). Protist Diversity and Metabolic Strategy in Freshwater Lakes Are Shaped by Trophic State and Watershed Land Use on a Continental Scale. *mSystems*, *0*(0), e00316-22. https://doi.org/10.1128/msystems.00316-22

Gayer, P. R., Guimarães, P. S., Albertoni, E. F., & Hepp, L. U. (2021). Bacterial decomposition of allochthonous organic matter in shallow subtropical lakes. *Lakes & Reservoirs: Science, Policy and Management for Sustainable Use*, *26*(4), e12389. https://doi.org/10.1111/lre.12389

Gibbs, K. E., Mackey, R. L., & Currie, D. J. (2009). Human land use, agriculture, pesticides and losses of imperiled species. *Diversity and Distributions*, *15*(2), 242–253. https://doi.org/10.1111/j.1472-4642.2008.00543.x

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, *345*(6270), Article 6270. https://doi.org/10.1038/345060a0

Glińska-Lewczuk, K., Gołaś, I., Koc, J., Gotkowska-Płachta, A., Harnisz, M., & Rochwerger, A. (2016). The impact of urban areas on the water quality gradient along a lowland river. *Environmental Monitoring and Assessment*, *188*(11), 624. https://doi.org/10.1007/s10661-016-5638-z

Glöckner, F. O., Zaichikov, E., Belkova, N., Denissova, L., Pernthaler, J., Pernthaler, A., & Amann, R. (2000). Comparative 16S rRNA Analysis of Lake Bacterioplankton Reveals Globally Distributed Phylogenetic Clusters Including an Abundant Group of Actinobacteria. *Applied and Environmental Microbiology*, *66*(11), 5053–5065. https://doi.org/10.1128/AEM.66.11.5053-5065.2000

Gonzalez Mateu, M., Park, C. E., McAskill, C. P., Baldwin, A. H., & Yarwood, S. A. (2019). Urbanization Altered Bacterial and Archaeal Composition in Tidal Freshwater Wetlands Near Washington DC, USA, and Buenos Aires, Argentina. *Microorganisms*, *7*(3), Article 3. https://doi.org/10.3390/microorganisms7030072

Grimm, N. B., Grove, J. G., Pickett, S. T. A., & Redman, C. L. (2000). Integrated Approaches to Long-Term Studies of Urban Ecological Systems: Urban ecological systems present multiple challenges to ecologists—pervasive human impact and extreme heterogeneity of cities, and the need to integrate social and ecological approaches, concepts, and theory. *BioScience*, *50*(7), 571–584. https://doi.org/10.1641/0006-3568(2000)050[0571:IATLTO]2.0.CO;2

Grossart, H.-P., Massana, R., McMahon, K. D., & Walsh, D. A. (2020). Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnology and Oceanography*, *65*(S1), S2–S20. https://doi.org/10.1002/lno.11382

Hale, R. L., Turnbull, L., Earl, S. R., Childers, D. L., & Grimm, N. B. (2015). Stormwater Infrastructure Controls Runoff and Dissolved Material Export from Arid Urban Watersheds. *Ecosystems*, *18*(1), 62–75. https://doi.org/10.1007/s10021-014-9812-2

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249. https://doi.org/10.1016/S1074-5521(98)90108-9

Hengy, M. H., Horton, D. J., Uzarski, D. G., & Learman, D. R. (2017). Microbial community diversity patterns are related to physical and chemical differences among temperate lakes near Beaver Island, MI. *PeerJ*, *5*, e3937. https://doi.org/10.7717/peerj.3937

Horton, D. J., Theis, K. R., Uzarski, D. G., & Learman, D. R. (2019). Microbial community structure and microbial networks correspond to nutrient gradients within coastal wetlands of the Laurentian Great Lakes. *FEMS Microbiology Ecology*, *95*(4), fiz033. https://doi.org/10.1093/femsec/fiz033

Hosen, J. D., Febria, C. M., Crump, B. C., & Palmer, M. A. (2017). Watershed Urbanization Linked to Differences in Stream Bacterial Community Composition. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01452

Huot, Y., Brown, C. A., Potvin, G., Antoniades, D., Baulch, H. M., Beisner, B. E., Bélanger, S., Brazeau, S., Cabana, H., Cardille, J. A., del Giorgio, P. A., Gregory-Eaves, I., Fortin, M.-J., Lang, A. S., Laurion, I., Maranger, R., Prairie, Y. T., Rusak, J. A., Segura, P. A., … Walsh, D. A. (2019). The NSERC Canadian Lake Pulse Network: A national assessment of lake health providing science for water management in a changing climate. *Science of The Total Environment*, *695*, 133668. https://doi.org/10.1016/j.scitotenv.2019.133668

Ives, J. T., McMeans, B. C., McCann, K. S., Fisk, A. T., Johnson, T. B., Bunnell, D. B., Frank, K. T., & Muir, A. M. (2019). Food-web structure and ecosystem function in the Laurentian Great Lakes—Toward a conceptual model. *Freshwater Biology*, *64*(1), 1–23. https://doi.org/10.1111/fwb.13203

Jankowiak, J., Hattenrath-Lehmann, T., Kramer, B. J., Ladds, M., & Gobler, C. J. (2019). Deciphering the effects of nitrogen, phosphorus, and temperature on cyanobacterial bloom intensification, diversity, and toxicity in western Lake Erie. *Limnology and Oceanography*, *64*(3), 1347–1370. https://doi.org/10.1002/lno.11120

Jeziorski, A., Yan, N. D., Paterson, A. M., DeSellas, A. M., Turner, M. A., Jeffries, D. S., Keller, B., Weeber, R. C., McNicol, D. K., Palmer, M. E., McIver, K., Arseneau, K., Ginn, B. K., Cumming, B. F., & Smol, J. P. (2008). The Widespread Threat of Calcium Decline in Fresh Waters. *Science*, *322*(5906), 1374–1377. https://doi.org/10.1126/science.1164949

Jurasinski, G., & Koch, M. (2011). Commentary: Do we have a consistent terminology for species diversity? We are on the way. *Oecologia*, *167*(4), 893–902. https://doi.org/10.1007/s00442-011-2126-6

Karamigolbaghi, M., Feng, Y., Atkinson, J. F., Verhamme, E. M., & Boyer, G. L. (2019). Circulation and mixing in Sodus Bay due to water exchange with Lake Ontario. *Journal of Great Lakes Research*, *45*(6), 1090–1102. https://doi.org/10.1016/j.jglr.2019.09.020

Keatley, B. E., Bennett, E. M., MacDonald, G. K., Taranu, Z. E., & Gregory-Eaves, I. (2011). Land-Use Legacies Are Important Determinants of Lake Eutrophication in the Anthropocene. *PLOS ONE*, *6*(1), e15913. https://doi.org/10.1371/journal.pone.0015913

Keshri, J., Pradeep Ram, A. S., Nana, P. A., & Sime-Ngando, T. (2018). Taxonomical Resolution and Distribution of Bacterioplankton Along the Vertical Gradient Reveals Pronounced Spatiotemporal Patterns in Contrasted Temperate Freshwater Lakes. *Microbial Ecology*, *76*(2), 372–386. https://doi.org/10.1007/s00248-018-1143-y

Kiersztyn, B., Chróst, R., Kaliński, T., Siuda, W., Bukowska, A., Kowalczyk, G., & Grabowska, K. (2019). Structural and functional microbial diversity along a eutrophication gradient of interconnected lakes undergoing anthropopressure. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-47577-8

Konstantinidis, K. T., Ramette, A., & Tiedje, J. M. (2006). The bacterial species definition in the genomic era. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1475), 1929–1940. https://doi.org/10.1098/rstb.2006.1920

Kraemer, S. A., Barbosa da Costa, N., Oliva, A., Huot, Y., & Walsh, D. A. (2022). A resistome survey across hundreds of freshwater bacterial communities reveals the impacts of veterinary and human antibiotics use. *Frontiers in Microbiology*, *13*. https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2022.995418

Kraemer, S. A., Barbosa da Costa, N., Shapiro, B. J., Fradette, M., Huot, Y., & Walsh, D. A. (2020). A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. *The ISME Journal*, *14*(12), Article 12. https://doi.org/10.1038/s41396-020-0733-0

Li, X., Yu, R., Wang, J., Sun, H., Liu, X., Ren, X., Zhuang, S., Guo, Z., & Lu, X. (2024). Greenhouse gas emissions from Daihai Lake, China: Should eutrophication and salinity promote carbon emission dynamics? *Journal of Environmental Sciences*, *135*, 407–423. https://doi.org/10.1016/j.jes.2022.12.021

Lindström, E. S., Feng, X. M., Granéli, W., & Kritzberg, E. S. (2010). The interplay between bacterial community composition and the environment determining function of inland water bacteria. *Limnology and Oceanography*, *55*(5), 2052–2060. https://doi.org/10.4319/lo.2010.55.5.2052

Lindström, E. S., Forslund, M., Algesten, G., & Bergström, A.-K. (2006). External control of bacterial community structure in lakes. *Limnology and Oceanography*, *51*(1 I), 339–342. Scopus. https://doi.org/10.4319/lo.2006.51.1.0339

Lindström, E. S., Kamst-Van Agterveld, M. P., & Zwart, G. (2005). Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. *Applied and Environmental Microbiology*, *71*(12), 8201–8206. Scopus. https://doi.org/10.1128/AEM.71.12.8201-8206.2005

Marmen, S., Blank, L., Al-Ashhab, A., Malik, A., Ganzert, L., Lalzar, M., Grossart, H.-P., & Sher, D. (2020). The Role of Land Use Types and Water Chemical Properties in Structuring the Microbiomes of a Connected Lake System. *Frontiers in Microbiology*, *11*. https://www.frontiersin.org/articles/10.3389/fmicb.2020.00089

Martiny, A. C. (2019). High proportions of bacteria are culturable across major biomes. *The ISME Journal*, *13*(8), 2125–2128. https://doi.org/10.1038/s41396-019-0410-3

Marzluff, J. M. (2001). Worldwide urbanization and its effects on birds. In J. M. Marzluff, R. Bowman, & R. Donnelly (Eds.), *Avian Ecology and Conservation in an Urbanizing World* (pp. 19–47). Springer US. https://doi.org/10.1007/978-1-4615-1531-9_2

Medeiros, J. D., Cantão, M. E., Cesar, D. E., Nicolás, M. F., Diniz, C. G., Silva, V. L., Vasconcelos, A. T. R. de, & Coelho, C. M. (2016). Comparative metagenome of a stream impacted by the urbanization phenomenon. *Brazilian Journal of Microbiology*, *47*(4), 835–845. https://doi.org/10.1016/j.bjm.2016.06.011

Methé, B. A., & Zehr, J. P. (1999). Diversity of bacterial communities in Adirondack lakes: Do species assemblages reflect lake water chemistry? *Hydrobiologia*, *401*, 77–96. Scopus. https://doi.org/10.1023/a:1003782209607

Minns, C. K., Moore, J. E., Shuter, B. J., & Mandrak, N. E. (2008). A preliminary national analysis of some key characteristics of Canadian lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(8), 1763–1778. https://doi.org/10.1139/F08-110

https://doi.org/10.1016/j.jglr.2013.03.014

Mullins, T. D., Britschgi, T. B., Krest, R. L., & Giovannoni, S. J. (1995). Genetic comparisons reveal the same unknown bacterial lineages in Atlantic and Pacific bacterioplankton communities. *Limnology and Oceanography*, *40*(1), 148–158. https://doi.org/10.4319/lo.1995.40.1.0148

Nanayakkara, L. (2018). Prairie Lakes: Model Systems for Integrating Ecology and Human Dimensions for Better Management. *Limnology and Oceanography Bulletin*, *27*(4), 110–113. https://doi.org/10.1002/lob.10274

Nekola, J. C., & White, P. S. (1999). The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, *26*(4), 867–878. https://doi.org/10.1046/j.1365-2699.1999.00305.x

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews: MMBR*, *75*(1), 14–49. https://doi.org/10.1128/MMBR.00028-10

Newton, R. J., Jones, S. E., Helmus, M. R., & McMahon, K. D. (2007). Phylogenetic Ecology of the Freshwater Actinobacteria acI Lineage. *Applied and Environmental Microbiology*, *73*(22), 7169–7176. https://doi.org/10.1128/AEM.00794-07

Newton, R. J., & McLellan, S. L. (2015). A unique assemblage of cosmopolitan freshwater bacteria and higher community diversity differentiate an urbanized estuary from oligotrophic Lake Michigan. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.01028

Nguyen, J., Lara-Gutiérrez, J., & Stocker, R. (2021). Environmental fluctuations and their effects on microbial communities, populations and individuals. *FEMS Microbiology Reviews*, *45*(4), fuaa068. https://doi.org/10.1093/femsre/fuaa068

Niño-García, J. P., Ruiz-González, C., & del Giorgio, P. A. (2016). Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *The ISME Journal*, *10*(7), 1755–1766. https://doi.org/10.1038/ismej.2015.226

Obieze, C. C., Wani, G. A., Shah, M. A., Reshi, Z. A., Comeau, A. M., & Khasa, D. P. (2022). Anthropogenic activities and geographic locations regulate microbial diversity, community assembly and species sorting in Canadian and Indian freshwater lakes. *Science of The Total Environment*, *826*, 154292. https://doi.org/10.1016/j.scitotenv.2022.154292

Ormerod, S. J., Dobson, M., Hildrew, A. G., & Townsend, C. R. (2010). Multiple stressors in freshwater ecosystems. *Freshwater Biology*, *55*(s1), 1–4. https://doi.org/10.1111/j.1365-2427.2009.02395.x

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C., & Iliopoulos, loannis. (2015). Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinformatics and Biology Insights*, *9*, BBI.S12462. https://doi.org/10.4137/BBI.S12462

Ozbayram, E. G., Koker, L., Akçaalan, R., Aydın, F., Ertürk, A., Ince, O., & Albay, M. (2021). Contrasting the Water Quality and Bacterial Community Patterns in Shallow and Deep Lakes: Manyas vs. Iznik. *Environmental Management*, *67*(3), 506–512. https://doi.org/10.1007/s00267-020-01357-7

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science (New York, N.Y.)*, *276*(5313), 734–740. https://doi.org/10.1126/science.276.5313.734

Paruch, L., Paruch, A. M., Eiken, H. G., Skogen, M., & Sørheim, R. (2020). Seasonal dynamics of lotic bacterial communities assessed by 16S rRNA gene amplicon deep sequencing. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-73293-9

Paver, S. F., Newton, R. J., & Coleman, M. L. (2020). Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environmental Microbiology*, *22*(1), 433–446. https://doi.org/10.1111/1462-2920.14862

Pearce, D. A. (2008). Climate change and the microbiology of the Antarctic Peninsula region. *Science Progress*, *91*(2), 203–217. Scopus. https://doi.org/10.3184/003685008X332534

Pearce, D. A., van der Gast, C. J., Lawley, B., & Ellis-Evans, J. C. (2003). Bacterioplankton community diversity in a maritime Antarctic lake, determined by culture-dependent and

culture-independent techniques. *FEMS Microbiology Ecology*, *45*(1), 59–70. https://doi.org/10.1016/S0168-6496(03)00110-7

Peter, H., Beier, S., Bertilsson, S., Lindström, E. S., Langenheder, S., & Tranvik, L. J. (2011). Function-specific response to depletion of microbial diversity. *The ISME Journal*, *5*(2), 351–361. https://doi.org/10.1038/ismej.2010.119

Pham, S. V., Leavitt, P. R., McGowan, S., & Peres-Neto, P. (2008). Spatial variability of climate and land-use effects on lakes of the northern Great Plains. *Limnology and Oceanography*, *53*(2), 728–742. https://doi.org/10.4319/lo.2008.53.2.0728

Pilla, R. M., Williamson, C. E., Adamovich, B. V., Adrian, R., Anneville, O., Chandra, S., Colom-Montero, W., Devlin, S. P., Dix, M. A., Dokulil, M. T., Gaiser, E. E., Girdner, S. F., Hambright, K. D., Hamilton, D. P., Havens, K., Hessen, D. O., Higgins, S. N., Huttula, T. H., Huuskonen, H., … Zadereev, E. (2020). Deeper waters are changing less consistently than surface waters in a global analysis of 102 lakes. *Scientific Reports*, *10*(1), 20514. https://doi.org/10.1038/s41598-020-76873-x

Pinder, K. C., Eimers, M. C., & Watmough, S. A. (2014). Impact of wetland disturbance on phosphorus loadings to lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *71*(11), 1695–1703. https://doi.org/10.1139/cjfas-2014-0143

Polz, M. F., & Cavanaugh, C. M. (1998). Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied and Environmental Microbiology*, *64*(10), 3724–3730. https://doi.org/10.1128/AEM.64.10.3724-3730.1998

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial Community Dynamics. *PLOS ONE*, *9*(4), e93827. https://doi.org/10.1371/journal.pone.0093827

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), Article 9. https://doi.org/10.1038/nbt.3935

Quinlan, R., Leavitt, P. R., Dixit, A. S., Hall, R. I., & Smol, J. P. (2002). Landscape effects of climate, agriculture, and urbanization on benthic invertebrate communities of Canadian prairie lakes. *Limnology and Oceanography*, *47*(2), 378–391. https://doi.org/10.4319/lo.2002.47.2.0378

Quinlan, R., Paterson, A. M., Hall, R. I., Dillon, P. J., Wilkinson, A. N., Cumming, B. F., Douglas, M. S. V., & Smol, J. P. (2003). A landscape approach to examining spatial patterns of limnological variables and long-term environmental change in a southern Canadian lake district. *Freshwater Biology*, *48*(9), 1676–1697. https://doi.org/10.1046/j.1365-2427.2003.01105.x

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T. J., Kidd, K. A., MacCormack, T. J., Olden, J. D., Ormerod, S. J., Smol, J. P., Taylor, W. W., Tockner, K., Vermaire, J. C., Dudgeon, D., & Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews*, *94*(3), 849–873. https://doi.org/10.1111/brv.12480

Reis, P. C. J., Thottathil, S. D., & Prairie, Y. T. (2022). The role of methanotrophy in the microbial carbon metabolism of temperate lakes. *Nature Communications*, *13*(1), 43. https://doi.org/10.1038/s41467-021-27718-2

Riera, JoaN. L., Magnuson, JohN. J., Kratz, TiM. K., & Webster, K. E. (2000). A geomorphic template for the analysis of lake districts applied to the Northern Highland Lake District, Wisconsin, U.S.A. *Freshwater Biology*, *43*(3), 301–318. https://doi.org/10.1046/j.1365-2427.2000.00567.x

Ritchie, H., Samborska V., & Roser, M. (2018). “Urbanization” Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/urbanization' [Online Resource]

Robertson, D. M., Saad, D. A., Benoy, G. A., Vouk, I., Schwarz, G. E., & Laitta, M. T. (2019). Phosphorus and Nitrogen Transport in the Binational Great Lakes Basin Estimated Using SPARROW Watershed Models. *JAWRA Journal of the American Water Resources Association*, *55*(6), 1401–1424. https://doi.org/10.1111/1752-1688.12792

Sadeghi, J., Chaganti, S. R., Shahraki, A. H., & Heath, D. D. (2021). Microbial community and abiotic effects on aquatic bacterial communities in north temperate lakes. *Science of The Total Environment*, *781*, 146771. https://doi.org/10.1016/j.scitotenv.2021.146771

Salcher, M. M., Pernthaler, J., & Posch, T. (2010). Spatiotemporal distribution and activity patterns of bacteria from three phylogenetic groups in an oligomesotrophic lake. *Limnology and Oceanography*, *55*(2), 846–856. https://doi.org/10.4319/lo.2010.55.2.0846

Schindler, D. W., Hecky, R. E., & McCullough, G. K. (2012). The rapid eutrophication of Lake Winnipeg: Greening under global change. *Journal of Great Lakes Research*, *38*, 6–13. https://doi.org/10.1016/j.jglr.2012.04.003

Seitzinger, S., Harrison, J. A., Böhlke, J. K., Bouwman, A. F., Lowrance, R., Peterson, B., Tobias, C., & Van Drecht, G. (2006). Denitrification across landscapes and waterscapes: A synthesis. *Ecological Applications: A Publication of the Ecological Society of America*, *16*(6), 2064–2090. https://doi.org/10.1890/1051-0761(2006)016[2064:dalawa]2.0.co;2

Shade, A., Peter, H., Allison, S. D., Baho, D., Berga, M., Buergmann, H., Huber, D. H., Langenheder, S., Lennon, J. T., Martiny, J. B., Matulich, K. L., Schmidt, T. M., & Handelsman, J. (2012). Fundamentals of Microbial Community Resistance and Resilience. *Frontiers in Microbiology*, *3*. https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2012.00417

Shafi, S., Kamili, A. N., Shah, M. A., Parray, J. A., & Bandh, S. A. (2017). Aquatic bacterial diversity: Magnitude, dynamics, and controlling factors. *Microbial Pathogenesis*, *104*, 39–47. https://doi.org/10.1016/j.micpath.2017.01.016

Shahraki, A. H., Chaganti, S. R., & Heath, D. (2021). Spatio-temporal dynamics of bacterial communities in the shoreline of Laurentian great Lake Erie and Lake St. Clair’s large freshwater ecosystems. *BMC Microbiology*, *21*, 253. https://doi.org/10.1186/s12866-021-02306-y

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Stadler, M., Ejarque, E., & Kainz, M. J. (2020). In-lake transformations of dissolved organic matter composition in a subalpine lake do not change its biodegradability. *Limnology and Oceanography*, *65*(7), 1554–1572. https://doi.org/10.1002/lno.11406

Stehle, S., & Schulz, R. (2015). Agricultural insecticides threaten surface waters at the global scale. *Proceedings of the National Academy of Sciences*, *112*(18), 5750–5755. https://doi.org/10.1073/pnas.1500232112

Steinman, A. D., Cardinale, B. J., Munns, W. R., Ogdahl, M. E., Allan, J. D., Angadi, T., Bartlett, S., Brauman, K., Byappanahalli, M., Doss, M., Dupont, D., Johns, A., Kashian, D., Lupi, F., McIntyre, P., Miller, T., Moore, M., Muenich, R. L., Poudel, R., … Washburn, E. (2017). Ecosystem services in the Great Lakes. *Journal of Great Lakes Research*, *43*(3), 161–168. https://doi.org/10.1016/j.jglr.2017.02.004

Sterner, R. W., Keeler, B., Polasky, S., Poudel, R., Rhude, K., & Rogers, M. (2020). Ecosystem services of Earth's largest freshwater lakes. *Ecosystem Services*, *41*, 101046. https://doi.org/10.1016/j.ecoser.2019.101046

Straškrábová, V., Izmest'yeva, L. R., Maksimova, E. A., Fietz, S., Nedoma, J., Borovec, J., Kobanova, G. I., Shchetinina, E. V., & Pislegina, E. V. (2005). Primary production and microbial activity in the euphotic zone of Lake Baikal (Southern Basin) during late winter. *Global and Planetary Change*, *46*(1), 57–73. https://doi.org/10.1016/j.gloplacha.2004.11.006

Strayer, D. L., & Dudgeon, D. (2010). Freshwater biodiversity conservation: Recent progress and future challenges. *Journal of the North American Benthological Society*, *29*(1), 344–358. https://doi.org/10.1899/08-171.1

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., … Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1261359. https://doi.org/10.1126/science.1261359

Taranu, Z. E., & Gregory-Eaves, I. (2008). Quantifying Relationships Among Phosphorus, Agriculture, and Lake Depth at an Inter-Regional Scale. *Ecosystems*, *11*(5), 715–725. https://doi.org/10.1007/s10021-008-9153-0

Taranu, Z. E., Köster, D., Hall, R. I., Charette, T., Forrest, F., Cwynar, L. C., & Gregory-Eaves, I. (2010). Contrasting responses of dimictic and polymictic lakes to environmental change: A spatial and temporal study. *Aquatic Sciences*, *72*(1), 97–115. https://doi.org/10.1007/s00027-009-0120-4

Toming, K., Kotta, J., Uuemaa, E., Sobek, S., Kutser, T., & Tranvik, L. J. (2020). Predicting lake dissolved organic carbon at a global scale. *Scientific Reports*, *10*(1), 8471. https://doi.org/10.1038/s41598-020-65010-3

Van der Gucht, K., Cottenie, K., Muylaert, K., Vloemans, N., Cousin, S., Declerck, S., Jeppesen, E., Conde-Porcuna, J.-M., Schwenk, K., Zwart, G., Degans, H., Vyverman, W., & De Meester, L. (2007). The power of species sorting: Local factors drive bacterial community composition over a wide range of spatial scales. *Proceedings of the National Academy of Sciences*, *104*(51), 20404–20409. https://doi.org/10.1073/pnas.0707200104

Vaz-Moreira, I., Egas, C., Nunes, O. C., & Manaia, C. M. (2011). Culture-dependent and culture-independent diversity surveys target different bacteria: A case study in a freshwater sample. *Antonie van Leeuwenhoek*, *100*(2), 245–257. https://doi.org/10.1007/s10482-011-9583-0

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S. E., Sullivan, C. A., Liermann, C. R., & Davies, P. M. (2010). Global threats to human water security and river biodiversity. *Nature*, *467*(7315), 555–561. https://doi.org/10.1038/nature09440

Wang, L., Zhang, J., Li, H., Yang, H., Peng, C., Peng, Z., & Lu, L. (2018). Shift in the microbial community composition of surface water and sediment along an urban river. *Science of The Total Environment*, *627*, 600–612. https://doi.org/10.1016/j.scitotenv.2018.01.203

Wang, S., Zhao, D., Zeng, J., Xu, H., Huang, R., Jiao, C., & Guo, L. (2019). Variations of bacterial community during the decomposition of Microcystis under different temperatures and biomass. *BMC Microbiology*, *19*(1), 207. https://doi.org/10.1186/s12866-019-1585-5

Wang, S.-Y., Sudduth, E. B., Wallenstein, M. D., Wright, J. P., & Bernhardt, E. S. (2011). Watershed Urbanization Alters the Composition and Function of Stream Bacterial Communities. *PLOS ONE*, *6*(8), e22972. https://doi.org/10.1371/journal.pone.0022972

Wetzel, R.G. (2001). Limnology Lake and Reservoir Ecosystems. *Academic Press*, San Diego.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, *95*(12), 6578–6583. https://doi.org/10.1073/pnas.95.12.6578

Wiken, E. B. (1986). Terrestrial EcoZones of Canada. *Ecological Land Classification Series* No. 19. Hull.

Wiken, E. B., Gauthier, D., Marshall, I., Lawton, K., & Hirvonen, H. (1996). A perspective on Canada's ecosystems: An overview of the terrestrial and marine ecozones. In *Canadian Council on Ecological Areas Occasional Papers* (Vol. 14).

Williamson, C. E., Saros, J. E., Vincent, W. F., & Smol, J. P. (2009). Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnology and Oceanography*, *54*(6part2), 2273–2282. https://doi.org/10.4319/lo.2009.54.6_part_2.2273

Woese, C.R. (1987). Bacterial evolution. *Microbiol Rev*51:.https://doi-org.lib-ezproxy.concordia.ca/10.1128/mr.51.2.221-271.1987

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, *74*(11), 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A Primer on Metagenomics. *PLOS Computational Biology*, *6*(2), e1000667. https://doi.org/10.1371/journal.pcbi.1000667

Woolway, R. I., Sharma, S., & Smol, J. P. (2022). Lakes in Hot Water: The Impacts of a Changing Climate on Aquatic Ecosystems. *BioScience*, *72*(11), 1050–1061. https://doi.org/10.1093/biosci/biac052

Wu, S., Wu, Z., Liang, Z., Liu, Y., & Wang, Y. (2019). Denitrification and the controlling factors in Yunnan Plateau Lakes (China): Exploring the role of enhanced internal nitrogen cycling by algal blooms. *Journal of Environmental Sciences*, *76*, 349–358. https://doi.org/10.1016/j.jes.2018.05.028

Yan, N. D., Paterson, A. M., Somers, K. M., & Scheider, W. A. (2008). An introduction to the Dorset special issue: Transforming understanding of factors that regulate aquatic ecosystems on the southern Canadian Shield. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(5), 781–785. https://doi.org/10.1139/f08-077

Yang, Y., Li, Z., Song, W., Du, L., Ye, C., Zhao, B., Liu, W., Deng, D., Pan, Y., Lin, H., & Cao, X. (2019). Metagenomic insights into the abundance and composition of resistance genes in aquatic environments: Influence of stratification and geography. *Environment International*, *127*, 371–380. https://doi.org/10.1016/j.envint.2019.03.062

Yannarell, A. C., & Triplett, E. W. (2005). Geographic and Environmental Sources of Variation in Lake Bacterial Community Composition. *Applied and Environmental Microbiology*, *71*(1), 227–239. https://doi.org/10.1128/AEM.71.1.227-239.2005

Ylla, I., Peter, H., Romaní, A. M., & Tranvik, L. J. (2013). Different diversity–functioning relationship in lake and stream bacterial communities. *FEMS Microbiology Ecology*, *85*(1), 95–103. https://doi.org/10.1111/1574-6941.12101

Yu, M., Liu, S., Li, G., Zhang, H., Xi, B., Tian, Z., Zhang, Y., & He, X. (2020). Municipal wastewater effluent influences dissolved organic matter quality and microbial community composition in an urbanized stream. *The Science of the Total Environment*, *705*, 135952. https://doi.org/10.1016/j.scitotenv.2019.135952

Zhao, A., Lu, Y., Li, Q., Li, T., & Zhao, J. (2023). Metagenomics reveals the diversity and role of surface-water microbes in biogeochemical cycles in lakes at different terrain ladders. *Frontiers in Environmental Science*, *11*. https://www.frontiersin.org/articles/10.3389/fenvs.2023.1121775

Zwart, G., Crump, B. C., Agterveld, M. P. K., Hagen, F., & Han, S.-K. (2002). Typical freshwater bacteria: An analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquatic Microbial Ecology*, *28*(2), 141–155. https://doi.org/10.3354/ame028141

Chapter 2: A large-scale assessment of bacterial diversity and community structure across environmental gradients in Canadian lakes

2.1 Abstract

Freshwater bacteria are a critical component of lake ecosystems and play essential roles in nutrient cycling that supports aquatic food webs. However, bacterial biodiversity patterns and the factors that shape these patterns are not well described, especially at large landscape scales. In this chapter, we investigated bacterial diversity and distributions across Canadian lakes at the continental scale. We generated a 16S rRNA amplicon dataset encompassing 403 lakes situated across a large and environmentally heterogenous area of the Canadian landscape encompassing 7 distinct ecozones. We identified a broad scale pattern in diversity, where lakes located in the more northwestern ecozones exhibited higher richness than those in the southeastern ecozones. These change in diversity were linked to lake productivity, ion composition, and lake depth. Cyanobacteria, Bacteroidetes, and Firmicutes were enriched in nutrient-rich lakes, while Verrucomicrobia were enriched in nutrient-poor lakes. Variation in bacterial community structure was most strongly related to lake physicochemistry, particularly lake pH and trophic state. Moreover, nutrient rich lakes in the Prairies and Boreal Plains often exhibited the most taxonomically distinct communities. Overall, this chapter is the first to show the major drivers of bacterial diversity and community structure across Canadian lakes and will serve as a future resource in understanding human impacts on the bacterial component of lake ecosystems.

2.2 Introduction

Lakes cover less than 2% of Earth's surface area (Verpoorter et al., 2014; Messager et al., 2016) yet play integral roles in global biogeochemical cycles (Tranvik et al., 2018), and provide valuable ecosystem services (Sterner et al., 2020). Freshwater bacteria are a diverse component of lake ecosystems and are involved in essential ecological processes such as primary productivity, organic matter degradation, and nutrient cycling (Cotner & Biddanda, 2002; Linz et al., 2018); as such, understanding the diversity and environmental drivers of lake bacterial communities is important and an active area of research. Studies have demonstrated the relationship between environmental heterogeneity and spatial variation in microbial diversity patterns (J. B. H. Martiny et al., 2006; Ramette & Tiedje, 2007); for instance, Yannarell & Triplett, (2004) documented the influence of both distance and environmental factors on bacterial community composition across a

broad spatial scale (500 km) in northern and southern Wisconsin lakes. In addition to natural environmental gradients determined by geographical factors such as geology, topography, land cover and use, and climate, the effects of environmental stressors (chemical pollutants and excess nutrients) on bacterial communities have been demonstrated by regional water monitoring across Europe (Saccà et al., 2019; Sperlea et al., 2021); for example, bacterial diversity was shown to increase along a trophic gradient in which eutrophic lakes were observed to harbour higher diversity than meso-eutrophic lakes (Kiersztyn et al., 2019). Furthermore, a previous investigation of Eastern Canadian lakes revealed that local environmental conditions and geographical location altered bacterial diversity, community composition and interactions among bacterial taxa (Kraemer et al., 2020).

In this chapter, I present a continental scale study of lake bacterial communities across Canada with the aim of describing how environmental conditions shape bacterial community diversity and community structure at large spatial scale. The NSERC Canadian Lake Pulse Network (herein referred to as LakePulse) provides an opportunity to investigate lake bacterial communities at such continental scale. As part of LakePulse, hundreds of lakes differing in limnological conditions and watershed characteristics were sampled with the objective of assessing lake health through a multidisciplinary lens that includes studies in biogeochemistry, lake pathogens, pesticide pollution in lakes, as well as remote sensing (Huot et al., 2019). To date, LakePulse has contributed a number of insights into lake microbial ecology, including insights into the biogeography of bacteria in eastern Canada (Kraemer et al., 2020), protist diversity and metabolic patterns across trophic and land use gradients (Garner et al., 2022), and distributions of potential pathogenic bacteria, fungi and protists (Oliva et al., 2022, 2023).

The specific objective of this study is to investigate bacterial community diversity in surface waters of 403 freshwater lakes located across Canadian ecozones using 16S rRNA amplicon analysis. Ecozones are defined by geologic, landform, soil, vegetation, climatic, water and human factors present in an ecologically distinctive area, while characterising that area as a discrete system (CCEA 2016). These ecozones therefore, represent large ecological units of interacting biotic and abiotic factors. Here, we report on the variation in bacterial diversity across ecozones and the environmental drivers of bacterial community composition. Our study maps bacterial biogeography via a standardized assessment of diversity and composition across the country hosting the greatest abundance of lakes globally (Minns et al., 2008).

2.3 Results

2.3.1 Environmental context of lakes

Bacterial diversity was surveyed in the surface waters of 403 Canadian lakes (43 – 57 °N, 53 – 120 °W) located across 7 ecozones (**Figure 2.1**). Lakes ranged in surface area (0.007 – 99.66 km^2) and watershed area (0.2 – 37460 km^2). The lakes represented a broad gradient in trophic state (total phosphorus (TP) 3.3 – 2483.74 µg/L) and pH (pH 5.5 – 10.2) (**Figure 2.1**). Lakes varied in maximum depths (0.9 – 85 m). Lake and watershed characteristics are included as an extended set of environmental parameters summarized by ecozone in **Supplementary Figure S2.1**. Lakes in western Canada (Semi-Arid Plateaux, Prairies, and Boreal Plains ecoregions) notably exhibited the highest mean alkalinity, nutrient, and ion concentrations, and were generally shallower and more productive than lakes in eastern Canada (Mixedwood Plains, Boreal Shield, Atlantic Maritimes, and Atlantic Highlands ecozones) (**Supplementary Table S2.1**).

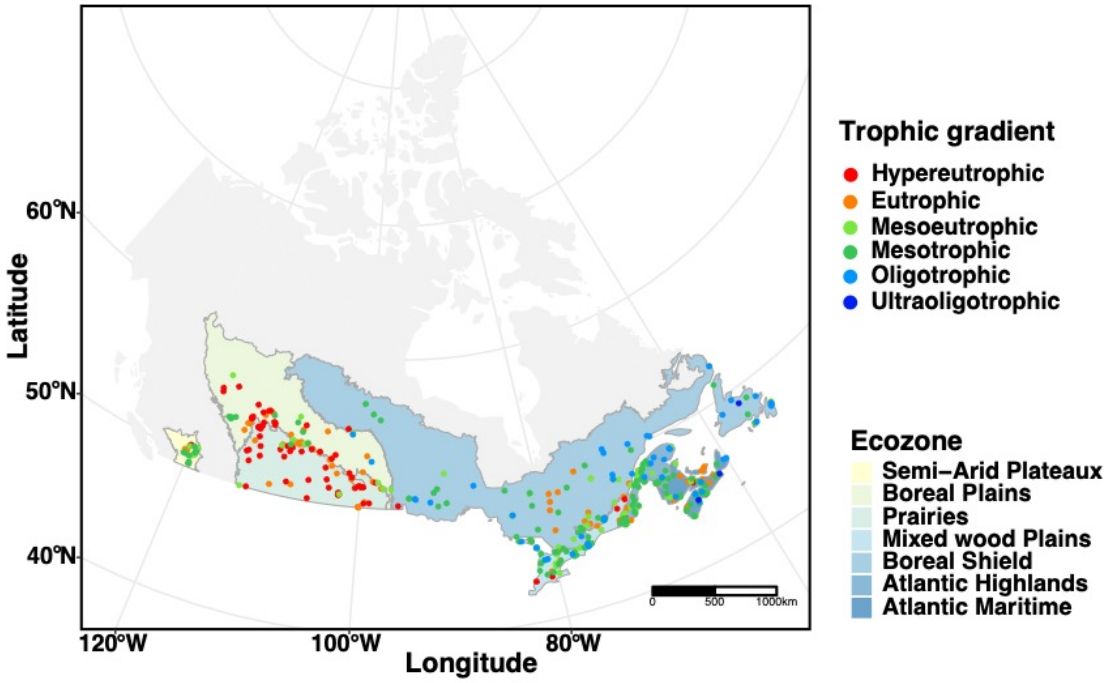


Figure 2.1 *Distribution of lakes across ecozones and trophic gradients. 403 lakes sampled across 7 Canadian ecozones, capturing a wide longitudinal and latitudinal expanse of the country. The points represent all sampled lakes within an ecozone and are color coded by trophic gradient spanning from high nutrient to oligotrophic lakes derived. This gradient was derived from lake phosphorus concentration – hypereutrophic (>100 µg/L; n = 65), eutrophic (35 to 100 µg/L; n = 71), mesoeutrophic (20 to 35 µg/L; n = 68), mesotrophic (10 to 20 µg/L; n = 147), oligotrophic (4 to 10 µg/L; n = 49) and ultraoligotrophic (TP*

concentration, <4 µg/L; n = 3). Light grey regions represent unsampled regions while colored backgrounds indicate sampled ecozones.

2.3.2 Bacterial diversity

Bacterial diversity was assessed by 16S rRNA analysis and 42,926 amplicon sequence variants (ASVs) were identified. An ASV accumulation curve showed that the sampling of new ASVs increased steeply in the first ~200 randomly ordered lakes, indicating that a large-scale sequencing effort was required to exhaustively capture the bacterial diversity targeted by the PCR primer pair (**Supplementary Figure S2.2.1a**). We analyzed ASV incidence to assess the contribution of individual ASVs to total landscape diversity. A large fraction of ASVs were restricted to one or a few lakes (**Supplementary Figure S2.2.1b)**. Others were distributed widely, such as ASVs assigned to Proteobacteria and Actinobacteria which were ubiquitous, yet highly variable in relative abundance across the 403 lakes. We assessed continental scale patterns in bacterial diversity (Shannon diversity index (**Supplementary Figure S2.2.2**), richness (Chao1 richness index) (**Supplementary Figure S2.2.3**), and evenness (Pielou's index) (**Supplementary Figure S2.2.4**). Shannon diversity and Pielou's evenness showed similar patterns being elevated in the eastern ecozones (Boreal Shield, Mixedwood Plains, Atlantic Highland, Atlantic Maritimes) compared to western ecozones (Semi-Arid Plateaux, Prairies, Boreal Plains) (ANOVA; $p = 2.2e^{-05}$) **(Figure 2.2a and 2c**), while richness was elevated in western lakes (ANOVA; $p = 1.88e^{-06}$) (**Figure 2.2b)**, with the exception of bacterial richness measured in lakes located in the Semi-Arid Plateaux.

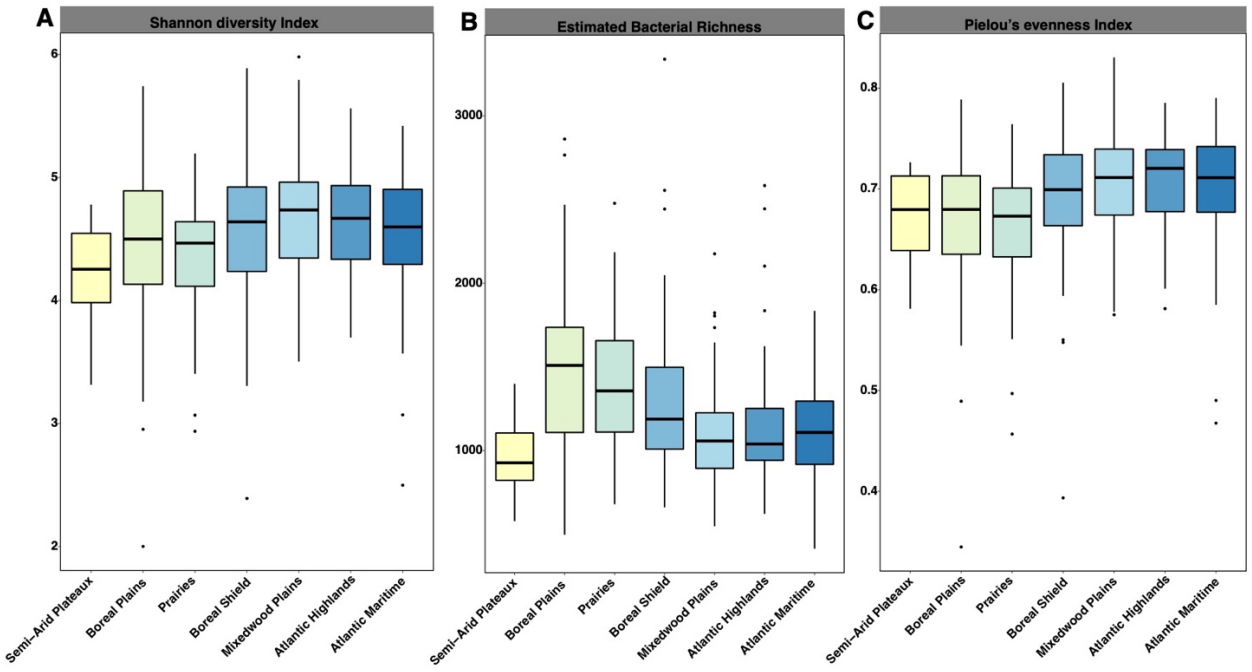


Figure 2.2 *Boxplots of bacterial diversity within lakes across ecozones ordered from western to eastern Canada. (A) Rarefied Shannon diversity index calculated for each bacterial assemblage across ecozones. (B). Rarefied Chao1 diversity index calculated for each bacterial assemblage across ecozones. (C) Pielou's evenness diversity index calculated for each bacterial assemblage across ecozones. Boxplots represents Minimum, Q1 (1st quartile), Median, Q3 (3rd quartile), and Maximum. The whiskers represent the minimum value (lower) and the maximum value (upper) in the data. The bottom edge of the "box" is Q1 and the top edge is Q3.*

2.3.3 Environmental predictors of bacterial diversity

We investigated the environmental variables associated with diversity patterns across lakes using a random forest (RF) modelling approach. Models explained a similar degree of variation for Shannon, Chao1, and Pielou's diversity indices (55-64 % variation explained, out of bag error rate of 17-19%) (**Table 2.1**). The strengths of environmental variables within models differed in their contribution to the three different diversity metrics. To explore variable importance, we categorized variables as either strong (=>5% variation explained) or intermediate (2-5%) predictors of diversity. In support of the observation of diversity differences between western and eastern ecozones, longitude was an intermediate to strong predictor of all diversity metrics (3.2-5.1% variation explained). Partial dependency plots (PDP) showed an increase in richness and decrease in evenness at 80ºW (**Figure 2.3a**). Latitude was a strong predictor of Chao1 richness (8.9%), and an intermediate predictor of

variation in Pielou's evenness (4.7%) (**Figure 2.3b**). We observed a large shift in Chao1 at 50ºN, with higher richness in the north, corresponding to the lakes within the Prairies and Boreal Plains.

Upon observing the difference in diversity among lakes in western and eastern ecozones, we then explored which environmental variables were most important in explaining the difference. A major difference between western and eastern lakes is their inorganic ion concentration and productivity level. Ion-rich and highly productive lakes are common in the western ecozones, and ion-poor and lower productivity lakes are common in the eastern ecozones (**Supplementary Figure S2.1**). Total phosphorus did not explain significant variation in diversity, however chlorophyll-*a* did (**Table 2.1**), and there was a maximum in Shannon diversity at high to intermediate trophic state (eutrophic to mesoeutrophic lakes) (**Figure 2.3c**). In addition, potassium and sulfate concentration were strong predictors of diversity (11.7% and 6.4%) and evenness (17.4% and 6.6%), but not richness (0.7% for both variables). PDPs showed a decline of all diversity metrics with increasing potassium and sulfate concentrations in lakes starting at low levels of potassium (1-10 mg/L) and sulfate (100-500 mg/L) (**Figures 2.3d-e**).

The RF analysis identified several additional variables strongly related to diversity patterns. Maximum lake depth was a strong predictor, particularly for Chao1 richness (21.8%) (**Table 2.1**). Richness was highest in shallow lakes and declined steeply at a lake depth of 20 m (**Figure 2.3f**). In addition, surface water temperature was a strong predictor of variation in bacterial richness (7.8%). PDPs showed that bacterial richness was lower at temperatures greater than 20ºC (**Figure 2.3g**). Dissolved organic carbon (DOC) concentrations was a strong predictor of Chao1 Richness (6.5%). Chao1 was positively associated with DOC concentration and a pronounced increase in diversity occurred over the DOC concentration range of 1-10 mg/L (**Figure 2.3h**).

Table 2.1 Summary of random forest result on bacterial diversity indices. The most influential variables are highlighted as well as the overall RF model fit and out of bag error.

Predictor	Shannon diversity	Chao1 Richness	Pielou's evenness
Rsquared_fitted	**0.63**	**0.63**	**0.55**
Rsquared_OOB	**0.19**	**0.26**	**0.17**
Weather			
Ice_disappearance_julianday	0.76	**6.99**	1.90
Precipitation_total_7d	0.81	0.31	0.28
Solar_radiation_net_7d	0.72	1.59	0.00
Temperature_mean_7d	0.25	0.30	0.43
Windspeed_mean_7d	0.32	1.50	0.29
Geographic Variables			
Altitude	1.12	2.31	0.88
Latitude	1.71	**8.98**	4.65
Longitude	3.24	**5.08**	3.69
Lake/Physical morphometry			
Area	1.18	0.14	1.26
Circularity	0.38	0.00	0.48
Discharge	0.51	0.07	0.25
Lakewatershed_area_ratio	0.82	0.69	0.35
Lake depth	**6.42**	**21.75**	1.64
Residence time	0.50	0.22	0.80
Slope_100m	1.79	**6.79**	0.20
Volume	0.31	0.27	0.23
Watershed area	0.73	0.20	0.33

Predictor	Shannon diversity	Chao1 Richness	Pielou's evenness
Physicochemical properties			
Calcium	3.68	0.46	5.75
Chlorophyll-*a*	3.69	0.51	2.79
Chloride	1.58	0.20	2.30
Colour	1.37	1.26	1.21
DIC	3.30	0.56	**5.37**
DOC	0.91	**6.46**	0.88
Magnesium	4.27	0.64	**5.58**
pH epilimnion	3.03	0.66	**5.02**
Potassium	**11.65**	0.73	**17.43**
Surface temperature	0.51	**7.80**	0.02
Sodium	2.82	0.15	3.28
TN	0.89	2.47	0.82
TP	0.95	2.18	1.01
Sulfate	**6.38**	0.67	**6.57**
Watershed surface soil properties			
cec_mean_0_5	0.19	0.32	0.23
cfvo_mean_0_5	2.85	1.01	1.02
clay_mean_0_5	0.99	0.11	2.19
nitrogen_mean_0_5	3.12	0.54	2.62
ocd_mean_0_5	1.54	1.88	0.06
phh2o_mean_0_5	0.63	2.78	1.01
sand_mean_0_5	**6.71**	1.48	4.55
silt_mean_0_5	4.31	**6.61**	1.46
soc_mean_0_5	2.00	0.10	2.02
bdod_mean_0.5	0.44	0.64	0.55

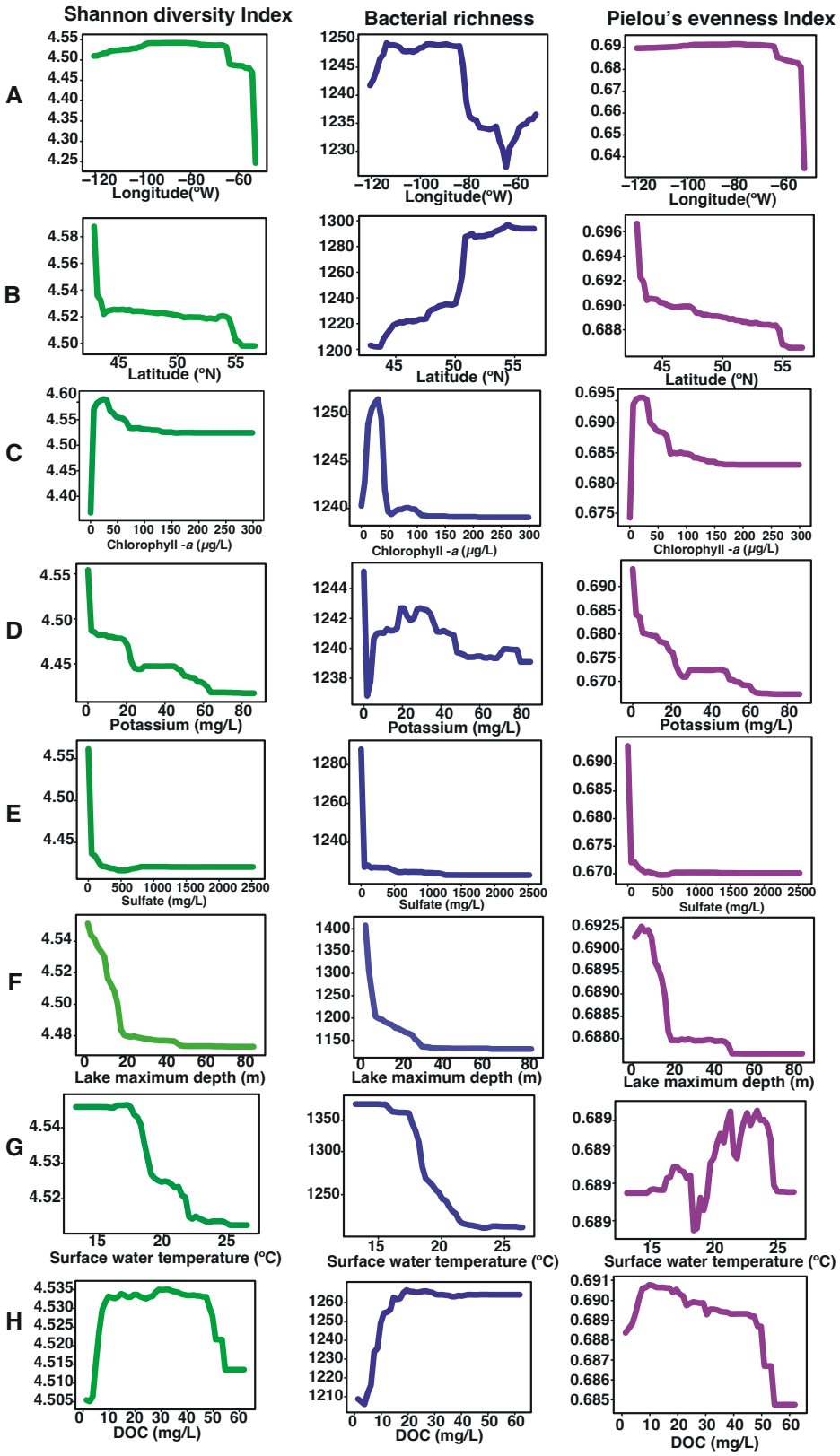

Shannon diversity Index
Bacterial richness
Pielou's evenness Index
A
B
C
D
E
F
G
H
Longitude(°W)
Latitude (°N)
Chlorophyll -a (μg/L)
Potassium (mg/L)
Sulfate (mg/L)
Lake maximum depth (m)
Surface water temperature (°C)
DOC (mg/L)

Figure 2.3 *Effect of geographic and environmental variables on bacterial diversity, richness, and evenness at the continental scale. Each horizontal panel depicts the relationship between the three diversity indices (response variables) and a specific geographic or environmental variable. Partial dependence plots (PDP), based on results from random forest analysis, reveal the mean marginal influence of explanatory variables on bacterial diversity indices. Each alphabetically labelled horizontal panel represents the effect of one geographic or environmental variable on response variables. A-B) are PDPs showing the direction of relationship between Shannon diversity, richness, and Pielou's evenness Index and geography (longitude and latitude). C) shows bacterial diversity relationships with chlorophyll-a concentrations. D-E) show the direction of the relationships between bacterial diversity indices and ionic concentrations of potassium and sulfate. F-G) show the direction of relationship between bacterial diversity lake depth and surface water temperature and H) shows relationships with DOC concentrations.*

2.3.4 Bacterial community composition

An overview of bacterial community composition in lakes according to trophic state is presented in **Figure 2.4.** Actinobacteriota, Proteobacteria and Bacteroidota dominated bacterial communities. At the phylum level, observable patterns of taxonomic composition along the trophic gradient included an increase in abundance of Cyanobacteria and a similar increase of the Bacteroidota phylum. Firmicutes were most abundant in hypereutrophic lakes (**Figure 2.4**). Conversely, Actinobacteriota, Verrucomicrobiota and Proteobacteria phyla were most abundant in oligotrophic to ultraoligotrophic lakes.

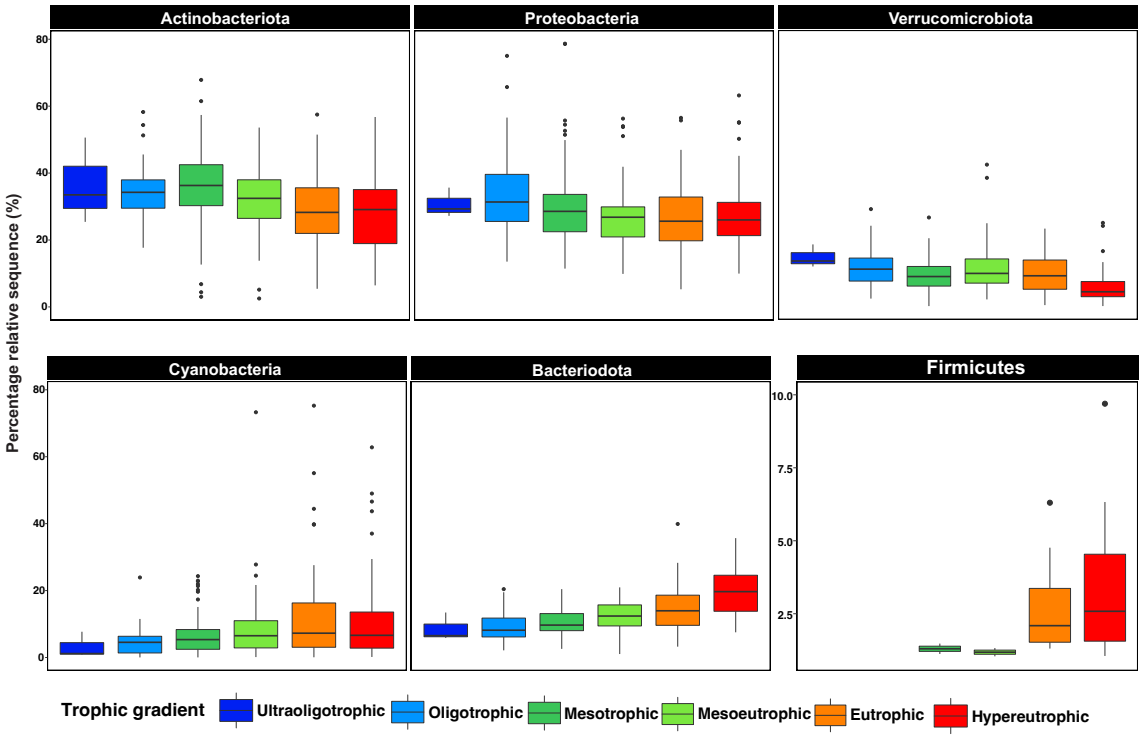


Figure 2.4 *Taxonomic composition of lakes across trophic gradient. Trophic state are as follows: ultraoligotrophic (TP concentration, <4 µg/L), oligotrophic (4 to 10 µg/L), mesotrophic (10 to 20 µg/L), mesoeutrophic (20 to 35 µg/L), eutrophic (35 to 100 µg/L), and hypereutrophic (>100 µg/L).*

We explored the variation in bacterial community structure across lakes using non-metric multi-dimensional scaling (NMDS) analysis (**Figure 2.5a**). The NMDS ordination showed a distribution of lakes along axis 1 that was related to lake geographic location and trophic state. Bacterial community distribution in lakes along NMDS axes 1 were correlated with latitude (r_1 = 0.44) and longitude (r_1 = -0.66). Distributions of bacterial communities in lakes along NMDS axes 1 were correlated with trophic gradient represented by total phosphorus (TP) concentration (r_1 = 0.46), and to a lesser extent to Chl-*a* concentrations (r_1 = 0.27) (**Figure 2.5b**).

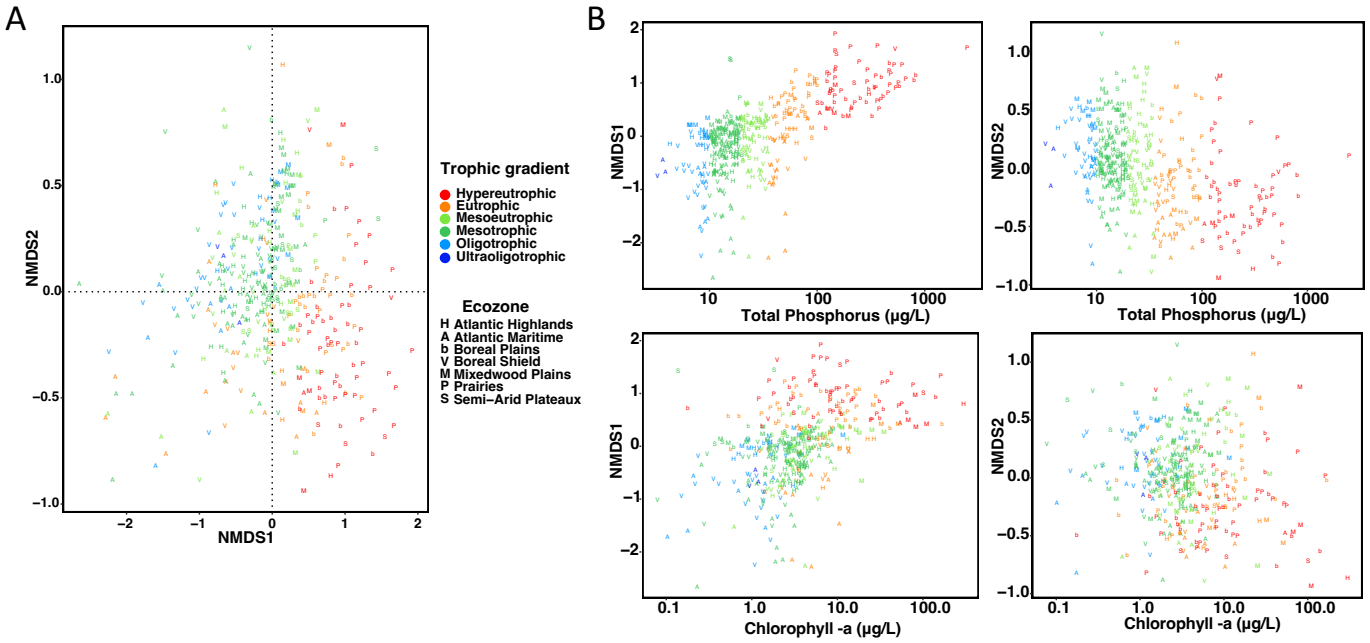


Figure 2.5 *Taxonomic composition and distribution of bacteria across 403 Canadian lakes. (A). NMDS ordination of the taxonomic variation among bacterial communities across 403 lakes (stress = 0.14). (B). Correlations between sample position along the first two ordination axes and either total phosphorus or chlorophyll –a.*

In the NMDS ordination, we observed a dispersion of hypereutrophic and eutrophic lakes, suggesting that community composition varies across these high nutrient state lakes. To examine the taxonomic distinctiveness of individual communities, we quantified local contribution to β-diversity (LCBD) (**Supplementary Figure S2.3**). Ninety-two lakes in our dataset had significantly distinct taxonomic composition. Such unique communities were found in the Prairies (27 lakes), Atlantic Maritimes (20 lakes), Boreal Shield (15 lakes) and Boreal Plains (13 lakes), with a few

others scattered across other ecozones (**Supplementary Figure S2.3**). The variation in species composition among lakes was decomposed into replacement and richness. The result showed that taxonomic dissimilarities between communities were primarily generated through ASV replacement (70.1% of total variance) but also, to a lesser extent, differences in ASV richness (29.9%). Positive correlations were detected between LCBD and a few physicochemical variables, including lake colour (r = 0.38), potassium (r = 0.35), TN (r = 0.32), DOC (r = 0.30), TP (r = 0.29), sodium (r = 0.25), magnesium (r = 0.24), DIC (r = 0.23) and chlorophyll–*a* (r = 0.20). Furthermore, LCBD values showed positive correlations with watershed land use type, agriculture (r = 0.25) but were negatively correlated with lake morphometry, lake depth (r = -0.29).

2.3.5 Environmental predictors of bacterial community composition

To elucidate the drivers of community composition, we investigated the importance of lake physicochemistry, watershed soil properties, morphometry, geography, and climate conditions using generalized dissimilarity models (GDMs). The GDM inferred from lake physicochemistry explained the largest amount of taxonomic turnover in lake bacterial communities (D^2_{taxon} = 48%) (**Figure 2.6a).** Lake pH was the most important variable in the physicochemical GDM, and continuous taxonomic turnover was observed along the full pH gradient of lakes (pH 5.54-10.17) (**Table 2.2; Figure 2.6b**); Chlorophyll-*a*, lake colour, DIC, and potassium concentrations were additional significant predictors (**Table 2.2; Figure 2.6 c-f**); total nitrogen and temperature were also significant, but weaker predictors of community turnover (**Table 2.2; Figure 2.6 g-h**). In addition, GDMs pointed to an influence of lake morphometric characteristics on bacterial communities. Lake morphometry and physical features, comprising maximum depth, and water residence time, also had relatively strong effects on turnover (D^2_{taxon} = 12.7%) (**Figure 2.6 i-j**). Geographic distance and longitude were important in the geography model. Turnover in bacterial community composition was observed along 80°W (**Figure 2.6 k-l**).

Table 2.2 Percent deviance explained by generalized dissimilarity models (GDMs) fitting the responses of bacterial communities to different categories of environmental gradients. NS, model was not statistically significant ($P \geq 0.05$).

Response variable	Deviance explained (%)	Explanatory variable	Predictors
Taxonomy	48	Physicochemistry	pH (1.51), Colour (0.94), K+ (0.93), Chl a (0.93), DIC (0.90), TN (0.37), Surface temp (0.35)
Taxonomy	12.7	Morphometry	Max depth (0.68), Residence time (0.45)
Taxonomy	7.6	Geography	Longitude (0.46), Geographic distance (0.30)
Taxonomy	NS	Climate	N/A
Taxonomy	NS	Soil	N/A

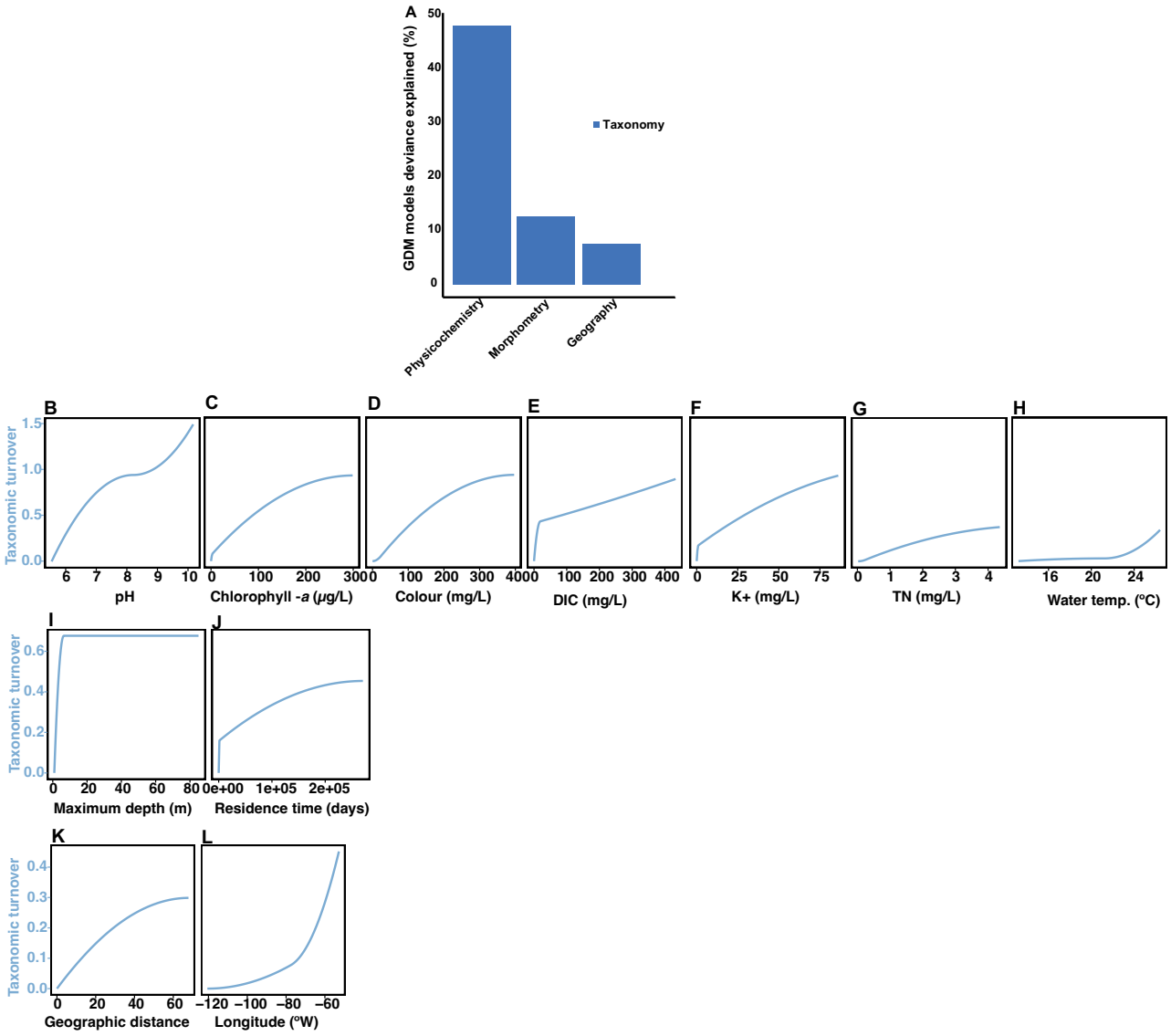


Figure 2.6 *Bacterial community turn over across geography and environmental conditions. Generalized dissimilarity models (GDMs) show taxonomic turnover across environmental gradients. A) Summary of percent deviance explained by models shows that taxonomic composition is most responsive to lake physicochemistry. B-H) Taxonomic turnover across physicochemical gradients. I-J) Taxonomic turnover across lake morphometry and K-L) Taxonomic turnover across geography.*

2.4 Discussion

Our continental-scale survey of bacterial diversity presented here encapsulated a broad heterogeneity of freshwater lakes distributed across the most lake-rich region of Earth (Messager et al., 2016). We elucidated the predictors of bacterial diversity and community composition patterns across a collection of strategically selected 403 lakes (Huot et al., 2019). Overall, we detected a spatial structure of bacterial diversity (richness and evenness) related primarily by geography and lake physicochemistry. Changes in community composition were also largely related to lake physicochemistry. However, the large amount of variation left unresolved in our dataset may require further investigations.

2.4.1 A latitudinal gradient in bacterial diversity

One of the most commonly observed large scale spatial patterns in ecology is a latitudinal diversity gradient, with more species occurring towards the equator (B. Liu et al., 2022). Although higher diversity at lower latitude is common in macroorganisms, studies in microorganisms often report an opposite relationship. For example, global studies of marine and soil bacterial diversity have shown positive relationships between diversity and latitude wherein marine diversity peaked in temperate latitudes (Ladau et al., 2013) and soil diversity was higher in temperate forest soils than in tropical or subtropical soils (Fu et al., 2023; Tian et al., 2018). Interestingly, contrasting reports have also been made in the global oceans wherein a decline in ocean prokaryotic diversity towards the poles, driven by declining temperature, was observed (Ibarbalz et al., 2019). Compared to oceans and soils, knowledge about bacterial latitudinal diversity patterns in lakes remains fragmentary.

Although our study focused on bacterial diversity patterns across a relatively narrow latitudinal gradient (40-60°N), we observed a positive relationship between diversity and latitude, with a pronounced increase in bacterial richness at approximately 50°N. Lakes above 50°N were colder and shallower western lakes (Prairies and Boreal Plains), suggesting temperature and lake depth are variables contributing to the observed latitudinal pattern. Indeed, we found a moderate negative correlation between latitude and temperature (r = -0.35) and a weaker negative correlation with depth (-0.11) revealing a relationship in which bacterial richness declined with increasing

surface water temperature and maximum depth. In aquatic ecosystems, temperature is known to influence the rate of chemical and biological reactions (Sun et al., 2022). In our lakes, bacterial richness was highest between 15-18ºC surface water temperature with a decline in richness at approximately 20ºC that continued steadily at higher temperatures. Similarly, bacterial richness declined with increasing maximum depth, but remained stable beyond depths of 20 m. The high bacterial richness in western lakes may also result from physical mixing in polymictic lakes, with approximately 76% of lakes in Prairies and 53% of lakes in the Boreal Plains having no thermal stratification, in comparison with eastern lakes in which mixed lakes were less than 50% in each ecozone. Polymixis is present in some western prairies and plains lakes owing to their shallow depths and this mixing could enhance the resuspension of bacteria from the sediments and littoral zones into the water column. Also, shallow lakes may have shorter resident times, thereby also allowing for the infiltration of soil bacteria (Adams et al., 2010). On investigating 15 shallow lakes with different trophic states in Hubei Province (China), Wang et al., 2022 demonstrated that Chao1 index in shallow, slightly eutrophic lakes were generally elevated (Y. Wang et al., 2022).

2.4.2 A productivity gradient in bacterial diversity

In addition to latitude, productivity-diversity relationships have received considerable research attention in ecology (Smith, 2007). However, understanding the relative roles of ecological processes, such as environmental filtering, versus competitive exclusion that could bring about observable changes in bacterial diversity along productivity gradients remains elusive. The Lakepulse dataset captured lake productivity along a gradient from ultraoligotrophic to hypereutrophic. This allowed us to investigate diversity patterns across the full spectrum of lake productivity. For this purpose, we used chlorophyll-*a* concentrations as a productivity proxy. We observed a hump-shaped relationship, wherein diversity showed a clear maximum at intermediate productivity levels. Bacterial diversity (i.e., Shannon diversity and species evenness) peaked between 10 to 45 μg/L of chlorophyll-a concentrations and decreased at concentrations greater than 50 μg/L. Bacterial richness was weakly predicted by chlorophyll-*a* concentrations but declined along the gradient. While we sampled a wide range of chlorophyll-*a* gradient from 1 to 300 μg/L, it is noteworthy that diversity changes were only observed at the lower ends of productivity and no change in diversity was observed at the higher end productivity. Jankowski and colleagues reported a positive linear relationship between bacterial

richness and productivity in oligotrophic to eutrophic lakes where epilimnetic chlorophyll-*a* concentrations ranged from 0.23 to 10.2 µg/L (Jankowski et al., 2014). However, this study, like some others (Korhonen et al., 2011) was a regional study within the Puget Sound region of western Washington (USA) and southern British Columbia with a narrower range of productivity compared to that captured in our study.

In addition to diversity relationships with productivity, we detected a pattern of increasing bacterial richness as levels of DOC increased. Over the past two decades, rising DOC levels in freshwater ecosystems related to climate change has been reported across Europe and North America (Evans et al., 2006). Bacterial diversity is linked with DOC degradation not just in lakes (Lambert & Perga, 2019) but other bacterial habitats such as the ocean (Chen et al., 2020) and soil (J. Wang et al., 2021). Increasing bacterial richness in lakes could be a biological mechanism useful for reducing nutrient concentrations in DOC-laden freshwaters owing to greater variation in species foraging physiology (Saleem et al., 2016).

2.4.3 Phylum-level variation in bacterial communities

Our results regarding bacterial community composition at a continental scale was generally consistent with previous reports of bacterial phyla found in other North American lakes (Kraemer et al., 2020; Linz et al., 2017; Morrison et al., 2017; Mou et al., 2013; Paver et al., 2020; Sadeghi et al., 2021; Shahraki et al., 2021). Our lakes were replete with major freshwater lake taxa, including Actinobacteriota and Proteobacteria (Liu et al., 2021; Mateus-Barros et al., 2021; Newton et al., 2011). We observed a composition shift from communities enriched in Actinobacteriota (a group of organisms sensitive to nutrient overloading) and Proteobacteria (adapted to some level of nutrient overloading) to those enriched by Cyanobacteria (phototrophic nuisance implicated in algal blooms and capable of producing toxins), Bacteroidota (proficient in the degradation of complex biopolymers and dissolved organic matter) (Newton et al., 2011) and Firmicutes (possess diverse metabolic capacities) (Martiny et al., 2006) over the trophic gradient. Obieze and colleagues (Obieze et al., 2022) reported the high abundance of Bacteriodota, Cyanobacteria and others in Lake Winnipeg which is characterized by high concentrations of nitrogen and phosphorus.

The observed shifts in relative bacterial phylum abundances indicated that nutrient loadings increased within lakes located in the western ecozones of the Prairies, Boreal Plains and Semi-Arid Plateaux may be attributed to runoffs from agricultural practices within the surrounding watershed surface soils. The relationship between agriculture and soil bacterial community composition has been reported with distinct bacterial compositions in agricultural and non-agricultural soils, even in proximity (Lauber et al., 2013). This may point to the uniqueness of soil composition in agricultural soil due to the use of fertilizers, most of which can run off into lake water columns, thereby introducing specific bacterial groups where watershed land use includes agriculture. The Prairies ecozone, commonly referred to as the "food basket" of the nation is an agricultural landscape where more than 85% of croplands and pasture in Canada is located (Rchaid et al., 2022). Therefore, such run-off mechanisms may be influencing the taxonomic distinctness of bacterial communities.

Unexpectedly, given that they are not typically components of freshwater pelagic communities, Firmicutes were abundant in hyper-eutrophic agricultural lakes. Abundant Firmicutes have been detected in soils in long-term agricultural field experiments in Belgium (Liu et al., 2022). This discovery of Firmicutes in high abundance in high-nutrient agricultural lakes in our study can be attributed to the diverse metabolic capacity of this group and because species within this group are known to be motile, tolerant to extreme environmental conditions, metabolically diverse, and possess specialized carbohydrate decomposition machinery that makes them highly competitive in nutrient-rich environments (Newton et al., 2011).

At the other end of the gradient, ultra-oligotrophic to oligotrophic lakes in our study were enriched with Verrucomicrobiota, a metabolically diverse group that has been reported to be active during winter in ice-covered lakes (Tran et al., 2018). The shift in bacterial community composition along trophic gradients that was observed may be better explained by focused investigations into the processes that influence their assembly patterns. Therefore, specific mechanisms for the observed bacterial community composition shifts requires further investigation.

2.4.4 Taxonomic distinctness across lakes

Intriguingly, we found that bacterial communities in eutrophic to hypereutrophic lakes located within agricultural watershed were the most taxonomically distinct (i.e., significant LCBD values, n=92) across Canada. Of these high LCBD lakes, 29 lakes (31.5% of lakes with significant LCBD values) had high Firmicute abundance, ranging from 943 (a Semi-Arid Plateaux Lake) to 41,620 (a Prairie Lake) ASVs. These lakes with high Firmicute abundances were all located within the western ecozones of the Prairies, Boreal Plains and Semi-Arid Plateaux. This supports our previous observations of distinct bacterial communities being associated with nutrient-rich lakes. We detected that shifts in bacterial community composition in our lakes were predominantly a result of ASV turnover along environmental gradients, with less compositional shifts attributable to a greater number of bacterial ASVs in some lakes relative to others. For instance, the abundance of Firmicutes as previously reported in these western lakes displays the observed ASV turnover within lakes. This explains why some groups of ASVs became dominant along the trophic gradient, phasing out other groups that are less tolerant to prevailing lake conditions.

2.4.5 Lake pH and other physicochemical conditions drive bacterial community structure

It is generally known that bacterial community structure in lakes is primarily driven by environmental gradients (Aguilar & Sommaruga, 2020). In this study, we found that the influence of local environmental factors (i.e., physicochemistry) outweighed the effects of geographic and lake morphometric/physical features across the continental scale. Water chemistry (but also lake morphometric characteristics) were similarly reported to be significant drivers of bacterial community composition in a comparative analysis of bacterioplankton assemblages from six sites located along coastal regions of Lakes Michigan, Huron, and Erie (Olapade, 2018). In a study of eighteen southern Canada lakes confined to a single region, comparing the influences of spatial position and water chemistry on a range of lake communities within various trophic levels – from bacteria to fish, only local environmental physicochemical predictors explained bacterial community composition, while higher trophic groups were influenced by spatial predictors (Beisner et al. 2006). In this study, we found an inter-regional distinction in the taxonomic distributions of bacterial communities associated to nutrient concentrations within lakes. Clearly, our high throughput method revealed that lakes were aggregated based on their trophic states (i.e.,

a clear clustering of bacterial communities within the Prairies and Boreal Shield ecozones characterized by eutrophic to hypereutrophic lakes as opposed to lakes in eastern ecozones).

The strongest drivers of bacterial community structure in our study lakes were pH, chlorophyll-*a*, lake colour, DIC, TN, potassium, and surface water temperature. Our results suggest that lake pH drives turnover in bacterial community composition. Studies on bacterial communities in shallow lakes have mainly focused on sediments at the bottom of the water column (Pinnell & Turner, 2020; Z. Yang et al., 2021). While similar observations have been made in lake sediments such as a link between the microbial community structure of Lake Hazen and pH, in the Canadian Nunavat region (Ruuskanen et al., 2018), relationships between pH and water column bacterial composition turnover have rarely been reported. Uncovering this trend in Canadian lakes may be attributed to the wide pH range sampled across the continental scale. This may have made the uncovering of such links possible in the water column. Ruiz-Gonzales and colleagues (2015) similarly demonstrated that aquatic bacterial community composition was linked to physicochemical drivers, primarily pH, water temperature and water residence time, in 296 boreal rivers and lakes across five regions in northern Quebec (Canada).

Furthermore, we discovered that chlorophyll-*a* concentrations and lake colour explained a similar amount of turnover on lake bacterial community structure. Research has demonstrated that lake nutrients like TP and TN, water colour and chlorophyll-a concentrations exhibit a nested relationship, described as the nutrient-water color paradigm (Webster et al., 2008). This framework has been used to characterize lake trophic conditions by relating lake primary productivity to both nutrients and water colour associated with coloured DOC (Nürnberg & Shaw, 1998). This paradigm has shown that TP, a limiting nutrient, and water colour, a strong light attenuator, influence lake chlorophyll-*a* concentrations (Fergus et al., 2016). However, it is noteworthy that these relationships could be highly variable depending on specific lake and catchment geomorphology (Fergus et al., 2016). In mesotrophic Lake Diefenbaker located in the Canadian Prairies, turnover in phytoplankton biomass was associated with chlorophyll-*a* concentrations and lake turbidity; although no relationship with bacterioplankton was investigated (Abirhire et al., 2015, 2023) as we have done in this study, thereby contributing new knowledge of Canadian lake bacterioplankton communities.

2.5 Conclusion

This study reports on the first pan-Canadian examination of lake bacterial diversity and community structure spanning continental-wide environmental and geographical gradients. We revealed a latitudinal temperature gradient along the continental scale correlated with lake maximum depth. This pattern in bacterial richness was attributed to a broad range of environmental gradients in lakes across the landscape. Water chemistry had the greatest impact on lake bacterial communities. Our large-scale survey provides new information on prevailing shifts in bacterial diversity and community composition resulting from long environmental gradients. Our dataset represents an important new microbial diversity resource from hundreds of lakes in Canada. To expand these investigations, the next chapter will incorporate land use types across Canada into a larger dataset consisting of over 600 lakes with the goal of identifying the more specific role of land use type and water quality in shaping lake bacterial diversity and community composition.

2.6 Caveats

Our analysis explored the influence of geographic and environmental factors on bacterial diversity and community composition. This limited our ability to investigate land use, biotic interactions, and their influence on diversity and community composition. Future work is aimed at examining bacterial co-occurrence patterns, resolving bacterial indicators of lake health as well as investigating the functional potential of lake communities. As an integral component of lake communities, bacteria are a valuable tool for detecting the impacts of anthropogenic stressors, including climate change, on biodiversity and ecosystem services. Their rapid proliferation and high metabolic rates (Sagova-Mareckova et al., 2021) make them responsive to environmental shifts such as changes in the physical and chemical characteristics of freshwaters, including the introduction of pollutants (Pernthaler, 2017). Therefore, more frequent sampling would need to be done to better resolve bacteria community responses to environmental gradients and change at the continental scale.

2.7 Methods

2.7.1 Lake selection and sampling

Surface water was collected from 664 lakes over a three-year period (2017-2019) by the Natural Sciences and Engineering Research Council of Canada (NSERC) Canadian Lake Pulse Network field campaigns (Huot et al., 2019). To minimize seasonal variability, sampling was done in each set of lakes (approximately one third of the lakes were sampled per year) at the period of peak thermal stratification (July to early September in each year) across 12 terrestrial ecozones – regions characterised by landform, geology, and vegetation (Wiken et al., 1996). Lakes were selected based on a random sampling design that was stratified across sizes (0.1-1km, 1-10km, 10-100km) and watershed human impact index (0-1 HII) to reflect both natural and human-induced alterations of lake systems. Furthermore, only lakes within 1km from a road, having at least 1m maximum depth were considered. Experimentally acidified and nutrient-enriched lakes in the Experimental Lakes Area were not included in our analyses. Saline lakes, identified as having conductivity ≥ 8,000 μS/cm or total major ions ≥4,000 mg/L, were likewise removed. For our study of bacterial communities, a total of 403 lakes were analysed.

All sampling equipment were acid-washed and triple rinsed with lake water before use. Sampling in each lake was done at the site of maximum depth located by depth sounding with the aid of available bathymetric maps. Water for assessing bacterial communities was collected from the euphotic zone (estimated as twice the Secchi disk depth) over a depth of up to 2 m below the surface using an integrated tube sampler. Carboys were stored in ice-pack-chilled coolers until water could be filtered on the lake-shore later in the day. Water was prefiltered through 100 μm synthetic nylon mesh and vacuum-filtered on 47 mm-diameter 0.22 μm Durapore membranes through a glass funnel apparatus at a maximum pressure of 8 inHg. Filtration concluded either at 500 mL or upon clogging of the filter. Filters were stored in sterile cryovials at -80 °C. Details for environmental sampling and field protocols can be found in the NSERC Canadian Lake Pulse Network field manual 2017 - 2018 - 2019 surveys prepared by Varin and colleagues (NSERC Canadian Lake Pulse Network, 2021).

2.7.2 DNA extraction, amplification and sequencing of the bacterial 16S rRNA gene

Bacterial diversity was assessed through the sequencing of 16S rRNA gene fragments amplified from DNA collected in 0.22 – 100 µm surface water samples., DNA was extracted from filters with PowerWater kits (Mobio Technologies Inc., Vancouver, Canada) using the manufacturer protocol including the optional Step 7 described in the manufacturer's detailed protocol (*i.e.*, addition of 1 µL ribonuclease A followed by incubation at 37 °C for 30 min). This was eluted into 50 µl of buffer.

A ~300 bp fragment of the 16S rRNA gene V4 region was amplified with the primer set 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3') (Caporaso et al., 2011) under the following conditions: 5 µl Phusion High Fidelity Buffer, 0.5 µl dNTPs (10 mM), 1.8 µl of each primer (5 µM), 0.25 µl Phusion polymerase, 13.65 µl ddH_2O and 2 µl of DNA. PCR conditions were 30 seconds of 98˚C, followed by 22 cycles of 98˚C for 20 seconds, 54˚C for 35 seconds, 72˚C for 30 seconds, and a final elongation at 72˚C for one minute. All pre-PCR DNA dilutions and liquid transfers were performed under positive pressure in a UV cabinet. PCR products were loaded into 1.5% agarose gel and electrophoresed at 80 V for 60 min. Samples were sequenced on an Illumina MiSeq machine as three separate sequencing runs.

Additional details of PCR step 2 and purification are as follows: Products from four reactions per sample were pooled and cleaned with the Zymo research DNA purification kit (Zymo Research, Irvine, USA) according to the standard protocol and eluted into a volume of 30 µl. Subsequently, barcodes and Illumina adaptors were added in a second PCR reaction (5 µl High fidelity Phusion buffer, 0.5 µl dNTPs (10 mM), 1.8 µl primer each of PE-PCR-III-F (5 µM) and PE-PCR-III-XXX (5 µM), 0.25 µl Phusion polymerase, 11.65 µl ddH_2O and 4 µl cleaned PCR product).

PCR conditions were 30 seconds of 98˚C, seven cycles of 30 seconds of 98˚C, 30 seconds of 83˚C and 30 seconds of 72˚C, followed by cooling to 10˚C. After the second PCR, the products were purified using the AMPure kit (Beckman Coulter Diagnostics, Montreal, Canada), following the standard protocol (except for using 0.8X AMPure XP beads instead of 1.8X). DNA concentrations were measured using a nanodrop and reactions pooled in volumes containing equal

quantities of DNA. Pooled samples were diluted to 10 nM and sequenced using an Illumina MiSeq machine (three runs total). Each sequencing plate contained two negative controls (ddH_2O) and one DNA sample of a mock community for sequencing quality control.

2.7.3 Processing of sequence data

Primer sequences were removed in Cutadapt v. 3.1 (Martin, 2011). Trimmed reads were processed into ASVs in R through DADA2 v. 1.16 (Callahan et al., 2016). The DADA2 pipeline consisted of trimming low-quality end positions, inferring denoised ASVs based on learned sequencing error rates, merging paired forward and reverse reads, eliminating chimaeras, and assigning taxonomy. Samples were pooled for ASV inference using otherwise default parameters. Taxonomy was assigned in TaxAss which classified 16S rRNA gene sequences using both the curated freshwater FreshTrain v. 2020/06/15 (specific for freshwater bacteria) and SILVA v. 138 reference databases (Rohwer et al., 2018). ASVs were aligned in SINA v. 1.7.2 (Pruesse et al., 2012) against the SILVA 138.1 SSU Ref NR 99 rRNA gene database (released August 27, 2020) (Quast et al., 2013). Positions outside a defined range were trimmed off, and ASVs with fragment lengths under 250 bp or over 260 bp were removed in R. Samples representing negative controls (sequencing blanks) and mock communities were removed. To create a dataset of bacterial assemblages, ASVs not assigned at the kingdom rank and ASVs assigned to archaea, eukaryotes, and chloroplasts were removed. Finally, samples containing fewer than 10000 sequences were removed; saline, northern lakes (belonging to the Pacific Maritimes, Taiga Plains, Boreal, Taiga and Montane Cordillera) and experimental lakes were also excluded from this analysis, resulting in a final dataset of 403 freshwater lake ASV assemblages.

2.7.4 Rarefaction and accumulation curves

Rarefaction analysis on the total data set was done by measuring ASV richness in assemblages randomly subsampled at each 1,000-sequence step. Taxon accumulation was estimated in a random ordering of lakes using 100 permutations in the R package vegan (Oksanen et al., 2019).

2.7.5 Processing of spatial and environmental data

Meteorological conditions recorded over seven days leading up to sampling were accessed from European ReAnalysis (ERA5)-Land hourly reanalysis (Muñoz Sabater, 2019). Data on watershed slope and lake volume, discharge, and hydraulic residence time were accessed from HydroLAKES v. 1.0 (Messager et al., 2016). Watershed surface soil properties were accessed from SoilGrids resolved at 250 m (Hengl et al., 2017) and land cover information was compiled as described in (Huot et al., 2019).

Specific to GDM analysis, environmental data were categorized into broad groups of variables for environmental filtering analyses. Latitude and longitude were categorized as geography variables. Ice disappearance day and meteorological variables (air temperature, precipitations, wind speed and solar radiation) were categorized as weather variables. Lake surface area, circularity, watershed slope within 100 m of the shoreline, volume, maximum depth, discharge, residence time, watershed area and the ratio of lake to watershed area were categorised as lake physical and morphometric variables. Surface soil properties (mean bulk density, cation exchange capacity, total nitrogen, pH, organic carbon, coarse fragments, clay, sand, and silt) were categorized as watershed variables. Major ion (calcium, magnesium, potassium, sodium, chloride, and sulfate), nutrient (total phosphorus, total nitrogen, soluble reactive phosphorus, DIC, and DOC), and concentrations, pH, lake colour and surface water temperature were categorized as physicochemical variables. Missing water chemistry and lake physical variable data were replaced with ecozone median values. Maps were constructed in R with the NAD 83 coordinate reference system and using the coordinates of Canada from the package maps and ecozone shape files sourced from the Canada Council of Ecological Areas (Wiken et al., 1996).

2.7.6 Estimation of bacterial diversity

ASV composition was randomly subsampled (i.e., "rarefied") to an equal sampling depth of 12,138 sequences specifically for the estimation of α-diversity indices. ASV richness (represented by the Chao1 richness index), Pielou's evenness and Shannon diversity indices were computed in the *R* package *microbiome* (Lahti & Shetty et al., 2017).

2.7.7 Random (RF) analyses and determination of relationship using partial dependence plots

RF analysis was used to analyze the influence of preselected groups of variables on bacterial diversity and richness represented by three main response variables (Shannon diversity, Richness, and Phylogenetic diversity). RF is advantageous in comparison to traditional regression techniques because it is considerably less vulnerable to overfitting when processing a large number of predictor variables as is the case in our study (Matsuki et al., 2016; Ryo & Rillig, 2017).

Here, we used a RF technique based on conditional inference regression trees (Strobl et al., 2009) developed by Ryo and Rillig (2017). A measure of importance was calculated for each predictor variable by cross-validating each tree with data not used when the tree was constructed, referred to as the out-of-bag (OOB) data (Breiman 2001). We conducted separate RF analyses for each of the three main response variables. In each analysis, 5000 regression trees were used to obtain a stable prediction using the party package in R (Horton et al., 2019; Ryo & Rillig, 2017; Strobl et al., 2007; Zeileis et al., 2008). To visualize our results, we created partial dependency plots (PDPs) of the relationships between each of response variable against scoring predictor variable from the RF using the R pdp package (Greenwell, 2017). PDPs do not display the data directly but are projections based on the model inferred by the RF. As such, the range of the variation displayed on the y-axis is the proportion of the range explained by the variable in question (according to the RF) i.e. PDPs for the more important predictors will often cover a wider range of the response scale compared to the less important predictors (Carlisle et al., 2009; Leach et al., 2018).

2.7.8 Factors relating to bacterial community composition

Ecological community analyses were conducted using the vegan package version 2.5-3 (Oksanen et al., 2019). To visualize dissimilarities in community composition across lakes, we generated a nonmetric multidimensional scaling (nMDS) plot with a two-dimensional solution based on Bray–Curtis dissimilarity of Hellinger-transformed community data (*metaMDS* function in vegan) for taxonomic variations (n = 403, stress = 0.140).

LCBD analyses were performed on bacterial composition using 999 permutations on Hellinger-transformed community data in the R package adespatial as well as β-diversity decomposition into replacement and richness components. Pairwise dissimilarities between assemblages were calculated as taxonomic turnover using Bray-Curtis dissimilarities.

Nonlinear relationships between β-diversity and untransformed environmental gradients were modeled in GDMs in the R package gdm. GDMs work by fitting compositional turnover to environmental gradients with flexible, monotonic I-splines (Ferrier et al., 2007; Rosauer et al., 2014). Site-pairs for computing pairwise dissimilarities between sites were weighted proportionally to the total number of sequences associated with each sample. Variable selection for GDMs was performed using backward elimination with 100 permutations per step. GDMs were computed for taxonomic turnover across each environmental category.

2.7.9 Statistical analyses

Data wrangling and statistical analysis were performed in R v. 4.0.4 (R_Core_Team, 2018).

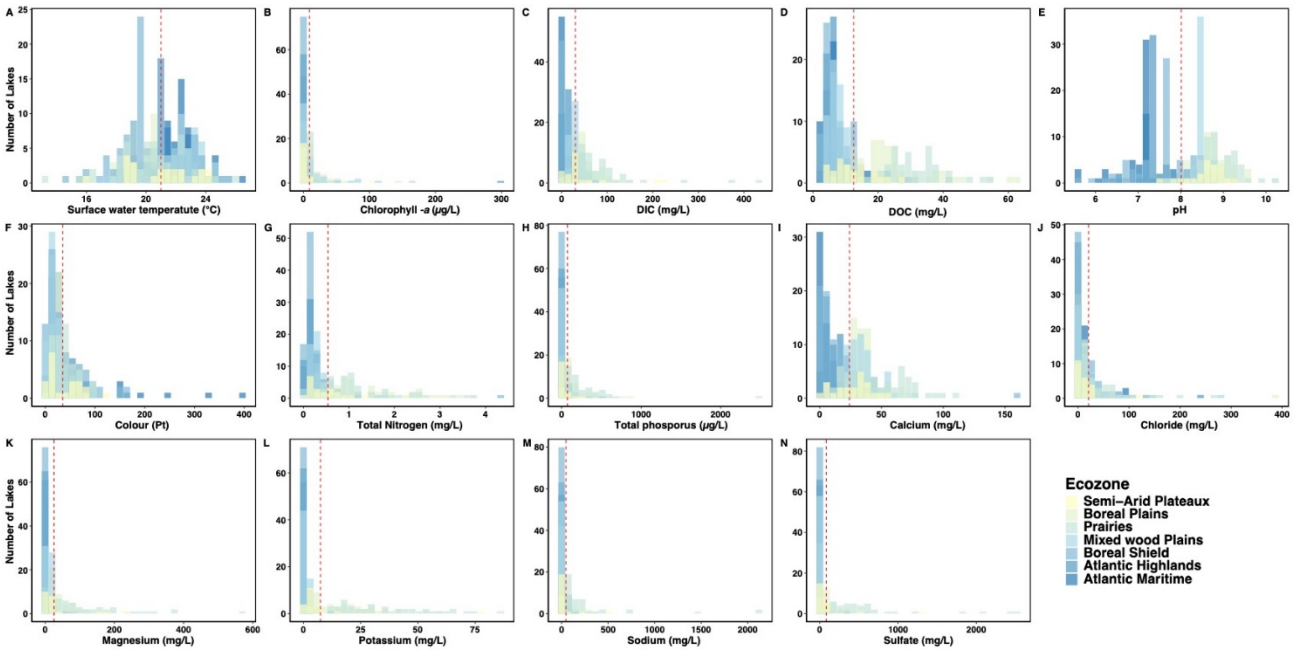
2.8 Data availability

Sequence data have been deposited in the European Nucleotide Archive under study accession PRJEB47327 (www.ebi.ac.uk).

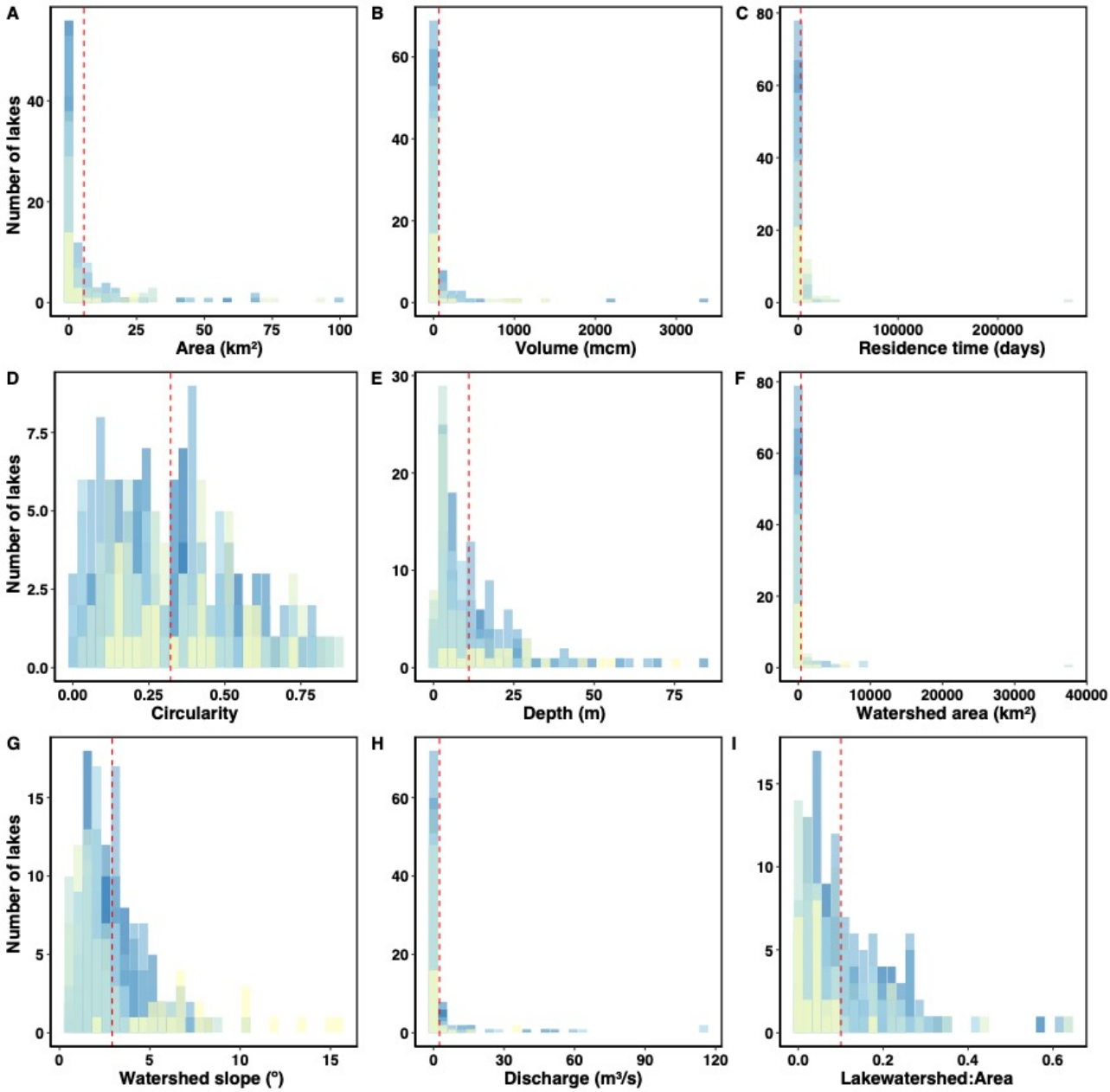
2.9 Funding and acknowledgement

This chapter was funded by the NSERC Canadian LakePulse Network (Strategic Network Grant NETGP-479720). I received scholarship from the NSERC CREATE ÉcoLac Training Program in Lake and Fluvial Ecology. I appreciate the LakePulse researchers involved in the generation, management, and quality control of the data set. Thanks to Genevieve Bourette, UQAM for support with illumina sequencing and Wentworth Brookes for bioinformatics technical support.

2.10 Supplementary Figures and Tables



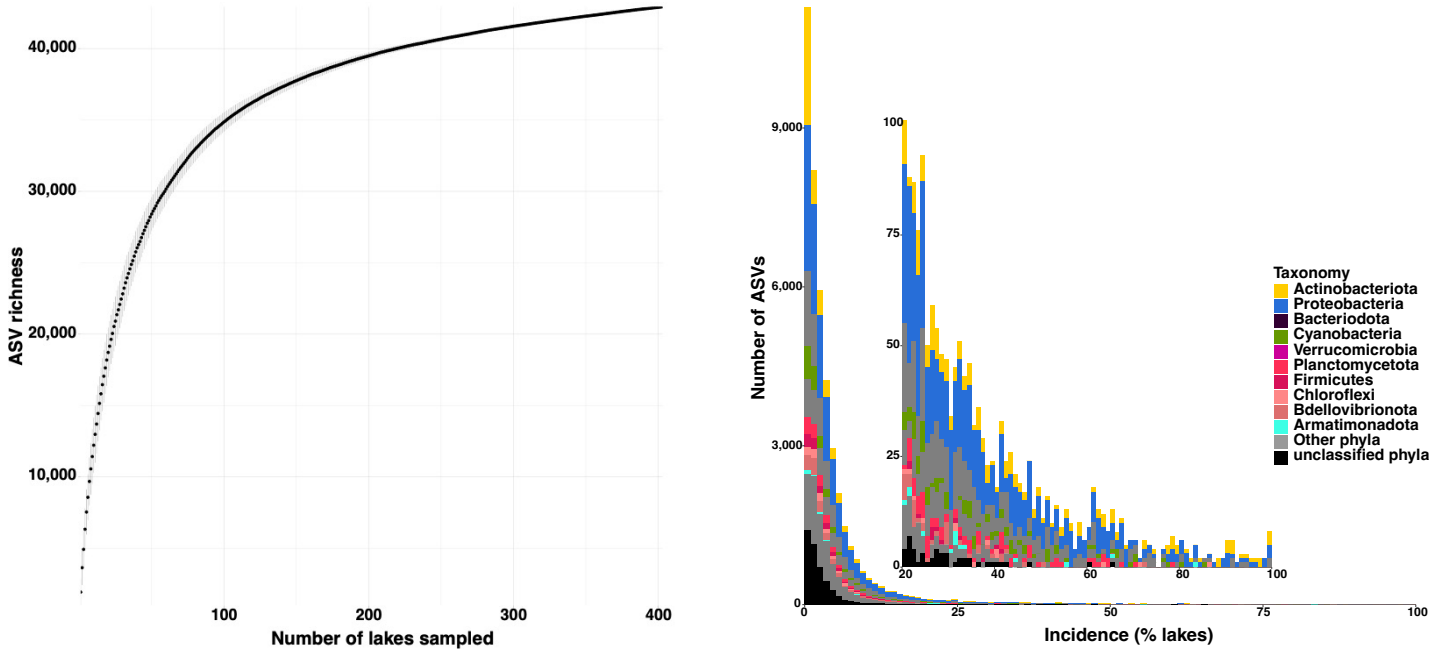
Supplementary Figure S2.1 *Ranges for physicochemistry and morphometric variables across ecozones (n =403). Red dotted line represents the mean value of a variable across lakes.*



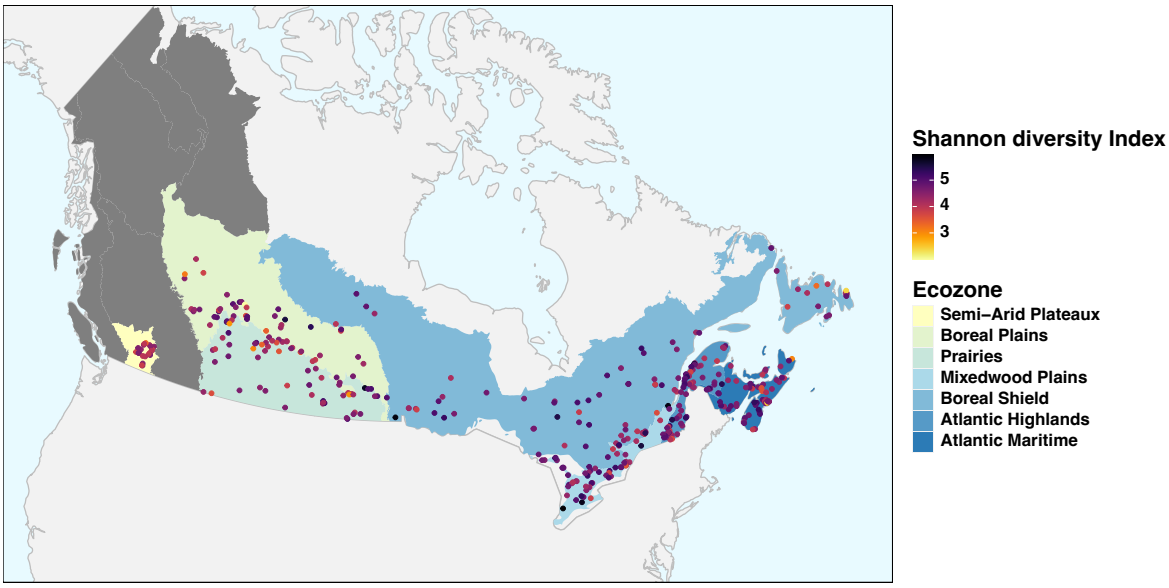
Supplementary Figure S2.1 *Ranges for physicochemistry and morphometric variables across ecozones (n =403). Red dotted line represents the mean value of a variable across lakes.*

Supplementary Table S2.1 *Mean pH, nutrient, and ion concentration across ecozones; highlighted ecozone has highest mean values across abiotic variables*

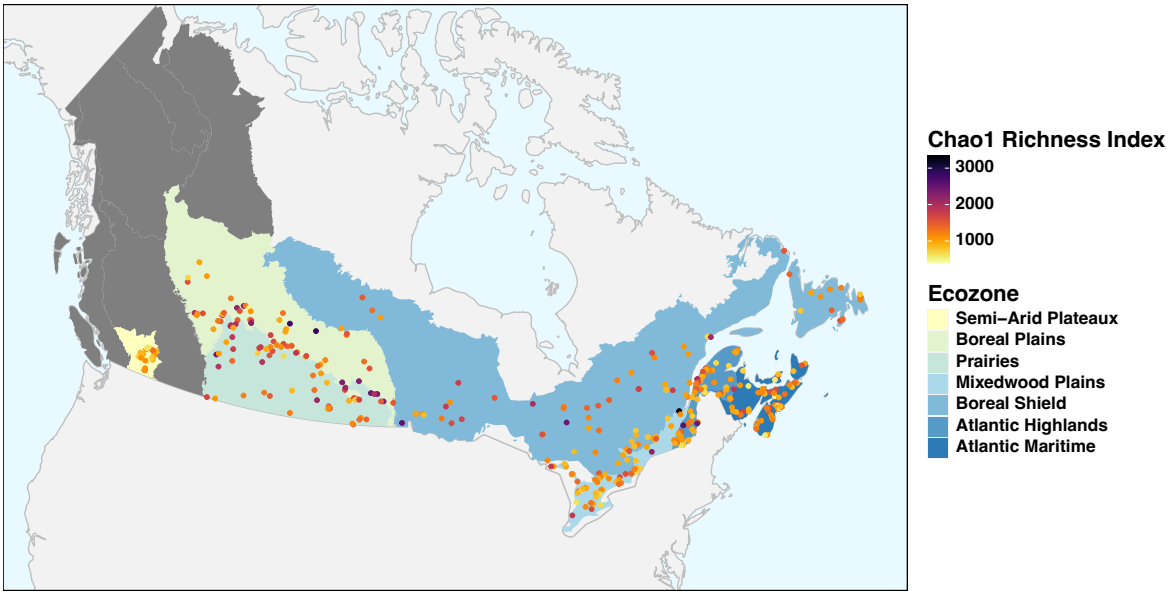
	Ions (mg/L)						Nutrients (mg/L; ug/L)					pH
	Cl^-	SO_4^{2-}	Ca^{2+}	K^+	Mg^{2+}	Na^{2+}	TN	TP	DOC	DIC	Chla	
Semi-Arid Plateaux	15.1	109.3	27.1	8.2	35.8	46.8	0.5	42.3	12.3	46.5	7.7	8.6
Boreal Plains	21.66	75.13	35.5	15.9	40.6	42.3	1.2	144	24.4	51.3	15.0	8.7
Prairies*	**40.3**	**487.0**	**46.9**	**32.3**	**114.2**	**251.1**	**1.3**	**263.5**	**26.6**	**92.5**	**15.7**	**8.9**
Boreal Shield	15.4	8.8	14.5	1.0	3.7	10.1	0.2	25.9	7.5	10.0	2.5	7.6
Mixedwood Plains	23.7	8.4	32.8	1.3	11.3	14.1	0.4	28.9	7.5	28.7	12.8	8.5
Atlantic Highland	8.7	4.2	15.4	1.0	4.2	4.5	0.2	26.0	8.2	11.7	9.4	7.5
Atlantic Maritime	22.1	5.4	7.8	0.7	1.5	15.4	0.2	22.3	6.3	4.6	3.7	7.1



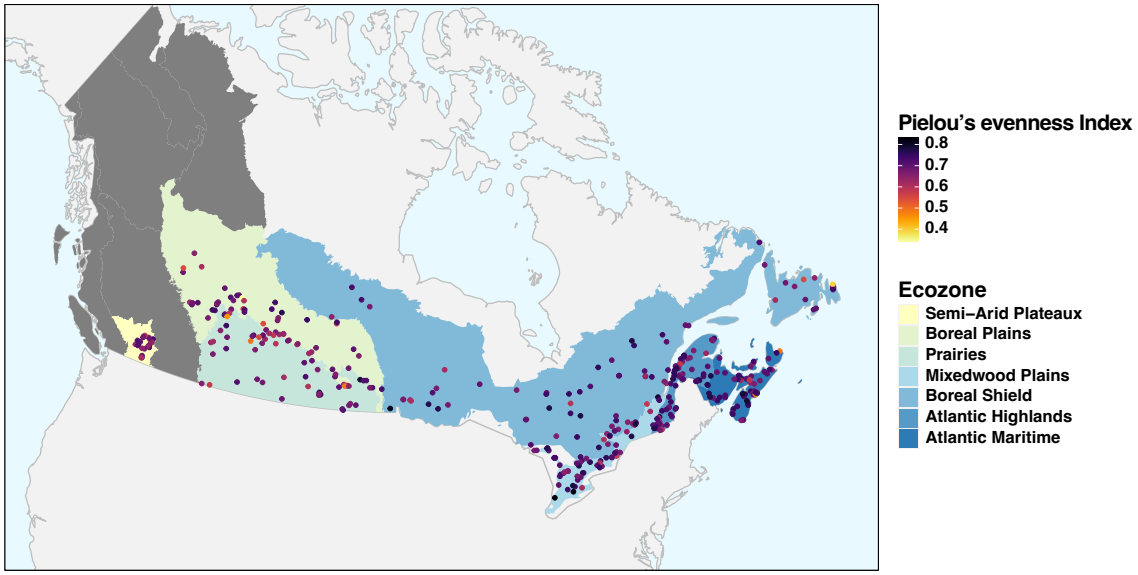
Supplementary Figures S2.2.1a and S2.2.1b *Accumulation curve and incidence plot of ASVs across lakes. (A) Accumulation curve of ASVs in a random ordering of lakes. Vertical bars are standard deviations. (B) Incidence of ASVs across lakes.*



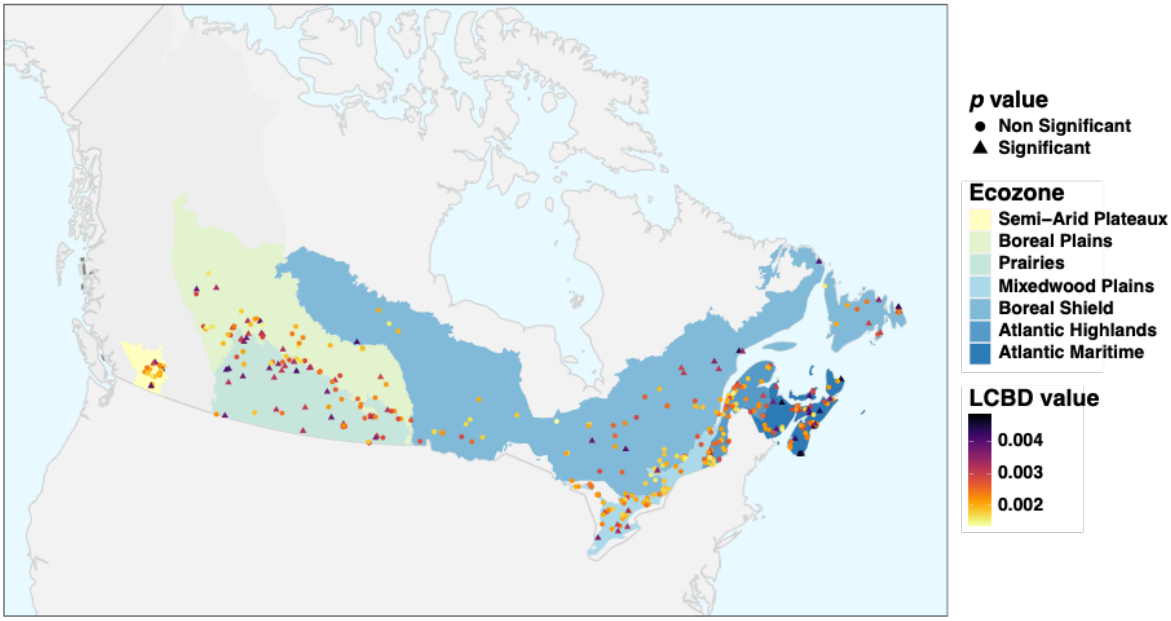
Supplementary Figure S2.2.2 *Bacterial diversity across lakes based on rarefied Shannon diversity index. Dark grey regions represent unsampled and lakes within Northern ecozones excluded from this study, while colored backgrounds indicate sampled ecozones.*



Supplementary Figure S2.2.3 *Bacterial richness across lakes based on rarefied Chao1 richness index. Dark grey regions represent unsampled and lakes within Northern ecozones excluded from this study, while colored backgrounds indicate sampled ecozones.*



Supplementary Figure S2.2.4 *Bacterial evenness across lakes based on rarefied Pielou's evenness index. Dark grey regions represent unsampled and lakes within Northern pristine ecozones excluded from this study, while colored backgrounds indicate sampled ecozones.*



Supplementary Figure S2.3 *Local contribution of bacterial assemblages to beta diversity (LCBD) across lakes. Dark grey regions represent unsampled and lakes within Northern ecozones excluded from this study, while colored backgrounds indicate sampled ecozones.*

Bibliography

Abirhire, O., Davies, J.-M., Imtiazy, N., Hunter, K., Emmons, S., Beadle, J., & Hudson, J. (2023). Response of phytoplankton community composition to physicochemical and meteorological factors under different hydrological conditions in Lake Diefenbaker. *Science of The Total Environment*, *856*, 159210. https://doi.org/10.1016/j.scitotenv.2022.159210

Abirhire, O., North, R. L., Hunter, K., Vandergucht, D. M., Sereda, J., & Hudson, J. J. (2015). Environmental factors influencing phytoplankton communities in Lake Diefenbaker, Saskatchewan, Canada. *Journal of Great Lakes Research*, *41*, 118–128. https://doi.org/10.1016/j.jglr.2015.07.002

Adams, H. E., Crump, B. C., & Kling, G. W. (2010). Temperature controls on aquatic bacterial production and community dynamics in arctic lakes and streams. *Environmental Microbiology*, *12*(5), 1319–1333. https://doi.org/10.1111/j.1462-2920.2010.02176.x

Aguilar, P., & Sommaruga, R. (2020). The balance between deterministic and stochastic processes in structuring lake bacterioplankton community over time. *Molecular Ecology*, *29*(16), 3117–3130. https://doi.org/10.1111/mec.15538

Beisner, B. E., Peres-Neto, P. R., Lindström, E. S., Barnett, A., & Longhi, M. L. (2006). The Role of Environmental and Spatial Processes in Structuring Lake Communities from Bacteria to Fish. *Ecology*, *87*(12), 2985–2991. https://doi.org/10.1890/0012-9658(2006)87[2985:TROEAS]2.0.CO;2

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32. https://doi-org.lib-ezproxy.concordia.ca/10.1023/A:1010933404324

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Carlisle, D. M., Falcone, J., & Meador, M. R. (2009). Predicting the biological condition of streams: Use of geospatial indicators of natural and anthropogenic characteristics of watersheds. *Environmental Monitoring and Assessment*, *151*(1), 143–160. https://doi.org/10.1007/s10661-008-0256-z

CCEA. (2016). Ecozone introduction. de http://www.ccea.org/ecozones-introduction/

Chen, J., Li, H., Zhang, Z., He, C., Shi, Q., Jiao, N., & Zhang, Y. (2020). DOC dynamics and bacterial community succession during long-term degradation of Ulva prolifera and their implications for the legacy effect of green tides on refractory DOC pool in seawater. *Water Research*, *185*, 116268. https://doi.org/10.1016/j.watres.2020.116268

Cotner, J. B., & Biddanda, B. A. (2002). Small Players, Large Role: Microbial Influence on Biogeochemical Processes in Pelagic Aquatic Ecosystems. *Ecosystems*, *5*(2), 105–121. https://doi.org/10.1007/s10021-001-0059-3

Evans, C. D., Chapman, P. J., Clark, J. M., Monteith, D. T., & Cresser, M. S. (2006). Alternative explanations for rising dissolved organic carbon export from organic soils. *Global Change Biology*, *12*(11), 2044–2053. https://doi.org/10.1111/j.1365-2486.2006.01241.x

Fergus, C. E., Finley, A. O., Soranno, P. A., & Wagner, T. (2016). Spatial Variation in Nutrient and Water Color Effects on Lake Chlorophyll at Macroscales. *PLOS ONE*, *11*(10), e0164592. https://doi.org/10.1371/journal.pone.0164592

Ferrier, S., Manion, G., Elith, J., & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity

assessment. *Diversity and Distributions*, *13*(3), 252–264. https://doi.org/10.1111/j.1472-4642.2007.00341.x

Fu, X.-Y., Cheng, Z.-C., Ni, H.-W., & Zhang, R.-T. (2023). Latitude variations of soil bacterial community diversity and composition in three typical forests of temperate, northeastern of China. *Frontiers in Earth Science*, *10*. https://www.frontiersin.org/articles/10.3389/feart.2022.1096931

Garner, R. E., Kraemer, S. A., Onana, V. E., Huot, Y., Gregory-Eaves, I., & Walsh, D. A. (2022). Protist Diversity and Metabolic Strategy in Freshwater Lakes Are Shaped by Trophic State and Watershed Land Use on a Continental Scale. *mSystems*, *0*(0), e00316-22. https://doi.org/10.1128/msystems.00316-22

Greenwell, B., M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, *9*(1), 421. https://doi.org/10.32614/RJ-2017-016

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Horton, D. J., Theis, K. R., Uzarski, D. G., & Learman, D. R. (2019). Microbial community structure and microbial networks correspond to nutrient gradients within coastal wetlands of the Laurentian Great Lakes. *FEMS Microbiology Ecology*, *95*(4), fiz033. https://doi.org/10.1093/femsec/fiz033

Huot, Y., Brown, C. A., Potvin, G., Antoniades, D., Baulch, H. M., Beisner, B. E., Bélanger, S., Brazeau, S., Cabana, H., Cardille, J. A., del Giorgio, P. A., Gregory-Eaves, I., Fortin, M.-J., Lang, A. S., Laurion, I., Maranger, R., Prairie, Y. T., Rusak, J. A., Segura, P. A., … Walsh, D. A. (2019a). The NSERC Canadian Lake Pulse Network: A national assessment of lake health providing science for water management in a changing climate. *Science of The Total Environment*, *695*, 133668. https://doi.org/10.1016/j.scitotenv.2019.133668

Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., Coelho, L. P., Endo, H., Gasol, J. M., Gregory, A. C., Mahé, F., Rigonato, J., Royo-Llonch, M., Salazar, G., Sanz-Sáez, I., Scalco, E., Soviadan, D., Zayed, A. A., Zingone, A., … Zinger, L. (2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, *179*(5), 1084-1097.e21. https://doi.org/10.1016/j.cell.2019.10.008

Jankowski, K., Schindler, D. E., & Horner-Devine, M. C. (2014). Resource Availability and Spatial Heterogeneity Control Bacterial Community Response to Nutrient Enrichment in Lakes. *PLOS ONE*, *9*(1), e86991. https://doi.org/10.1371/journal.pone.0086991

Kiersztyn, B., Chróst, R., Kaliński, T., Siuda, W., Bukowska, A., Kowalczyk, G., & Grabowska, K. (2019). Structural and functional microbial diversity along a eutrophication gradient of interconnected lakes undergoing anthropopressure. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-47577-8

Korhonen, J. J., Wang, J., & Soininen, J. (2011). Productivity-Diversity Relationships in Lake Plankton Communities. *PLOS ONE*, *6*(8), e22041. https://doi.org/10.1371/journal.pone.0022041

Kraemer, S. A., Barbosa da Costa, N., Shapiro, B. J., Fradette, M., Huot, Y., & Walsh, D. A. (2020). A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. *The ISME Journal*. https://doi.org/10.1038/s41396-020-0733-0

Ladau, J., Sharpton, T. J., Finucane, M. M., Jospin, G., Kembel, S. W., O'Dwyer, J., Koeppel, A. F., Green, J. L., & Pollard, K. S. (2013). Global marine bacterial diversity peaks at high latitudes in winter. *The ISME Journal*, *7*(9), Article 9. https://doi.org/10.1038/ismej.2013.37

Lahti L, Sudarshan Shetty S., et al. (2017). Tools for microbiome analysis in R. Version . URL: http://microbiome.github.com/microbiome.

Lambert, T., & Perga, M.-E. (2019). Non-conservative patterns of dissolved organic matter degradation when and where lake water mixes. *Aquatic Sciences*, *81*(4), 64. https://doi.org/10.1007/s00027-019-0662-z

Lauber, C. L., Ramirez, K. S., Aanderud, Z., Lennon, J., & Fierer, N. (2013). Temporal variability in soil microbial communities across land-use types. *The ISME Journal*, *7*(8), Article 8. https://doi.org/10.1038/ismej.2013.50

Leach, T. H., Beisner, B. E., Carey, C. C., Pernica, P., Rose, K. C., Huot, Y., Brentrup, J. A., Domaizon, I., Grossart, H.-P., Ibelings, B. W., Jacquet, S., Kelly, P. T., Rusak, J. A., Stockwell, J. D., Straile, D., & Verburg, P. (2018). Patterns and drivers of deep chlorophyll maxima structure in 100 lakes: The relative importance of light and thermal stratification. *Limnology and Oceanography*, *63*(2), 628–646. https://doi.org/10.1002/lno.10656

Linz, A. M., Crary, B. C., Shade, A., Owens, S., Gilbert, J. A., Knight, R., & McMahon, K. D. (2017). Bacterial Community Composition and Dynamics Spanning Five Years in Freshwater Bog Lakes. *mSphere*, *2*(3), e00169-17. https://doi.org/10.1128/mSphere.00169-17

Linz, A. M., He, S., Stevens, S. L. R., Anantharaman, K., Rohwer, R. R., Malmstrom, R. R., Bertilsson, S., & McMahon, K. D. (2018). Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *PeerJ*, *6*, e6075. https://doi.org/10.7717/peerj.6075

Liu, B., Arlotti, D., Huyghebaert, B., & Tebbe, C. C. (2022). Disentangling the impact of contrasting agricultural management practices on soil microbial communities – Importance of rare bacterial community members. *Soil Biology and Biochemistry*, *166*, 108573. https://doi.org/10.1016/j.soilbio.2022.108573

Liu, K., Yao, T., Pearce, D. A., Jiao, N., Zeng, Y., Guo, B., & Liu, Y. (2021). Bacteria in the lakes of the Tibetan Plateau and polar regions. *Science of The Total Environment*, *754*, 142248. https://doi.org/10.1016/j.scitotenv.2020.142248

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H., & Staley, J. T. (2006). Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology*, *4*(2), Article 2. https://doi.org/10.1038/nrmicro1341

Mateus-Barros, E., de Melo, M. L., Bagatini, I. L., Caliman, A., & Sarmento, H. (2021). Local and Geographic Factors Shape the Occupancy-Frequency Distribution of Freshwater Bacteria. *Microbial Ecology*, *81*(1), 26–35. https://doi.org/10.1007/s00248-020-01560-3

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of*

Reading : The Official Journal of the Society for the Scientific Study of Reading, *20*(1), 20–33. https://doi.org/10.1080/10888438.2015.1107073

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13603

Minns, C. K., Moore, J. E., Shuter, B. J., & Mandrak, N. E. (2008). A preliminary national analysis of some key characteristics of Canadian lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(8), 1763–1778. https://doi.org/10.1139/F08-110

Morrison, J. M., Baker, K. D., Zamor, R. M., Nikolai, S., Elshahed, M. S., & Youssef, N. H. (2017). Spatiotemporal analysis of microbial community dynamics during seasonal stratification events in a freshwater lake (Grand Lake, OK, USA). *PLOS ONE*, *12*(5), e0177488. https://doi.org/10.1371/journal.pone.0177488

Mou, X., Jacob, J., Lu, X., Robbins, S., Sun, S., & Ortiz, J. D. (2013). Diversity and distribution of free-living and particle-associated bacterioplankton in Sandusky Bay and adjacent waters of Lake Erie Western Basin. *Journal of Great Lakes Research*, *39*(2), 352–357. https://doi.org/10.1016/j.jglr.2013.03.014

Muñoz Sabater, J. (2019). *ERA5-Land hourly data from 1981 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). https://doi.org/10.24381/cds.e2161bac

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews: MMBR*, *75*(1), 14–49. https://doi.org/10.1128/MMBR.00028-10

NSERC Canadian Lake Pulse Network. (2021). *NSERC Canadian Lake Pulse Network field manual 2017—2018—2019 surveys* (M.-P. Varin, M.-L. Beaulieu, & Y. Huot, Eds.). Université de Sherbrooke. https://doi.org/10.17118/11143/18662

Nürnberg, G. K., & Shaw, M. (1998). Productivity of clear and humic lakes: Nutrients, phytoplankton, bacteria. *Hydrobiologia*, *382*(1), 97–112. https://doi.org/10.1023/A:1003445406964

Obieze, C. C., Wani, G. A., Shah, M. A., Reshi, Z. A., Comeau, A. M., & Khasa, D. P. (2022). Anthropogenic activities and geographic locations regulate microbial diversity, community assembly and species sorting in Canadian and Indian freshwater lakes. *Science of The Total Environment*, *826*, 154292. https://doi.org/10.1016/j.scitotenv.2022.154292

Oksanen, J., Blanchet, F. G., Kindt, R., Legen-, P., Minchin, P. R., Hara, R. B. O., Simpson, G. L., Soly-, P., Stevens, M. H. H., & Wagner, H. (2019). vegan: Community ecology package. *R Package Version 2.5-4*. https://doi.org/ISBN 0-387-95457-0

Olapade, O. A. (2018). Community Composition and Diversity of Coastal Bacterioplankton Assemblages in Lakes Michigan, Erie, and Huron. *Microbial Ecology*, *75*(3), 598–608. https://doi.org/10.1007/s00248-017-1081-0

Oliva, A., Garner, R. E., Walsh, D., & Huot, Y. (2022). The occurrence of potentially pathogenic fungi and protists in Canadian lakes predicted using geomatics, in situ and satellite-derived variables: Towards a tele-epidemiological approach. *Water Research*, *209*, 117935. https://doi.org/10.1016/j.watres.2021.117935

Oliva, A., Onana, V. E., Garner, R. E., Kraemer, S. A., Fradette, M., Walsh, D. A., & Huot, Y. (2023). Geospatial analysis reveals a hotspot of fecal bacteria in Canadian prairie lakes linked to agricultural non-point sources. *Water Research*, *231*, 119596. https://doi.org/10.1016/j.watres.2023.119596

Paver, S. F., Newton, R. J., & Coleman, M. L. (2020). Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environmental Microbiology*, *22*(1), 433–446. https://doi.org/10.1111/1462-2920.14862

Pernthaler, J. (2017). Competition and niche separation of pelagic bacteria in freshwater habitats. *Environmental Microbiology*, *19*(6), 2133–2150. https://doi.org/10.1111/1462-2920.13742

Pinnell, L. J., & Turner, J. W. (2020). Temporal changes in water temperature and salinity drive the formation of a reversible plastic-specific microbial community. *FEMS Microbiology Ecology*, *96*(12), fiaa230. https://doi.org/10.1093/femsec/fiaa230

Pruesse, E., Peplies, J., & Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, *28*(14), 1823–1829. https://doi.org/10.1093/bioinformatics/bts252

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

Ramette, A., & Tiedje, J. M. (2007). Multiscale responses of microbial life to spatial distance and environmental heterogeneity in a patchy ecosystem. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(8), 2761–2766. https://doi.org/10.1073/pnas.0610671104

Rchiad, Z., Dai, M., Hamel, C., Bainard, L. D., Cade-Menun, B. J., Terrat, Y., St-Arnaud, M., & Hijri, M. (2022). Soil Depth Significantly Shifted Microbial Community Structures and Functions in a Semiarid Prairie Agroecosystem. *Frontiers in Microbiology*, *13*. https://www.frontiersin.org/articles/10.3389/fmicb.2022.815890

R_Core_Team. (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*.

Rosauer, D. F., Ferrier, S., Williams, K. J., Manion, G., Keogh, J. S., & Laffan, S. W. (2014). Phylogenetic generalised dissimilarity modelling: A new approach to analysing and predicting spatial turnover in the phylogenetic composition of communities. *Ecography*, *37*(1), 21–32. https://doi.org/10.1111/j.1600-0587.2013.00466.x

Ruiz-González, C., Niño-García, J. P., Lapierre, J.-F., & del Giorgio, P. A. (2015). The quality of organic matter shapes the functional biogeography of bacterioplankton across boreal freshwater ecosystems. *Global Ecology and Biogeography*, *24*(12), 1487–1498. https://doi.org/10.1111/geb.12356

Ruuskanen, M. O., St. Pierre, K. A., St. Louis, V. L., Aris-Brosou, S., & Poulain, A. J. (2018). Physicochemical Drivers of Microbial Community Structure in Sediments of Lake Hazen, Nunavut, Canada. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.01138

Ryo, M., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, *8*(11), e01976. https://doi.org/10.1002/ecs2.1976

Saccà, M. L., Ferrero, V. E. V., Loos, R., Di Lenola, M., Tavazzi, S., Grenni, P., Ademollo, N., Patrolecco, L., Huggett, J., Caracciolo, A. B., & Lettieri, T. (2019). Chemical mixtures and fluorescence in situ hybridization analysis of natural microbial community in the Tiber river. *Science of The Total Environment*, *673*, 7–19. https://doi.org/10.1016/j.scitotenv.2019.04.011

Sadeghi, J., Chaganti, S. R., Shahraki, A. H., & Heath, D. D. (2021). Microbial community and abiotic effects on aquatic bacterial communities in north temperate lakes. *Science of The Total Environment*, *781*, 146771. https://doi.org/10.1016/j.scitotenv.2021.146771

Sagova-Mareckova, M., Boenigk, J., Bouchez, A., Cermakova, K., Chonova, T., Cordier, T., Eisendle, U., Elersek, T., Fazi, S., Fleituch, T., Frühe, L., Gajdosova, M., Graupner, N., Haegerbaeumer, A., Kelly, A.-M., Kopecky, J., Leese, F., Nõges, P., Orlic, S., … Stoeck, T. (2021). Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Research*, *191*, 116767. https://doi.org/10.1016/j.watres.2020.116767

Saleem, M., Fetzer, I., Harms, H., & Chatzinotas, A. (2016). Trophic complexity in aqueous systems: Bacterial species richness and protistan predation regulate dissolved organic carbon and dissolved total nitrogen removal. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1825), 20152724. https://doi.org/10.1098/rspb.2015.2724

Shahraki, A. H., Chaganti, S. R., & Heath, D. (2021). Spatio-temporal dynamics of bacterial communities in the shoreline of Laurentian great Lake Erie and Lake St. Clair's large freshwater ecosystems. *BMC Microbiology*, *21*, 253. https://doi.org/10.1186/s12866-021-02306-y

Smith, V. H. (2007). Microbial diversity–productivity relationships in aquatic ecosystems. *FEMS Microbiology Ecology*, *62*(2), 181–186. https://doi.org/10.1111/j.1574-6941.2007.00381.x

Sperlea, T., Kreuder, N., Beisser, D., Hattab, G., Boenigk, J., & Heider, D. (2021). Quantification of the covariation of lake microbiomes and environmental variables using a machine learning-based framework. *Molecular Ecology*, *30*(9), 2131–2144. https://doi.org/10.1111/mec.15872

Sterner, R. W., Keeler, B., Polasky, S., Poudel, R., Rhude, K., & Rogers, M. (2020). Ecosystem services of Earth's largest freshwater lakes. *Ecosystem Services*, *41*, 101046. https://doi.org/10.1016/j.ecoser.2019.101046

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Sun, P., Wang, Y., Huang, X., Huang, B., & Wang, L. (2022). Water masses and their associated temperature and cross-domain biotic factors co-shape upwelling microbial communities. *Water Research*, *215*, 118274. https://doi.org/10.1016/j.watres.2022.118274

Tian, J., He, N., Hale, L., Niu, S., Yu, G., Liu, Y., Blagodatskaya, E., Kuzyakov, Y., Gao, Q., & Zhou, J. (2018). Soil organic matter availability and climate drive latitudinal patterns in bacterial diversity from tropical to cold temperate forests. *Functional Ecology*, *32*(1), 61–70. https://doi.org/10.1111/1365-2435.12952

Tran, P., Ramachandran, A., Khawasik, O., Beisner, B. E., Rautio, M., Huot, Y., & Walsh, D. A. (2018). Microbial life under ice: Metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes. *Environmental Microbiology*, *20*(7), 2568–2584. https://doi.org/10.1111/1462-2920.14283

Tranvik, L. J., Cole, J. J., & Prairie, Y. T. (2018). The study of carbon in inland waters—From isolated ecosystems to players in the global carbon cycle. *Limnology and Oceanography Letters*, *3*(3), 41–48. https://doi.org/10.1002/lol2.10068

Verpoorter, C., Kutser, T., Seekell, D. A., & Tranvik, L. J. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters*, *41*(18), 6396–6402. https://doi.org/10.1002/2014GL060641

Wang, J., Wang, X., Liu, G., Zhang, C., & Wang, G. (2021). Bacterial richness is negatively related to potential soil multifunctionality in a degraded alpine meadow. *Ecological Indicators*, *121*, 106996. https://doi.org/10.1016/j.ecolind.2020.106996

Wang, Y., Guo, M., Li, X., Liu, G., Hua, Y., Zhao, J., Huguet, A., & Li, S. (2022). Shifts in microbial communities in shallow lakes depending on trophic states: Feasibility as an evaluation index for eutrophication. *Ecological Indicators*, *136*, 108691. https://doi.org/10.1016/j.ecolind.2022.108691

Webster, K. E., Soranno, P. A., Cheruvelil, K. S., Bremigan, M. T., Downing, J. A., Vaux, P. D., Asplund, T. R., Bacon, L. C., & Connor, J. (2008). An empirical evaluation of the nutrient-color paradigm for lakes. *Limnology and Oceanography*, *53*(3), 1137–1148. https://doi.org/10.4319/lo.2008.53.3.1137

Wiken, E. B., Gauthier, D., Marshall, I., Lawton, K., & Hirvonen, H. (1996). A perspective on Canada's ecosystems: An overview of the terrestrial and marine ecozones. In *Canadian Council on Ecological Areas Occasional Papers* (Vol. 14).

Yang, Z., Hosokawa, H., Kuroda, M., Inoue, D., & Ike, M. (2021). Microbial antimonate reduction and removal potentials in river sediments. *Chemosphere*, *266*, 129192. https://doi.org/10.1016/j.chemosphere.2020.129192

Yannarell, A. C., & Triplett, E. W. (2004). Within- and between-lake variability in the composition of bacterioplankton communities: Investigations using multiple spatial scales. *Applied and Environmental Microbiology*, *70*(1), 214–223. https://doi.org/10.1128/AEM.70.1.214-223.2004

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Chapter 3: Water quality and land use shape bacterial communities across 621 Canadian lakes

3.1 Abstract

Human activities such as agriculture and urban development are linked to water quality degradation. Degradation can occur through an influx of excess nutrients (i.e. eutrophication), as well as the influx of synthetic contaminants and pathogens. Bacterial communities are closely linked to water quality via nutrient cycling, degradation of contaminants, and as potential pathogens of animals and humans. Canada represents a large and heterogenous landscape of freshwater lakes, where variation in climate, geography, and geology interact with land cover alteration to influence water quality differently across regions. In this study, we investigated the influence of water quality and land use variables on bacterial communities across 12 ecozones that represent large regional difference of the Canadian landscape. At the pan-Canadian scale, total phosphorous (TP) was the most significant water quality variable influencing community structure, and the most pronounced shift was observed at 110 ug/L TP, corresponding to the transition from eutrophic to hypereutrophic conditions. At the regional scale, despite significant regional differences in environmental conditions, water quality significantly explained bacterial community structure in all ecozones. At the pan-Canadian scale, agriculture and, to a lesser extent, urbanization were significant land use variables influencing community structure. In ecozones characterized by extensive agriculture, this land cover variable was consistently significant in explaining community structure. Likewise, in extensively urbanized ecozones, urbanization was consistently significant in explaining community structure. Agriculture was associated with an increase in bacterial diversity, and we observed more taxa increase than decrease along the pan-Canadian and regional agricultural gradients. In contrast, bacterial diversity did not change significantly in relation to urbanization, and we observed a similar number of increasing and decreasing taxa along urbanization gradients. Overall, these results demonstrate that bacterial community diversity and community composition are influenced by water quality and shaped by agriculture and urban development in different ways.

3.2 Introduction

Human population growth is leading to an increase in agricultural and urban development (Vitousek et al., 1997). For example, Canada has witnessed a 34% growth in the urbanized areas of major cities over the past two decades (Bouchard and Shiab, 2022). This allocation of land to human endeavors is transforming terrestrial and aquatic ecosystems (Ahmed et al., 2022; Vitousek et al., 1997). Freshwater ecosystems are particularly susceptible to human influence primarily due to the process of cultural eutrophication (Smith, 2003). Eutrophication has increased the frequency of harmful algal blooms, posing threats to water security and the provision of ecosystem services (Ho et al., 2019). Water quality parameters such as phosphorus, nitrogen, organic carbon, chlorophyll-*a* concentrations, and light intensity (Yannarell & Triplett, 2005; Zeglin, 2015) may vary differently within regions compared to across regions. These variations are related to the input of nutrients and other materials from the surrounding terrestrial watershed (Chen et al., 2018), within regions governed by climatic and geological factors, thus affecting ecoregion definition at continental scales. Consequently, different land use types (such as agriculture and urbanization) may influence the quantity of watershed inputs into lakes (Solomon et al., 2015) in diverse and regionally specific ways.

Addressing the relationship between lake water quality and bacterial diversity is essential as lake bacteria play pivotal roles in ecological processes such as nutrient transformation and the decomposition of organic matter thereby contributing to biogeochemical cycles (Cotner & Biddanda, 2002; Linz et al., 2018). While there have been efforts to identify the environmental factors influencing freshwater bacterial communities (Bock et al., 2020; Langenheder & Lindström, 2019; Liu et al., 2020; Newton et al., 2011; Williamson et al., 2009), it is critical to acknowledge the multi-scale nature of these factors—whether within specific regions such as ecozones or across regions at a larger continental scale. Existing studies have explored changes in bacterial community structure in response to diverse environmental and watershed land use conditions in different regions. For example, Kraemer et al., (2020) investigated influences within eastern ecozone regions of Canada; other studies have investigated regions within Northern and Southern Ontario as well as the Laurentians (MacLeod et al., 2017; Paver et al., 2020; Sadeghi et al., 2021; Shahraki et al., 2021). In contrast, Garner et al. (2023) investigated microbial community composition at much larger (i.e. continental) scale in Canada, but did not investigate how communities varied in different regions. These studies have addressed either the regional or the continental scale influences of environmental conditions exacerbated by land use on bacterial diversity and community composition. However,

depending on spatial scale and regional differences, observations do not always result in the same environmental drivers of bacterial community shifts. A comprehensive comparison of changes within the different regional ecozones and across the country remains to be performed.

Recently, a coordinated study of over 600 Canadian lakes found that agricultural and urban land use are primary factors explaining variation in water quality among a diversity of land use and cover categories (Schacht et al., 2023). In this chapter, I build on this study by addressing the influence of water quality and land use on bacterial community diversity and composition in these same lakes. Specifically, I examined lake bacterial community composition and its potential environmental drivers across the country and within the different regional ecozones. This is the first such regional comparison, drawing data from hundreds of lakes (with varying limnological conditions and watershed characteristics) sampled by the LakePulse Network in a collective effort aimed at assessing lake health across Canada (Huot et al., 2019). We investigated bacterial diversity and community composition in surface waters across 621 freshwater lakes situated in 12 Canadian ecozone using 16S rRNA amplicon analyses. We hypothesize that such varying patterns in water quality will elicit responses from lake bacterial communities, and that the LakePulse lakes thus also provide a unique opportunity to identify specific bacterial taxa exhibiting increases or decreases along gradients of water quality and land use.

3.3 Results

3.3.1 Variation in lake water quality and watershed land use

The study encompassed 621 lakes from 12 ecozones (**Figure 3.1a**). To summarize the variation in lake water quality, we conducted a Principal Component Analysis (PCA) on the water quality data. Principal Component axis 1 (PC1) explained 50.4% of the variance. Total phosphorus (TP), total nitrogen (TN), chlorophyll a (Chl-*a*), and dissolved organic carbon (DOC) had large positive loadings on PC1 and were associated with higher proportions of agriculture and pasture (**Figure 3.1b**). Specific conductance, ions (Na^+ and K^+), and dissolved inorganic carbon (DIC) also loaded strongly on PC1 and were associated with agriculture and urban development. Lakes in the Prairies and Boreal Plains ecozones (positive PC1 loadings) tended to differ from lakes in the Mixedwood Plains and the Atlantic and Pacific Maritimes (negative PC1 loadings) (**Figure 3.1c**). The difference was supported by a one-way ANOVA with post hoc Tukey Test ($p < 2e\text{-}16$).

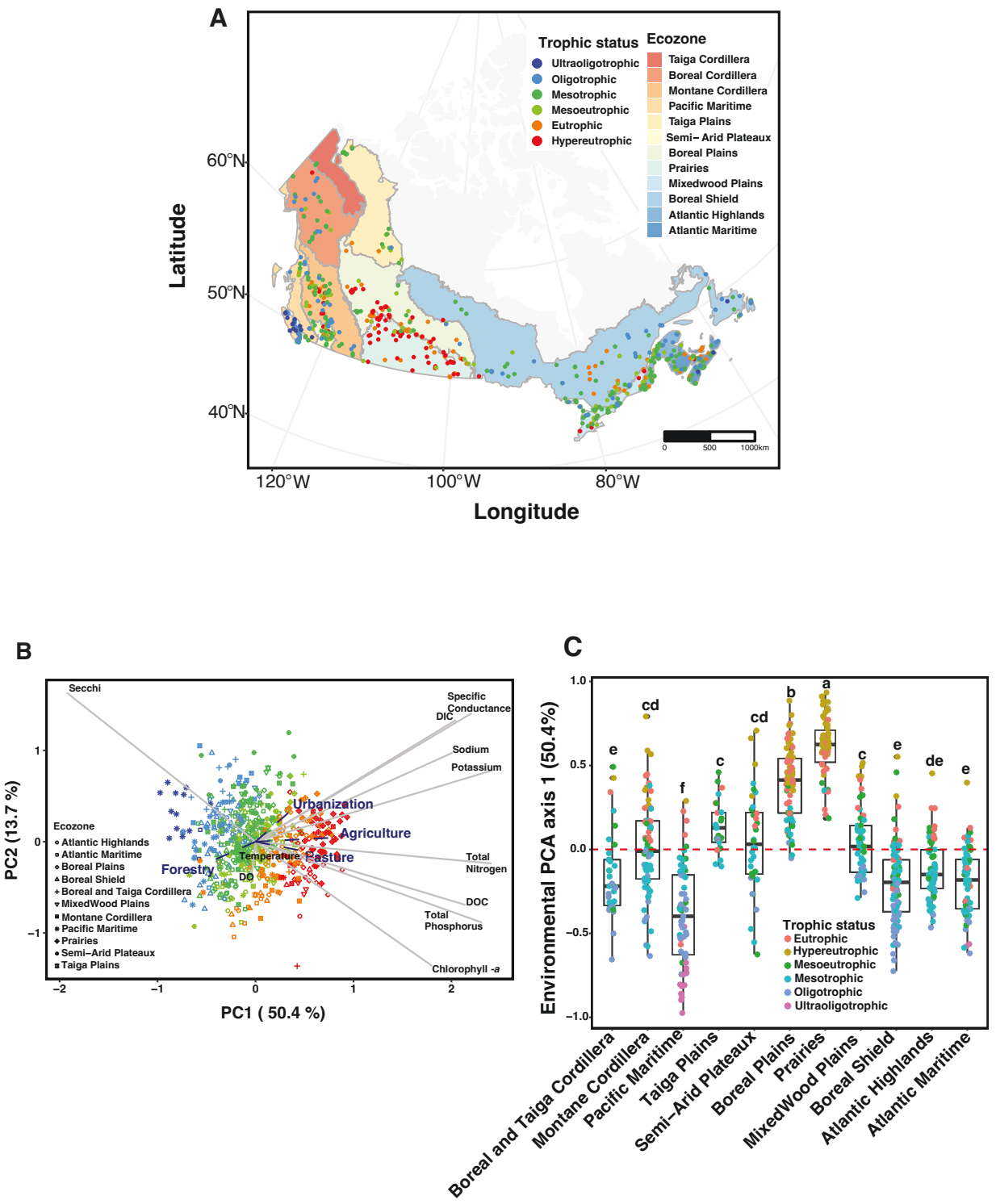


Figure 3.1 *Distribution and environmental conditions of lakes at the continental scale. A) Map showing the distribution of the 621 lakes across 12 Canadian ecozones (43 – 57 °N, 53 – 120 °W). Lakes are coloured*

according to trophic status. B) Principal component analysis of lake water quality variables from the LakePulse dataset (N = 621). The PCA highlights the distribution of land use types plotted passively (coloured in blue, dashed lines) over site scores, colour-coded by lake trophic status and shaped according to ecozone. C) Boxplot of site scores from PC axis 1 grouped and colour-coded by lake trophic status, where the width of the boxes reflects sample size ranging from 25 to 88 lakes per ecozone. Letter superscripts denote significantly different groups determined via a post-hoc Tukey's test.

3.3.2 Bacterial community diversity

Bacterial diversity was assessed using the V4 region of the 16S rRNA gene. After rarefaction we identified a total of 26,196 amplicon sequence variant (ASVs). Taxonomic classification of ASVs using the FreshTrain taxonomy identified 1,063 clade/genera level taxonomic units. The clades encompassed 27% of the ASVs, and 76% of the sequences, and therefore represent the most common and abundant taxa found in freshwaters. Chao-1 estimates of clade diversity varied across ecozones (**Figure 3.2a**). Higher richness was associated with the Prairies and Boreal Plains and lower richness was associated with the Mixedwood Plains. The significance of these differences was supported by a one-way ANOVA ($p < 2e\text{-}16$) and Tukey Test. Random Forest analysis showed that water quality and land use explained 58% of the variation in bacterial diversity (OOB 29%). Water quality variables, specifically potassium (K^+) (26%) and TP (18%) were the most important factors affecting Chao-1 richness, with a lesser importance of land use variables (**Supplementary Figure S3.1**).

To summarize the variation of bacterial community composition, we conducted PCoA (**Figure 3.2b**). The first axis (PCoA1) explained 15.9% of the variance and tended to separate bacterial communities in Prairies and Boreal Plains lakes (positive PCoA1 loading) from those in lakes in the Mixedwood Plains and the Atlantic and Pacific Maritimes).

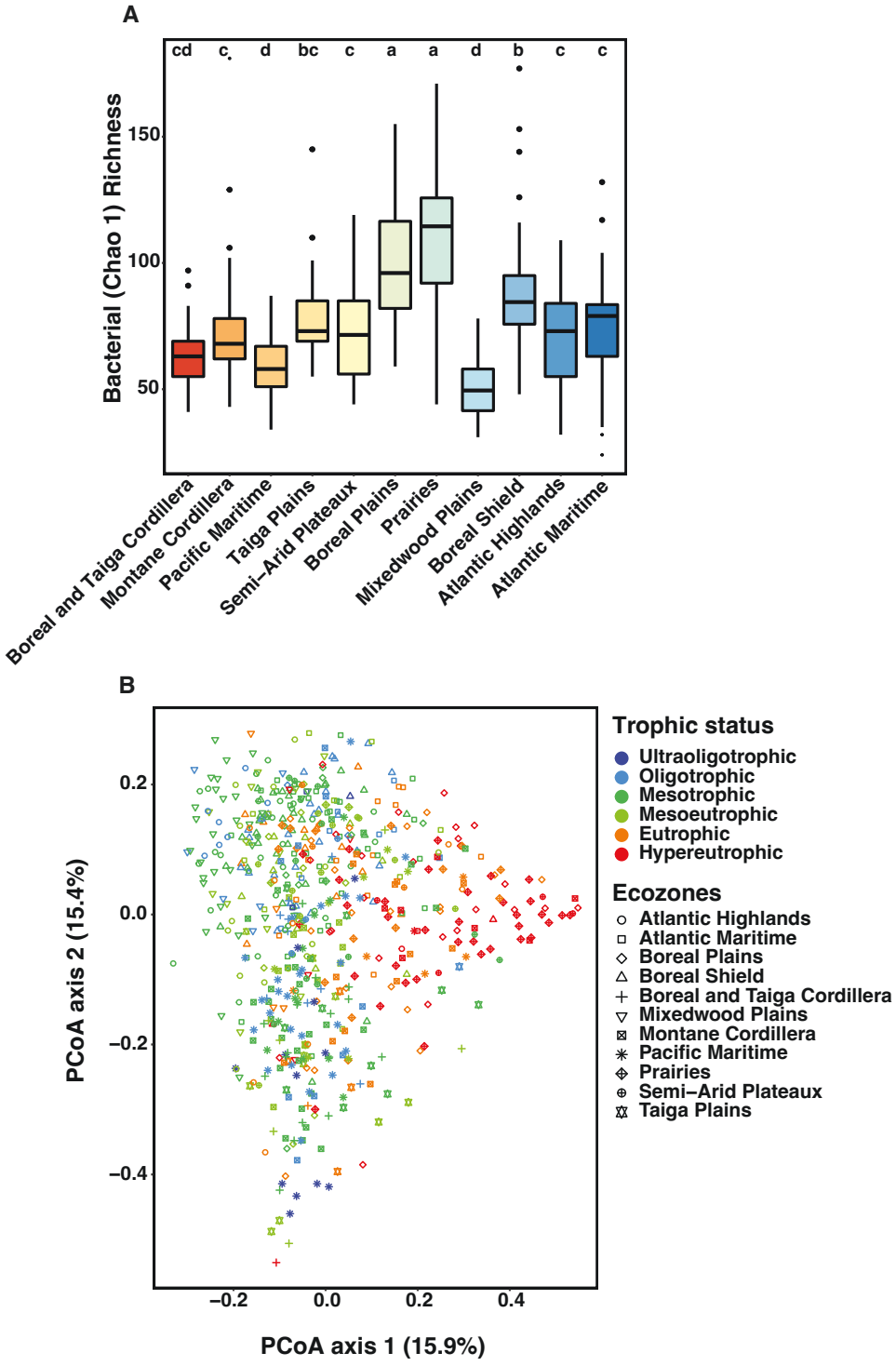


Figure 3.2 *Bacterial diversity and community composition across lakes. A) Boxplot of bacterial richness colour-coded by ecozones. Letter superscripts denote significantly different ecozones determined via a post-hoc Tukey's test. B) Principal component analysis (PCOA) highlighting the Bray-Curtis dissimilarity of bacterial community composition. Lakes are coloured according to trophic status and shaped according to ecozones.*

3.3.3 Water quality influence on bacterial community structure across ecozones

We assessed the importance of water quality for bacterial community structure using distance-based Redundancy Analysis (dbRDA) (**Figure 3.3a**). The water quality model was significant ($p < 0.001$) and explained 18 % of the variation. The dbRDA ordination revealed similar gradients as the PCoA. Nutrients (TP and TN), Chl-*a*, and ions (Na^+ and K^+) had strong loadings on RDA axis 1, which separated lakes in the Prairies and Boreal Plains (positive loadings) ecozones from those in the Mixedwood Plains and the Atlantic and Pacific Maritimes (negative loadings). The significance of this difference between ecozones was supported by a one-way ANOVA and Tukey Test ($p < 2e\text{-}16$) (**Figure 3.3b**).

We used indicator species analysis (ISA) to characterize the distribution of bacterial taxa across lakes of different trophic state (based on TP) from ultraoligotrophic to hypereutrophic. ISA identified 184 taxa that were associated with each trophic state (**Figure 3.3c**). Almost all indicator taxa were associated with extreme trophic states, either hypereutrophic (106 taxa) or ultraoligotrophic (26 taxa). Indicator taxa represented a broad phylogenetic diversity.

The high number of indicator taxa for hypereutrophic conditions suggested a rapid change in the bacterial community at high nutrient concentrations. We used Threshold Indicator Taxa Analysis (TITAN) to better pinpoint the threshold of TP concentration where the abrupt change occurred and to identify the taxa associated with the change. We identified 228 increasers (positive (z+) indicators) and 67 decreasers (negative (z-) indicator) along the TP gradient (**Figure 3.3d**). Most positive indicator taxa increased sharply between 10 ug/L and 100 ug/L [TP], resulting in a distinct filtered sum (z+) peak at ~110 ug/L TP, corresponding to the TP difference between eutrophic and hypereutrophic lakes (**Figure 3.3e**). The negative indicator taxa declined at varying [TP], resulting in relatively flat sum (z-) peak at 100 ug/L TP.

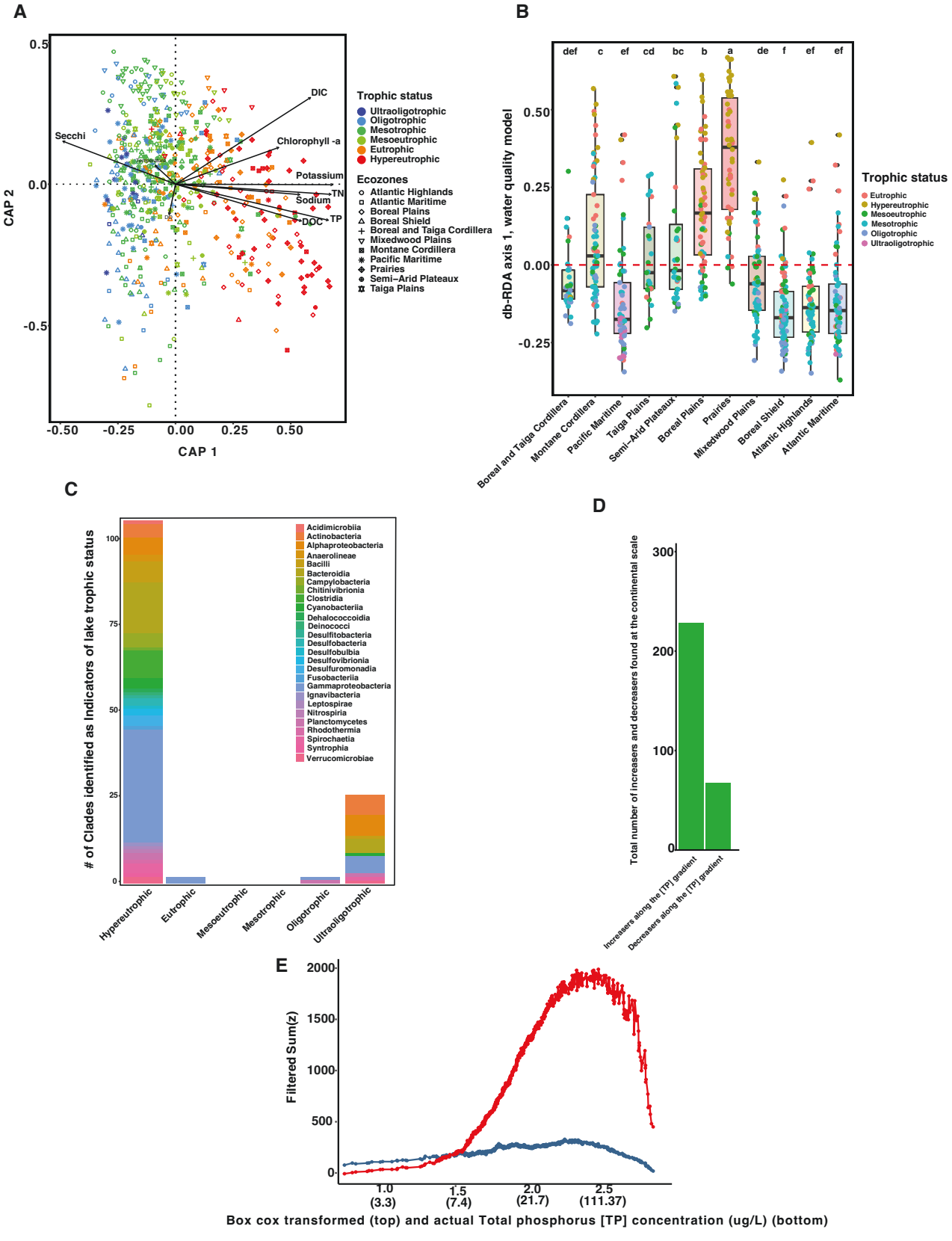

A
0.5
0.0
-0.5
CAP 2
-0.50
-0.25
0.00
0.25
0.50
CAP 1
Secchi
DIC
Chlorophyll -a
Potassium
TN
Sodium
TP
DOC
Trophic status
Ultraoligotrophic
Oligotrophic
Mesotrophic
Mesoeutrophic
Eutrophic
Hypereutrophic
Ecozones
Atlantic Highlands
Atlantic Maritime
Boreal Plains
Boreal Shield
Boreal and Taiga Cordillera
Mixedwood Plains
Montane Cordillera
Pacific Maritime
Prairies
Semi−Arid Plateaux
Taiga Plains
B
def
c
ef
cd
bc
b
a
de
f
ef
ef
0.50
0.25
0.00
-0.25
db-RDA axis 1, water quality model
Boreal and Taiga Cordillera
Montane Cordillera
Pacific Maritime
Taiga Plains
Semi−Arid Plateaux
Boreal Plains
Prairies
Mixedwood Plains
Boreal Shield
Atlantic Highlands
Atlantic Maritime
Trophic status
Eutrophic
Hypereutrophic
Mesoeutrophic
Mesotrophic
Oligotrophic
Ultraoligotrophic
C
# of Clades identified as Indicators of lake trophic status
100
75
50
25
0
Acidimicrobiia
Actinobacteria
Alphaproteobacteria
Anaerolineae
Bacilli
Bacteroidia
Campylobacteria
Chitinivibrionia
Clostridia
Cyanobacteriia
Dehalococcoidia
Deinococci
Desulfitobacteria
Desulfobacteria
Desulfobulbia
Desulfovibrionia
Desulfuromonadia
Fusobacteriia
Gammaproteobacteria
Ignavibacteria
Leptospirae
Nitrospiria
Planctomycetes
Rhodothermia
Spirochaetia
Syntrophia
Verrucomicrobiae
Hypereutrophic
Eutrophic
Mesoeutrophic
Mesotrophic
Oligotrophic
Ultraoligotrophic
D
Total number of increasers and decreasers found at the continental scale
300
200
100
0
Increasers along the [TP] gradient
Decreasers along the [TP] gradient
E
Filtered Sum(z)
2000
1500
1000
500
0
1.0
(3.3)
1.5
(7.4)
2.0
(21.7)
2.5
(111.37)
Box cox transformed (top) and actual Total phosphorus [TP] concentration (ug/L) (bottom)

Figure 3.3 *Water quality variables shaping community composition and bacterial indicators of changes. A) Distance based redundancy analysis (db-RDA) highlighting the bacterial community composition in lakes, constrained by water quality variables. Lakes are colour-coded by trophic status and shaped according to ecozone. B) Boxplot of site scores from db-RDA axis 1 grouped and colour-coded by lake trophic status, where the width of the boxes reflects sample size ranging from 25 to 88 lakes per ecozone. Letter superscripts denote significantly different ecozones determined via a post-hoc Tukey's test. C) Bar plot of bacterial indicator species associated with lake trophic status from hypereutrophic to ultraoligotrophic from indicator species analysis. D). Barplot showing the sum of increasing and decreasing bacterial taxa along a phosphorus gradient in lakes based on threshold indicator taxa analysis (TITAN). E) TITAN graph illustrating the magnitude of change for bacterial taxa in response to total phosphorus concentrations in lakes. The x-axis is the range of boxcox transformed (top) and actual values (bottom) for total phosphorus concentration in lakes. The y-axis is the sum of the z-scores of all bacterial species identified as pure and reliable indicators of either the z- group (decreasers; blue color) or the z+ group (increasers; red color).*

3.3.4 Land use influence on bacterial community structure across ecozones

We assessed the role of land use in explaining bacterial community structure using dbRDA and the fraction of agriculture, pasture, urban development, and forestry in lake watersheds as explanatory variables. The land use model was significant ($p < 0.001$), included all land use variables, and explained 6 % of the variance. In the dbRDA ordination plot, the agriculture vector was orthogonal to the urban development and pasture vectors, showing different influences of these land use types on bacterial community composition (**Figure 3.4a**). Axis 1 separated agriculturally rich lakes of the Prairies and Boreal Plains (positive axis 1 loading) from lakes of other ecozones while axis 2 correlated with highly urbanized lakes located in the Mixedwood Plains (**Figure 3.4b**).

We used TITAN to determine at which point along the agriculture and urban development gradients the largest changes occurred, and to identify the taxa associated with the change. We identified 164 increasing taxa and 69 decreasing taxa along the agricultural gradient (**Figure 3.4c**). Decreasers shifted in abundance across a wide range of agriculture, resulting in a broad peak in the sum (z-) centered near 20-40% agriculture (**Figure 3.4d**). For urban development, we identified 60 increasing taxa and 52 decreasing taxa (**Figure 3.4c**). We observed a strong filtered sum (z+) peak at very low urban development 10% (**Figure 3.4e**).

The most agriculturally influenced watersheds tended to be associated with nutrient rich lakes. Urban development was less associated with nutrient rich lakes. We therefore mapped the taxa that were responsive to TP and agriculture/urbanization gradients using Venn diagrams (**Figure 3.4fg**). The Venn diagrams illustrate the common and distinct taxa for TP and agriculture/urbanization across ecozones. Many taxa overlapped between TP concentration and agriculture (111 increasers), while no overlap was observed between TP and urbanization. The greatest number of uniquely increasing taxa (71 increasers) were found along the TP gradient with fewer unique taxa associated solely with agriculture and urbanization (**Figure 3.4f**). Similarly, decreasers (negative bacterial indicators) overlapped strongly between TP concentration and agriculture (26 decreasers) but an intersection of (8 decreasers) was found between TP concentration and agriculture. Unique negative indicators for urbanization were highest (20 decreasers) (**Figure 3.4g**).

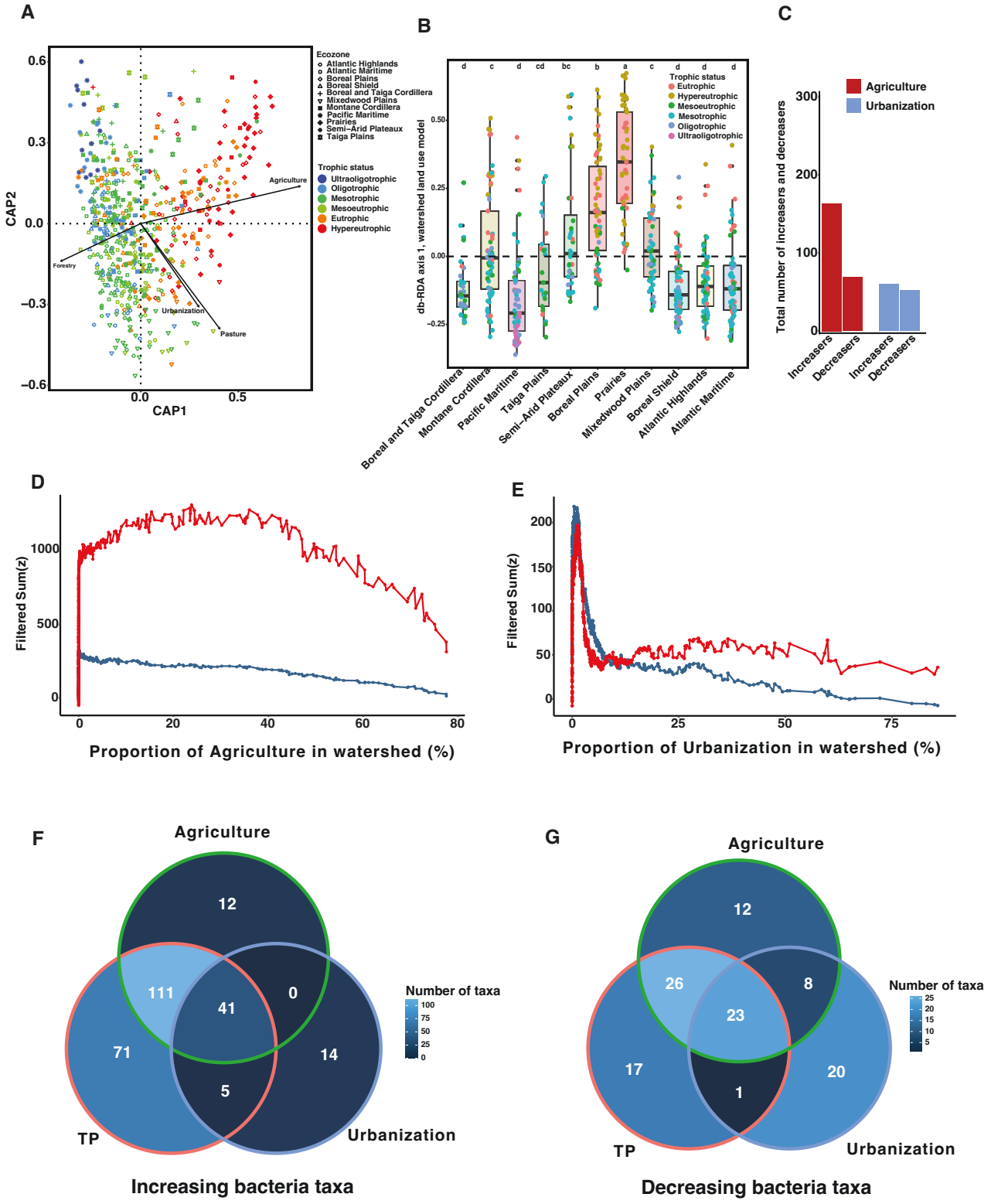


Figure 3.4 *Land use types shaping community composition and bacterial indicators of changes. A) Distance based redundancy analysis (db-RDA) highlighting the bacterial community composition in lakes, constrained*

by land use variables. Lakes are colour-coded by trophic status and shaped according to ecozone. B) Boxplot of site scores from land use model db-RDA axis 1 grouped and colour-coded by lake trophic status, where the width of the boxes reflects sample size ranging from 25 to 88 lakes per ecozone. Letter superscripts denote significantly different ecozones determined via a post-hoc Tukey's test. C) Barplot showing the sum of increasing and decreasing bacterial taxa along an agriculture and urbanization gradient in lakes based on threshold indicator taxa analysis (TITAN). D) TITAN graph highlights the magnitude of change for bacterial taxa in response to the proportion of agriculture within the watershed. The x-axis is the range for percentage agriculture within the watershed. The y-axis is the sum of the z-scores of all bacterial species identified as pure and reliable indicators of either the z- group (decreasers; blue color) or the z+ group (increasers; red color). Peaks indicate agriculture proportions at which there is lots of compositional change; plateaus indicate agriculture proportions in which there is little change in composition. E) TITAN graph highlights the magnitude of change for bacterial taxa in response to the proportion of urbanization within the watershed. The x-axis is the range for percentage urbanization within the watershed. The y-axis is the sum of the z-scores of all bacterial species identified as pure and reliable indicators of either the z- group (decreasers; blue color) or the z+ group (increasers; red color). Peaks indicate urbanization proportions at which there is lots of compositional change; plateaus indicate urbanization proportions in which there is little change in composition. F) Venn diagram representing the relationships observed among increasing bacterial taxa along water quality (total phosphorus concentrations) and land use (agriculture and urbanization) gradients at the different sets. Diagram is composed of circles that overlap, with each circle representing a set and the overlapping regions representing the intersections between sets. The purpose of this was to visually illustrate the commonalities and differences between water quality and land use bacterial indicators. G) Venn diagram representing the relationships observed among decreasing bacterial taxa along water quality (total phosphorus concentrations) and land use (agriculture and urbanization) gradients at the different sets. Diagram is composed of circles that overlap, with each circle representing a set and the overlapping regions representing the intersections between sets. The purpose of this was to visually illustrate the commonalities and differences between water quality and land use indicators.

3.3.5 Water quality influence on bacterial community structure within ecozones

Focusing our dbRDA analyses on the sets of lakes within each ecozone, we investigated how water quality influenced bacterial community structure in different regions. Water quality models were significant ($p < 0.01$) for all ecozones and explained between 14 and 28 % of the variation in community structure (**Table 3, Supplementary Figure S3.2**).

Lakes in the Prairies ecozone ranged from mesotrophic to hypereutrophic, while lakes in the Boreal Plains ecozone range from oligotrophic to hypereutrophic (**Supplementary Figure S3.2**). In both ecozones, TP and DOC had strong loadings on dbRDA axis 1, demonstrating the importance of nutrient and organic carbon in structuring bacterial communities in these regions. Furthermore, while lakes in the Montane Cordillera and Atlantic Highlands range from hypereutrophic to oligotrophic, they were predominantly of intermediate trophic states, containing fewer high nutrient lakes. Interestingly however, in both ecozones, TP and Secchi depth had strong loadings on dbRDA axis 1.

Having demonstrated the influence of nutrients on bacterial communities with dbRDAs, we sought to uncover specific TP thresholds at which changes may be occurring in each ecozone. We performed TITAN analysis in ecozones where dbRDA results showed that TP was among the strongest explanatory variables structuring bacterial communities (Prairies, Boreal Plains, (Montane Cordillera, and Atlantic Highlands). Regional TITAN analyses of the TP gradient for the Prairies and Boreal Plains ecozones produced a similar result as at the continental scale: For Prairies, a filtered sum (z+) peak at 167.4 ug/L TP (**Figure 3.5a**). For Boreal Plains, a filtered sum (z+) peak at 111.37 ug/L TP (**Figure 3.5b**). A notable difference between ecozones was observed for the filtered sum (z-) values: a strong filtered sum (z-) peak at 51.9 ug/L TP for the Boreal Plains, that was less pronounced for the Prairies. Also, across the larger trophic gradient of the Boreal Plains, there were a larger number of negative (z-) indicators than for the Prairies.

At the lower range of TP, regional TITAN analyses within the Montane Cordillera ecozone yielded different results compared to the continental scale and agriculture-rich ecozones. The filtered sum (z+) reached its peak at lower TP concentrations, ranging from 12 to 20 ug/L (**Figure 3.5c**). However, regional TITAN analyses in the Atlantic Highlands ecozone showed different responses to those observed at the continental scale and in the other examined ecozones with no filtered sum (z+) peak, but rather a broad and gradual change across TP concentrations (**Figure 3.5d**).

We constructed Venn diagrams to identify overlapping and unique indicator bacterial taxa for TP concentrations across the continental scale and for the four TP-influenced ecozones (Prairies, Boreal Plains, Montane Cordillera, and Atlantic Highlands) to uncover patterns driven by differences between the regions. We detected large number of increasing taxa distinct at the continental scale (145 increasers). An intersect between the Prairies and the continental scale showed 22 increasers most

responsive to high TP concentrations. We found an overlap of 16 increasers between the Atlantic Highland and the continental scale which were responsive to lower TP concentrations. Across ecozones, we found little regional specific taxa (i.e. increasers unique to regions) (**Figure 3.5e**). Similarly, we discovered 38 distinct negative indicators that persisted with alteration in total phosphorus concentrations at the continental scale but fewer regional specific decreasers across ecozones (**Figure 3.5f**).

Table 3 Summary of db-RDA modelling on bacterial community composition, percent variation explained by models and most influential variables are highlighted.

Ecozone	**% variation explained by water quality model**	**Significant Variables in order from forward selection step**	**ANOVA adjusted R^2 value from forward selection step**	***pvalue***	**% variation explained by land use model**	**Significant Variables in order from forward selection step**	**ANOVA adjusted R^2 value from forward selection**	***pvalue***
Boreal Plains	**20.4**	**TP**	0.073969	0.002	**4.6**	Agriculture	n/a	n/a
		Secchi	0.106369	0.002				
		DIC	0.127383	0.008				
		Na	0.151018	0.004				
		Temperature	0.173527	0.008				
		DO	0.192173	0.008				
		DOC	0.204981	0.042				
Prairies	**13.9**	Na	0.062215	0.002	**3.3**	Agriculture	n/a	n/a
		TP	0.091243	0.006				
		Chl-*a*	0.109962	0.022				
		DIC	0.124631	0.042				
		DOC	0.139911	0.034				
Mixed Wood Plains	**15.2**	Secchi	0.075327	0.002	**3.9**	Agriculture	n/a	n/a
		TP	0.100493	0.002				

Ecozone	% variation explained by water quality model	Significant Variables in order from forward selection step	ANOVA adjusted R^2 value from forward selection step	*pvalue*	% variation explained by land use model	Significant Variables in order from forward selection step	ANOVA adjusted R^2 value from forward selection	*pvalue*
		TN	0.123358	0.002				
		Chl-*a*	0.139921	0.014				
		Na	0.152163	0.032				
Pacific Maritimes	**23.5**	Secchi	0.13518	0.002	**9.9**	**Urbanization**	0.069993	0.002
		SCond	0.18519	0.002		Pasture	0.099480	0.010
		Temperature	0.23509	0.002				
Atlantic Maritime	**21.5**	DIC	0.10638	0.002	**6.3**	**Urbanization**	0.037065	0.002
		Secchi	0.14535	0.002		Forestry	0.062562	0.008
		Na	0.17133	0.002				
		Chl-*a*	0.19265	0.012				
		DOC	0.21532	0.004				
Boreal Shield	**18.5**	DIC	0.088137	0.002	**5.8**	**Urbanization**	0.022793	0.002
		Secchi	0.132129	0.002		Pasture	0.044813	0.016
		Temperature	0.148241	0.006		Forestry	0.057702	0.010
		Na	0.163597	0.004				
		Chl-*a*	0.176754	0.016				
		DOC	0.185456	0.034				
Boreal and Taiga Cordillera	**22.5**	Secchi	0.18371	0.002	**NS**	n/a	n/a	n/a
		Na	0.22485	0.012				
Montane Cordillera	**21.5**	**TP**	0.090439	0.002	**4.2**	**Urbanization**		
		SCond	0.128592	0.002				
		DOC	0.158087	0.004				
		Secchi	0.171226	0.032				

Ecozone	**% variation explained by water quality model**	**Significant Variables in order from forward selection step**	**ANOVA adjusted R^2 value from forward selection step**	***pvalue***	**% variation explained by land use model**	**Significant Variables in order from forward selection step**	**ANOVA adjusted R^2 value from forward selection**	***pvalue***
		DO	0.185310	0.018				
		DIC	0.200422	0.024				
		Temperature	0.215173	0.012				
Taiga Plains	**28.3**	DOC	0.11293	0.004	NS			
		TP	0.25944	0.002				
Semi-Arid Plateaux	**22.5**	**TN**	0.099763	0.002	**12.6**	Pasture	0.06708	0.014
		DIC	0.141188	0.006		Forestry	0.12644	0.010
		Temperature	0.169657	0.014				
		Secchi	0.200885	0.012				
		Chl-*a*	0.225410	0.034				
Atlantic Highlands	**21.6**	Secchi	0.10293	0.002	**8.9**	Agriculture	0.046905	0.002
		DIC	0.14570	0.002		Urbanization	0.089290	0.002
		TP	0.17297	0.004				
		Chl-*a*	0.19857	0.010				
		Na	0.21623	0.014				

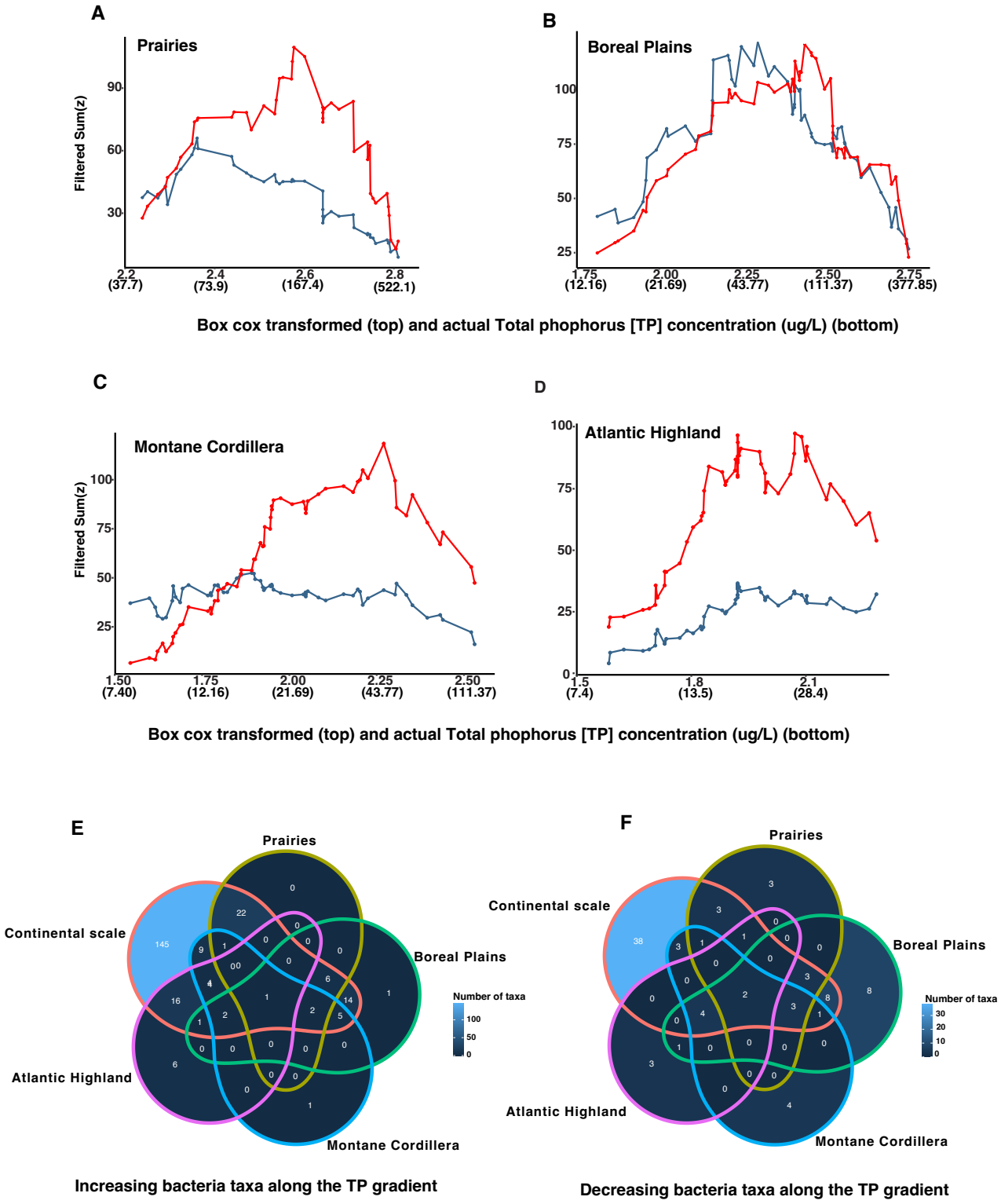


Figure 3.5 *Threshold of change in bacterial community within regions modelled by changes in water quality*

(total phosphorus concentrations) and overlaps in increasing and decreasing bacterial taxa across regions. A-D) TITAN graphs highlighting the magnitude of change for bacterial taxa in response to total phosphorus concentrations in the Prairies, Boreal Plains, Montane Cordillera, and Atlantic Highland lakes respectively. The x-axis is the range of boxcox transformed (top) and actual values (bottom) for total phosphorus concentration in lakes within the four ecozones. The y-axis is the sum of the z-scores of all bacterial species identified as pure and reliable indicators of either the z- group (decreasers; blue color) or the z+ group (increasers; red color). Peaks indicate total phosphorus concentration at which there are lots of compositional change within the four ecozones; plateaus indicate total phosphorus ranges in which there is little change in composition. E-F) Venn diagram representing the relationships observed among increasing and decreasing bacterial taxa respectively along water quality (total phosphorus concentrations) gradients at the continental scale and regionally across the Prairies, Boreal Plains, Montane Cordillera and Atlantic Highlands. Diagrams are composed of circles that overlap, with each circle representing a set and the overlapping regions representing the intersections between sets. The purpose of this was to visually illustrate the commonalities and differences between increasing and decreasing bacterial taxa at the continental scale and across ecozones.

3.3.6 Land use influence on bacterial community structure within ecozones

We assessed the importance of land use variables on bacterial community structure in different regions by focusing our dbRDA analyses on lakes within each ecozone. Land use models were significant ($p < 0.01$) for all ecozones except the Boreal and Taiga Cordillera and Taiga Plains and explained between 3 and 13 % of variation in bacterial communities (**Table 3, Supplementary Figure S3.3**). In the land use models, agriculture exhibited a substantial loading on dbRDA1, underscoring its significance in shaping bacterial communities within the Prairies, Boreal Plains, and Mixedwood Plains ecozones. Notably, agriculture stood out as the sole significant explanatory variable in these regions. In the Pacific Maritimes and Montane Cordillera, urbanization was instead the strongest significant variable explaining underlying structure in bacterial communities.

For land use, regional TITAN analyses of agriculture and urbanization gradients were restricted to ecozones where previous db-RDA analysis showed that either agriculture or urbanization had the greatest explanatory power for structuring bacterial communities. Regional TITAN analyses of agriculture was therefore performed for the Prairies (**Figure 3.6a**), Boreal Plains (**Figure 3.6b**) and Mixedwood Plains (**Figure 3.6c)** while analyses of urbanization were performed for the Pacific Maritimes (**Figure 3.6d)** and Montane Cordillera **(Figure 3.6e**).

Regional TITAN analyses of the agriculture gradient for the Prairies, Boreal Plains and Mixedwood plains ecozones produced the following results: gradual change along the agriculture gradient was identified in all agriculturally rich ecozones except the Boreal Plains where a peak occurred at low levels (10%) of agriculture. Venn diagrams illustrating commonalities and uniqueness between increasing (**Figure 3.6f**) and decreasing (**Figure 3.6g**) bacterial taxa within these three agriculture-rich ecozones revealed large number of increasing taxa distinct at the continental scale (132 increasers). An intersect between the Boreal Plains and the continental scale showed 14 increasers most responsive to increased agriculture proportions. Across ecozones, we found little regional specific taxa (i.e. increasers unique to regions). A similar trend was observed with decreasers. 46 distinct decreasing taxa were found, 10 decreasers were observed at the intersect of the Boreal Plains and the Continental scale.

Regional TITAN analyses of urbanization in the Pacific Maritimes and Montane Cordillera showed no peak and only a gradual change along the urbanization gradient for bacterial taxa. Venn diagrams illustrating commonalities and uniqueness between increasing (**Figure 3.6h**) and decreasing (**Figure 3.6i**) bacterial taxa within highly urbanized ecozones of the Pacific Maritimes and Montane Cordillera revealed high numbers of unique increasing (41) and decreasing (34) taxa at the continental scale but lower numbers of region-specific increasers and decreasers.

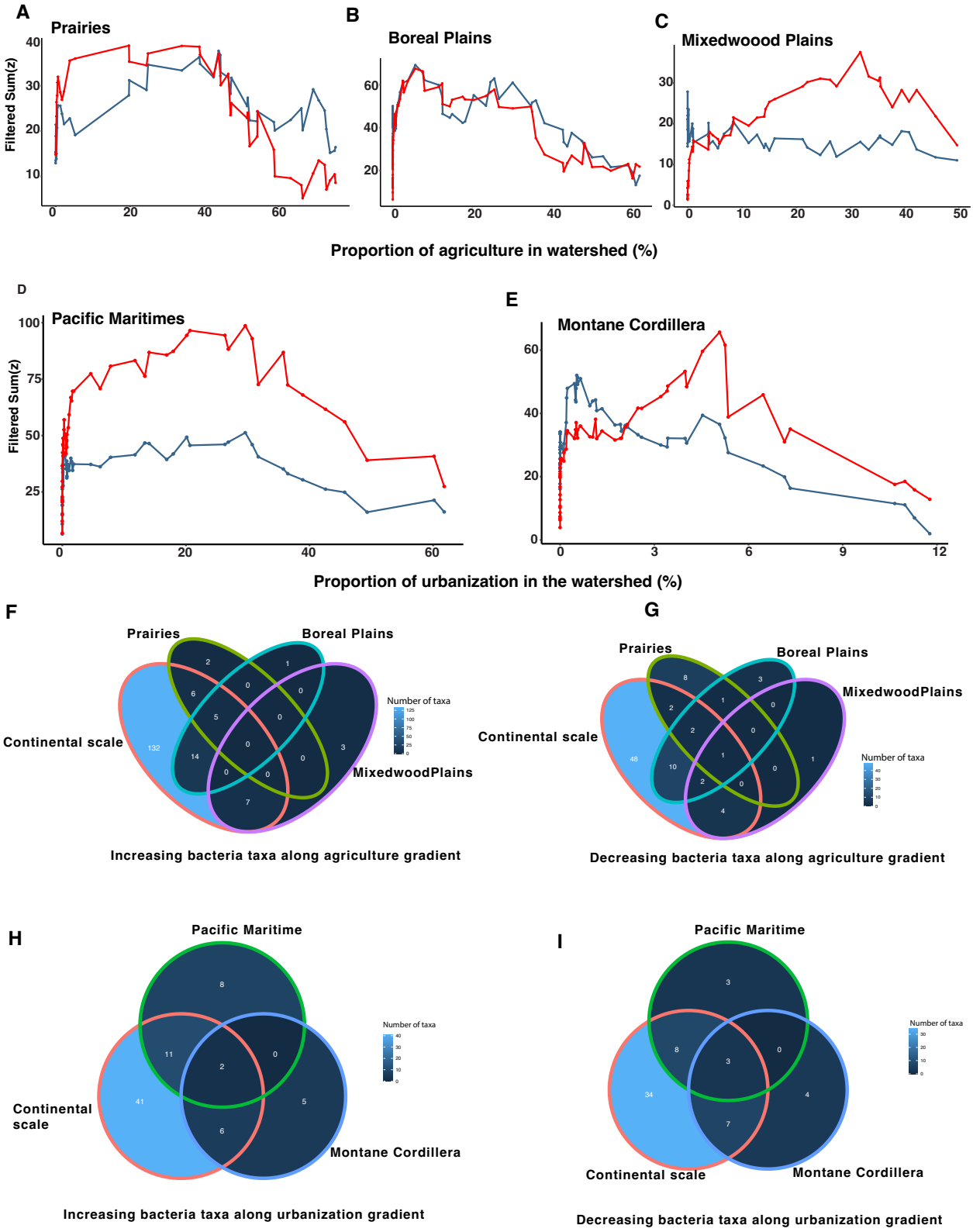


Figure 3.6 *Threshold of change in bacterial community within regions modelled by changes in land use and overlaps in increasing and decreasing bacterial taxa across regions. A-C) TITAN graphs highlighting the*

magnitude of change for bacterial taxa in response to proportions of agriculture within the watershed of lakes in the Prairies, Boreal Plains and Mixedwood Plains ecozones respectively. The x-axis is the range for percentage agriculture within the watershed for each ecozone. The y-axis is the sum of the z-scores of all bacterial species identified as pure and reliable indicators of either the z- group (decreasers; blue color) or the z+ group (increasers; red color). Peaks indicate agriculture proportions at which there is a lot of compositional change; plateaus indicate agriculture proportions in which there is little change in composition. D-E) TITAN graphs highlighting the magnitude of change for bacterial taxa in response to proportions of urbanization within the watershed of lakes in the Pacific Maritimes and Montane Cordillera ecozones respectively. The x-axis is the range for percentage urbanization within the watershed for each ecozone. The y-axis is the sum of the z-scores of all bacterial species identified as pure and reliable indicators of either the z- group (decreasers; blue color) or the z+ group (increasers; red color). Peaks indicate urbanization proportions at which there is a lot of compositional change; plateaus indicate urbanization proportions in which there is little change in composition. F-I) Venn diagrams representing the relationships observed among increasing and decreasing bacterial taxa respectively along land use gradients (agriculture and urbanization) at the continental scale and regionally across the Prairies, Boreal Plains, Mixedwood Plains, Pacific Maritimes and Montane Cordillera. Diagrams are composed of circles that overlap, with each circle representing a set and the overlapping regions representing the intersections between sets. The purpose of this was to visually illustrate the commonalities and differences between increasing and decreasing bacterial taxa at the continental scale and across ecozones.

3.4 Discussion

Although lakes occupy less than 2% of the earth's land surface (Messager et al., 2016; Verpoorter et al., 2014), their significance is paramount in global nutrient and carbon cycles (Tranvik et al., 2018). Canadian lakes, providing crucial ecosystem services, are confronting threats from anthropogenic pressures that disrupt surface water quality (Brauman et al., 2007; Dugan et al., 2017; Keeler et al., 2012). These challenges involve local changes in land use, heightened inputs of nutrients and pollutants from the watershed, and regional climate changes that can significantly disturb a lake's hydrological regime, thermal structure, and internal processes. Consequently, these alterations impact the chemical and biological properties – bacterial communities within lakes (Carpenter et al., 2007, 2011; Jeppesen et al., 2015; A. J. Reid et al., 2019). In this chapter, I analyzed 621 lakes across 12 ecozones, and observed substantial variations in environmental conditions across Canada's diverse ecozones. This work reinforces previous observations that lakes in the Prairies and Boreal Plains ecozones are most distinct from those in other ecozones in their water quality properties, especially

nutrient concentrations, and notably, they are situated within agricultural lands. Considering the intricate link between the quality and chemical properties of freshwater ecosystems, as well as of their communities, it is believed that these factors are closely connected to the distinctive features of the surrounding watershed. This chapter examines how such environmental distinctiveness influences variability in bacterial communities within regions, and across regions at the continental scale.

The relationship between water quality, as determined by nutrient concentrations, and the abundance of specific members of the microbial community has been identified (Davis et al., 2015; Hengy et al., 2017; Paver et al., 2020; Wang et al., 2022). An examination of prevailing environmental conditions revealed significant variation in water quality parameters across the continental scale in this study. Lakes within the Prairies and Boreal Plains ecozones exhibited elevated nutrient concentrations mostly in the eutrophic or hypereutrophic categories, in contrast to other ecozones that demonstrated intermediate (mesotrophic and mesoeutrophic) and low (oligotrophic and ultraoligotrophic) trophic states.

Bacterial diversity analysis in this chapter unveiled regional differences, with higher richness in the nutrient-rich lakes of the Prairies and the Boreal Plains. Indicator species analysis highlighted a greater number of indicators associated with hypereutrophic and eutrophic lakes. Additionally, a connection between water quality and land use was evident from the substantial number of shared bacterial indicator taxa between communities changing along gradients of TP and agricultural land use. Moreover, investigations identified the primary drivers of bacterial communities across the continent as either agriculture, or urbanization. Restricting the analyses to within regional ecozones revealed agriculture as the sole driver of community composition in nutrient-rich ecozones of the Prairies, Boreal Plains, and Mixedwood Plains. In contrast, urbanization was more prominent as a driver of bacterial community composition in the Montane Cordillera and Pacific Maritimes ecozones.

3.4.1 Environmental heterogeneity explains regional variation in patterns of bacterial diversity

Hotspots for high TP concentration and extensive agriculture, identified within the Prairies and Boreal Plains exhibited higher-than-average concentrations of TP, TN, major ions, dissolved carbon, chlorophyll-*a* concentrations, and trophic status (inferred from TP concentrations). These differences likely played a significant role in overall variation in the continental-scale dataset but importantly, these variations were primarily regionally driven. Studies of Canadian Prairies lakes have shown that high nutrient concentrations are associated with terrestrial land use, specifically agriculture (Dupont et al., 2023). In this dataset, the substantial differences in water quality observed in the Prairies and Boreal Plains ecozones are likely a result of a higher local prevalence of agriculture (median proportion of agriculture = 38.3 and 22.5% in the Prairies and Boreal Plains respectively) and variations in baseline nutrient conditions. Specifically, the median total phosphorus (TP) levels indicate significant distinctions: Prairies TP median = 338.0 µg/L, Boreal Plains TP median = 144.0 µg/L, compared to the national TP median of 18.2 µg/L (Taranu & Gregory-Eaves, 2008). Research suggests that the regional disparity in how lakes respond to land use within the surrounding terrestrial watershed may be influenced by various contributing factors. The region in focus possesses a flat catchment topography, with naturally nutrient-rich geology (Griffiths et al., 2021; Taranu & Gregory-Eaves, 2008).

Regarding urbanizing land uses, this study revealed that lakes in the Pacific Maritimes displayed lower nutrient concentrations despite highest local urbanization among all ecozones (18.2%). Supporting a population of two and a half million people, this ecozone experiences ongoing rapid population growth. Particularly noteworthy is the median total phosphorus (TP) level, showing a significant decrease compared to the earlier findings in the Prairies, with a TP median of 18.6 µg/L. Interestingly, this mirrors baseline nutrient conditions. Clearly, our analysis of lake environmental conditions across the nation highlights substantial variations on a continental scale, predominantly influenced by regional conditions, and impacted by terrestrial land use factors like agriculture and urbanization. Such extensive regional environmental variations are likely eliciting diverse responses from lake bacterial communities as would be revealed in this chapter.

3.4.2 Trophic status drives bacterial diversity and community composition at the continental scale

Continental scale studies revealed a widespread influence of water quality – specifically (TP, TN, chlorophyll-*a* concentrations, DIC, DOC and cations) as well as watershed land use type on bacterial diversity and community composition. We identified an intricate relationship wherein these water quality parameters were nested within human land use impacted lands. Variation in bacterial richness was equally strongly predicted by TP concentrations (mean random forest analysis prediction 18%). While the direction (increase or decrease) of variation in Chao-1 richness estimates across lakes was not further predicted in this chapter, one study reported a decline in bacterial richness associated with higher nutrient concentrations in eutrophic lakes (Ji et al., 2018). In contrast, my study pinpointed patterns deviating from this, as bacterial richness was notably highest within certain regions along the continental scale, and comparatively lower in others depicting ecozone specific differences. Similar trends have been documented in other studies. In a comparison of six lakes in Canada and India, Obieze et al., (2022) found that Lake Winnipeg, a eutrophic lake, exhibited increased microbial richness. Also, on examining determinants of bacterioplankton richness in Swedish lakes, researchers concluded a positive relationship between bacterioplankton richness and nutrient availability. They noted a significant increase in richness with rising nutrient levels, indicating a dependency between richness and productivity (Logue et al., 2012). In other Canadian lakes, increased bacterial richness was reported in tandem with lake trophic state and vertical heterogeneity in nutrients (Jankowski et al., 2014). This spatial resource variation could prompt habitat-specific responses of bacteria to elevated nutrient concentrations, indicative of lake productivity. Generally, the observed increased bacterial richness with higher nutrient concentrations aligns with our findings, potentially attributable to the presence of fast-growing copiotrophic bacteria thriving in nutrient-rich lakes.

A high number of bacterial indicators (106 taxa) were associated with hypereutrophic conditions but surprisingly, no taxa was associated with mesotrophic conditions. Among the hypereutrophic indicators, the most abundant clades belonged to the Proteobacteria (adapted to some level of nutrient overloading), Bacteroidota (proficient in the degradation of complex biopolymers and dissolved organic matter) (Newton et al., 2011), Firmicutes (possessing diverse metabolic capacities) (Martiny et al., 2006), and Desulfobacterota (sulfate-reducing bacteria) groups. Furthermore, five clades belonged to the Actinobacteriota group, known to be sensitive to high nutrient loadings but intriguingly, only three distinct species of Cyanobacteria were recognized as

indicators of hypereutrophic conditions – *Microcystis*, *Dolichospermum* and *Nodularia*. Apart from *Nodularia*, the remaining species were recently observed to have significant abundances in Canadian lakes. *Microcystis* concentrations exhibited variability by ecozone, but majority of all detections (67%) concentrated in the Prairies and Boreal Plains. This distribution also correlated with the trophic state, frequently occurring in lakes characterized as eutrophic or hypereutrophic, as detailed in the study by MacKeigan et al., 2023.

A lesser number of bacterial taxa were found under ultraoligotrophic conditions (26 taxa), primarily consisting of organisms from Proteobacteria (11 clades) and the second most abundant group was Actinobacteria (6 clades), and Bacteriodiota (4 clades). Other groups such as Verrucomicrobia, Spirochaetota, Cyanobacteria, Firmicutes, and Planctomycetota were each represented by one taxon. Importantly, Actinobacteriota are typically members of oligotrophic lakes, and we have evidence that supports this here as Actinobacteriota taxa were associated with ultraoligotrophic conditions.

3.4.3 Continental-scale patterns indicate a connection between lake trophic status and the extent of agricultural land use in the watershed

We identified bacterial taxa demonstrating a threshold response to changes in the proportion of agriculture and urbanization within the watershed. This helped identify change points in bacterial community composition along land use gradients. A broad peak was observed for agriculture gradients, but the peak was narrower with urbanization. Intriguingly, more bacterial taxa increased than decreased along the agriculture gradient, mirroring observations made along the phosphorus concentration gradient. In contrast, bacterial taxa associated with urbanization first exhibited a peak at very low levels of urbanization after which a steady increase or decrease was observed.

Upon examining specific taxa exhibiting increases or decreases along TP concentration, agriculture, and urbanization at the continental scale, we noted a substantial overlap of increasing taxa between TP and agriculture. However, no shared increasers were identified between TP and urbanization. The shared increasers were prominently represented by organisms within Proteobacteria (43 clades; 11 Alphaproteobacteria, 32 Gammaproteobacteria), Bacteroidota (16 clades), Firmicutes (14 clades), Desulfobacterota (7 clades), and Cyanobacteria (5 clades). This observed pattern concurs

with findings from our ISA and implies a potential connection to non-point source agricultural runoff from the surrounding watershed (Numberger et al., 2022; Obieze et al., 2022; Oliva et al., 2023). This finding further underscores the correlation between phosphorus concentration in lakes and agricultural land use.

The observed patterns in bacterial diversity and community composition at the continental scale are notably influenced by regional forcings within the Prairies and the Boreal Plains ecozones. However, it is essential to consider that these distinctive patterns within these ecozones may be introducing a bias in the continental-scale analysis. Exploring these patterns at regional scales holds the potential to unveil additional nuanced insights. The concentration of phosphorus in lakes is often linked to agricultural runoffs (K. Reid & Schneider, 2019) and is implicated as the dominant source in some of the most heavily impacted waters (Bunting et al., 2016). Land use associated with agriculture could be favouring higher immigration of soil species into lakes from runoff.

3.4.4 Agriculture and urbanization as region-specific drivers of community structure

Analysis within ecozone regions resolved that patterns of bacterial diversity and composition identified at the continental scale were driven by regional variation in the dataset. We identified four ecozones that are hotspots for elevated TP concentrations in lakes (Prairies, Boreal Plains, Atlantic Highlands and Montane Cordillera); two out of these ecozones were similarly hotspots for agricultural land use (Prairies, Boreal Plains and Mixedwood Plains). On the other hand, we identified predominantly urbanized ecozones as the Pacific Maritimes and Montane Cordillera (a high TP hotspot). Our investigations showed a variation in regional drivers of community composition based on prevalent land use type within the watershed as well as TP concentrations, similar to trends observed by Garner et al., 2023. In agricultural regions, increasing bacterial taxa were evident at the intersection of the continental scale with the Prairies and the Boreal Plains. Predominantly, the Boreal Plains exhibited overlaps characterized by organisms associated with elevated nutrient concentrations. This observation aligns with previous taxonomic profiling of anthropogenically impacted lakes (Fournier et al., 2021; Kraemer et al., 2020; Yang et al., 2020).

In urbanized environments however, shared taxa were confined to the Pacific Maritimes. An Actinobacteria taxon, indicative of oligotrophic conditions (Newton et al., 2011), was exclusive to this intersection and absent in all other scenarios. This distinction underscores the fact that bacterial taxa responding to high trophic states exhibit similarities with those in agricultural lands but may

diverge completely in urbanized contexts. In earlier studies conducted on urbanized Malaysian lakes, researchers found that urban land use significantly influenced the physicochemical properties and microbial dynamics of the water systems. These investigations revealed the coexistence of transient microbial communities alongside cosmopolitan communities (Ting et al., 2021). Similarly, in Northeast Germany, water samples were observed to harbor highly habitat-specific bacterial communities, with several genera exhibiting distinct urban signatures (Numberger et al., 2022).

3.5 Conclusion

Lakes offer valuable ecosystem services (Sterner et al., 2020), including provisioning (e.g., fisheries and hydroelectric power generation), regulating (e.g., local climate regulation), supporting (such as primary production, nutrient cycling, and ecosystem resilience), and cultural services (Aylward, 2005). Canada stands as the most lake-rich country globally, boasting over 900,000 lakes that intricately shape its landscape (Minns et al., 2008). Notably, over 90% of Canadians rely on lakes and rivers for their drinking water needs (Huot et al., 2019). This underscores the importance of our study in a region renowned for having the highest concentration of lakes on a global scale. This chapter highlights the intricate interconnections in water quality, land use, and bacterial communities within Canadian lakes. The multifaceted findings stress the necessity of acknowledging regional variations and discerning the unique influences exerted by agriculture and urbanization on bacterial diversity and composition. However, the study also advocates for more extensive regional analyses to delve deeper into localized patterns and influences in Canadian lake ecosystems. Moreover, it emphasizes the potential benefits of delving beyond taxonomic profiling to investigate the functional capabilities of diverse bacterial groups in response to changes in water quality and land use, enriching our understanding of these complex ecological dynamics.

3.6 Methods

3.6.1 Lake selection and sampling

As an integral part of the NSERC Canadian Lake Pulse Network (Huot et al., 2019), our research focused on 621 lakes, each sampled once across three consecutive summers (2017 - 2019). This endeavor adhered to a meticulous and standardized protocol outlined by the (NSERC Canadian Lake Pulse Network, 2021). The lakes, chosen deliberately, spanned twelve ecozones—distinct regions defined by unique climate, geology, and vegetation parameters (CCEA, 2016). The selection process employed a stratified random block design, strategically considering three different lake sizes (0.1–1 km2, 1–10 km2, 10–100 km2), and factoring in three watershed human impact categories (low, medium, high) as stratification groups (Huot et al., 2019). A prerequisite for inclusion was that the lakes possess a minimum depth of 1 m and be conveniently located within 1 km of road access. To mitigate the impact of seasonal variations, our sampling efforts were concentrated between the end of June and the commencement of September, aligning with the period of maximal thermal lake stratification.

Integrated surface water samples, spanning the euphotic zone up to 2 meters below the surface, were collected using an integrated tube sampler at the site of maximum lake depth. The euphotic zone, determined as twice the Secchi disk depth, guided the sampling approach. For full details of the field protocol, see the NSERC Canadian Lake Pulse Network Field Manual (2021). Collected samples were promptly stored in chilled coolers and, on the same day, filtered onshore. Pre-filtration through 100 μm nylon mesh preceded vacuum filtration through 0.22 μm Durapore membranes in glass funnels, maintaining a maximum pressure of 8 in Hg. Filtration concluded either at 500 mL or upon clogging of the filter. Strict cleanliness protocols were adhered to, including acid-washing, and rinsing all sampling equipment with lake water before use. Post-filtration, filters were stored in sterile cryovials at -80 °C for preservation.

3.6.2 DNA extraction, amplification and sequencing of the bacterial 16S rRNA gene

Bacterial diversity was assessed through the sequencing of 16S rRNA gene fragments, which were amplified from DNA obtained in 0.22 – 100 μm surface water samples. DNA extraction from filters utilized PowerWater kits (Mobio Technologies Inc., Vancouver, Canada), adhering to the manufacturer's protocol, including an optional Step 7 involving the addition of 1 μL ribonuclease A

followed by incubation at 37 °C for 30 minutes, and elution into 50 µl of buffer. For the amplification of a ~300 bp fragment of the 16S rRNA gene V4 region, the primer set 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'GGACTACHVGGGTWTCTAAT-3') was employed under specific conditions. The PCR protocol comprised 5 µl Phusion High Fidelity Buffer, 0.5 µl dNTPs (10 mM), 1.8 µl of each primer (5 µM), 0.25 µl Phusion polymerase, 13.65 µl ddH2O, and 2 µl of DNA. The PCR conditions involved an initial denaturation at 98˚C for 30 seconds, followed by 22 cycles of 98˚C for 20 seconds, 54˚C for 35 seconds, 72˚C for 30 seconds, and a final elongation at 72˚C for one minute. All pre-PCR DNA dilutions and liquid transfers were conducted under positive pressure in a UV cabinet. Subsequently, PCR products were electrophoresed in a 1.5% agarose gel at 80 V for 60 minutes, and samples were sequenced on an Illumina MiSeq machine across three separate sequencing runs.

Further details of PCR step 2 and purification involve pooling products from four reactions per sample, cleaning with the Zymo Research DNA purification kit (Zymo Research, Irvine, USA), and eluting into a volume of 30 µl. Barcodes and Illumina adaptors were then added in a second PCR reaction under specified conditions. Post-second PCR, the products underwent purification using the AMPure kit (Beckman Coulter Diagnostics, Montreal, Canada), with DNA concentrations measured using a nanodrop. Pooled samples, diluted to 10 nM, were sequenced using an Illumina MiSeq machine across three total runs. Each sequencing plate incorporated two negative controls (ddH2O) and one DNA sample of a mock community for sequencing quality control.

3.6.3 Processing of sequence data

Primer sequences were removed in Cutadapt v. 3.1 (Martin, 2011). Trimmed reads were processed into ASVs in R through DADA2 v. 1.16 (Callahan et al., 2016). The DADA2 pipeline consisted of trimming low-quality end positions, inferring denoised ASVs based on learned sequencing error rates, merging paired forward and reverse reads, eliminating chimaeras, and assigning taxonomy. Samples were pooled for ASV inference using otherwise default parameters. Taxonomy was assigned in TaxAss which classified 16S rRNA gene sequences using both the curated freshwater FreshTrain v. 2020/06/15 (specific for freshwater bacteria) and SILVA v. 138 reference databases (Rohwer et al., 2018). ASVs were aligned in SINA v. 1.7.2 (Pruesse et al., 2012) against the SILVA 138.1 SSU Ref NR 99 rRNA gene database (released August 27, 2020) (Quast et al., 2013). Positions outside a defined range were trimmed off, and ASVs with fragment lengths under

250 bp or over 260 bp were removed in R. Samples representing negative controls (sequencing blanks) and mock communities were removed. To create a dataset of bacterial assemblages, ASVs not assigned at the kingdom rank and ASVs assigned to archaea, eukaryotes, and chloroplasts were removed. Finally, samples containing fewer than 10000 sequences were removed, saline and experimental lakes were also excluded from this analysis, resulting in a final dataset of 621 freshwater lake ASV assemblages.

3.6.4 Estimation of bacterial diversity

ASV composition was randomly subsampled (i.e., "rarefied") to an equal sampling depth of 10,321 sequences specifically for the estimation of α-diversity indices. 899 clades were inferred from ASVs. Bacterial richness (represented by the Chao1 richness index) computed in the R package *vegan* (Oksanen et al. 2020).

3.6.5 Processing of water quality and land use data

Water quality variables included Secchi depth, dissolved oxygen, specific conductance (SC), potassium, sodium, sulphate, total P, total N, chlorophyll *a*, DOC, and DIC. Lake trophic state was categorized by TP concentrations according to the Canadian Water Quality guidelines: ultraoligotrophic (<4 μg/L), oligotrophic (4 – 10 μg/L), mesotrophic (10 – 20 μg/L), mesoeutrophic (20 – 35 μg/L), eutrophic (35 – 100 μg/L), and hypereutrophic (>100 μg/L) (CCME, 2004).

Land use information pertaining to each lake's watershed was collated from diverse sources, including the Annual Space-Based Crop Inventory for Canada 2016 (GOC 2017) and the Land Use 2010 database (GOC 2015a). These data underwent analyses employing geomatic software, as outlined in Huot et al. 2019. The watershed for each lake was delineated at a 30 m^2 pixel size, utilizing flow directions calculated with the Canadian Digital Elevation Model (GOC 2015b).

Land use types such as agriculture, pasture, urban, and forestry were extracted for each pixel from various public databases. An aggregate watershed human impact metric was then derived based on this land use data, with further details available in Huot et al. 2019. In instances where water quality and lake physical variable data were absent, they were substituted with ecozone median values. Mapping procedures were conducted in R, utilizing the NAD 83 coordinate reference system. The

coordinates of Canada were derived from the package *maps* (Becker, 2018), and ecozone shape files were sourced from the Canada Council of Ecological Areas (Wiken, 1996).

3.6.6 Statistical analyses

Data wrangling and statistical analyses were executed in R version 4.2.1 (R Core Team, 2022). Variables not conforming to the normality assumption, determined via the Shapiro–Wilks test ($p > 0.05$), underwent a Box Cox transformation. The R package *geoR* (Ribeiro et al., 2020) was employed for computing Box Cox transformations, following the methodology outlined in Schacht et al., 2023. Importantly, for the threshold indicator taxa analysis (TITAN), raw land use values were utilized without undergoing transformations.

3.6.7 Environmental ordination analysis

Environmental principal component analysis (PCA) was conducted using the *rda* function of the R package *vegan* (Oksanen et al., 2020). All water quality response variables underwent transformation and standardization (scaled to zero mean and unit variance) before performing the PCA. Land use variables were passively fitted onto the PCA to depict their association with water quality parameters. Given that most of the variation was explained by PC axis 1, an analysis of variance (ANOVA) approach was applied to quantify the observed variability in axis scores across ecozones, with post-hoc Tukey's tests used to identify significant differences in ecozones.

3.6.8 Random Forest (RF) analyses

RF analysis was employed to assess the impact of variables on bacterial richness. RF presents advantages over traditional regression techniques, particularly in mitigating the risk of overfitting when dealing with a large number of predictor variables, as is the case in our study (Matsuki et al., 2016; Ryo & Rillig, 2017). In this context, we utilized an RF technique based on conditional inference regression trees (Strobl et al., 2009), as developed by Ryo and Rillig (2017). Importance measures were computed for each predictor variable through cross-validation, utilizing data not employed in the tree construction, known as the out-of-bag (OOB) data (Breiman, 2001). The analysis involved the utilization of 5000 regression trees to ensure a robust prediction, implemented using the *party* package in R (Horton et al., 2019; Ryo & Rillig, 2017; Strobl et al., 2007; Zeileis et al., 2008).

3.6.9 Analysis relating to bacterial community composition

Principal Coordinates Analysis (PCoA) was employed to investigate and visualize dissimilarities in bacterial community data. A dissimilarity matrix was calculated using Bray-Curtis distance, and each item was assigned a location in a low-dimensional space using the R *vegan* package. Distance-based redundancy analysis (dbRDA) allows constrained ordinations on community data using non-Euclidean distance measures.

Two categories of db-RDA models were established: a water quality model encompassing surface water temperature, chlorophyll-*a* concentrations, specific conductance, dissolved oxygen, dissolved inorganic carbon, total nitrogen, total phosphorus, secchi depth, and potassium; and a land use model comprising proportions of agriculture, urbanization, pasture, and forestry in the watershed. A synthetic distance matrix was computed from bacterial community data using the *vegdist* function in R, employing Bray-Curtis as the distance method to capture dissimilarity in community composition. Subsequently, a forward selection distance-based redundancy analysis was performed to evaluate how water quality and land use influence community composition. The *capscale* function of the R *vegan* package (Oksanen et al., 2020) was utilized for this analysis. Explanatory variables were sequentially added to the models to assess their explanatory power within the response matrix.

3.6.10 Indicator Species analysis (ISA)

Indicator Value Species Analysis (IndVal) was conducted using *multipatt* (IndVal.g function) (Dufrêne & Legendre, 1997) implemented in the package *indicspecies* (ver. 1.7.14) (Cáceres & Legendre, 2009; De Cáceres et al., 2010) to assess bacterial taxa associations with trophic states. Indicator species were identified based on the product of their relative abundance and their occurrence in different sites within a sector. Values approaching 1 indicate the strongest predictive value for trophic states.

3.6.11 Threshold Indicator Taxa Analysis (TITAN)

For the TITAN analysis, we selected taxa with more than three occurrences, to meet TITAN's minimum taxon frequency criterion. The analysis involved 100 repetitions and 100 bootstraps, with purity and reliability thresholds set at 0.95. Results were visualized by plotting cumulative threshold frequencies for taxa displaying either an increase or decrease in abundance along the gradients of total

phosphorus concentration, agriculture, and urbanization. Total phosphorus concentration values were boxcox transformed but raw land use values were used. The *TITAN2* package in R, developed by Baker et al. (2019), facilitated these analytical procedures. TITAN (Baker & King, 2010) is a valuable tool for detecting non-linear responses of bacterial taxa to environmental gradients, such as water quality and land use. It identifies abrupt changes in taxa distribution along specific points of the gradient. TITAN utilizes indicator species analysis (IndVal) (Dufrêne and Legendre, 1997) to combine abundance and occurrence data, providing insights into associations between taxa and environmental conditions.

The process involves standardizing observed IndVals as z-scores, indicating negative (z-) and positive (z+) distribution changes in response to environmental gradient variations. By summing z-scores for each species, TITAN detects change points in assemblage composition. The methodology employs bootstrap resampling to estimate confidence limits for various parameters and taxon-specific change points.

Two key diagnostic indices, purity and reliability, are provided for each taxon. Purity assesses the consistency in response direction, while reliability evaluates the frequency of strong response magnitude. TITAN's thorough approach, incorporating resampling techniques and diagnostic indices, enhances its effectiveness in revealing intricate relationships between bacterial taxa and environmental gradients.

3.6.12 Venn Diagrams

Venn diagrams were generated to illustrate the overlap of indicators identified through the TITAN analysis. Positive indicators (increasers) and negative indicators (decreasers) were employed separately to construct Venn diagrams, forming sets based on water quality and land use variables at the continental scale. Additionally, sets were created to compare environmental variables regionally. The *ggVennDiagram* package (ver. 1.2.3) in R, developed by Gao et al. (2021), was utilized for computation and visualization of intersections (Gao et al., 2021).

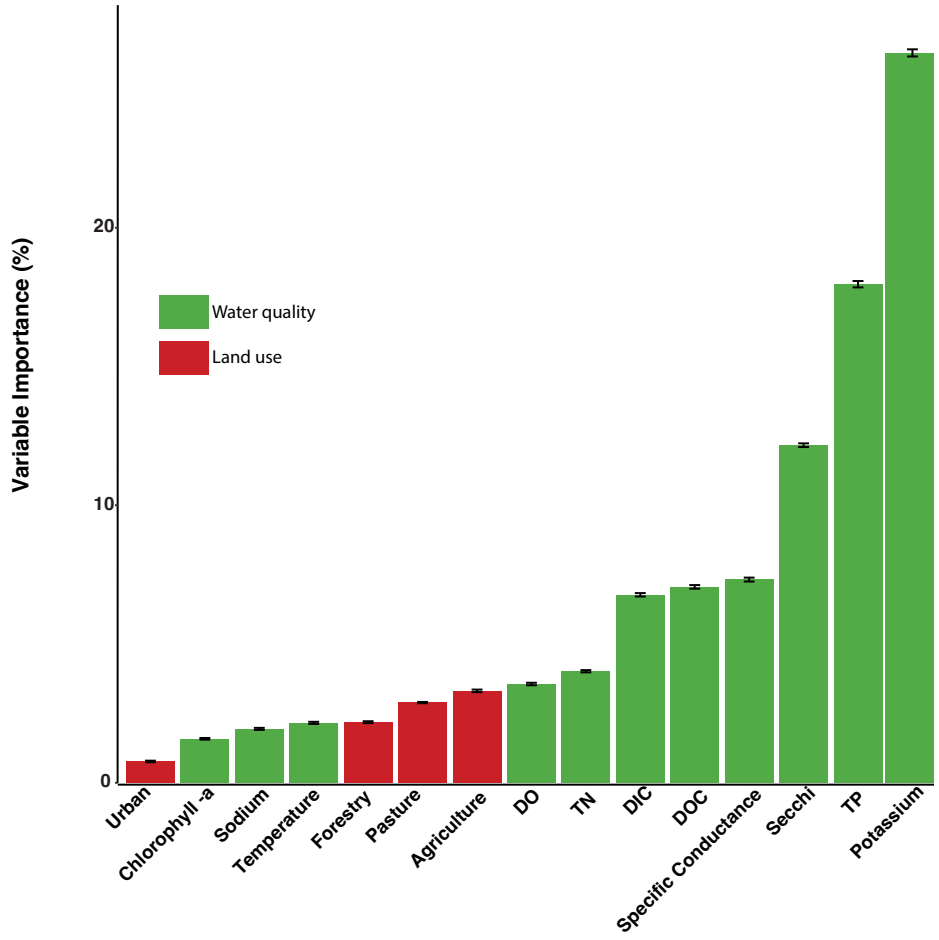
3.7 Data availability

Sequence data have been deposited in the European Nucleotide Archive under study accession PRJEB47327 (www.ebi.ac.uk).

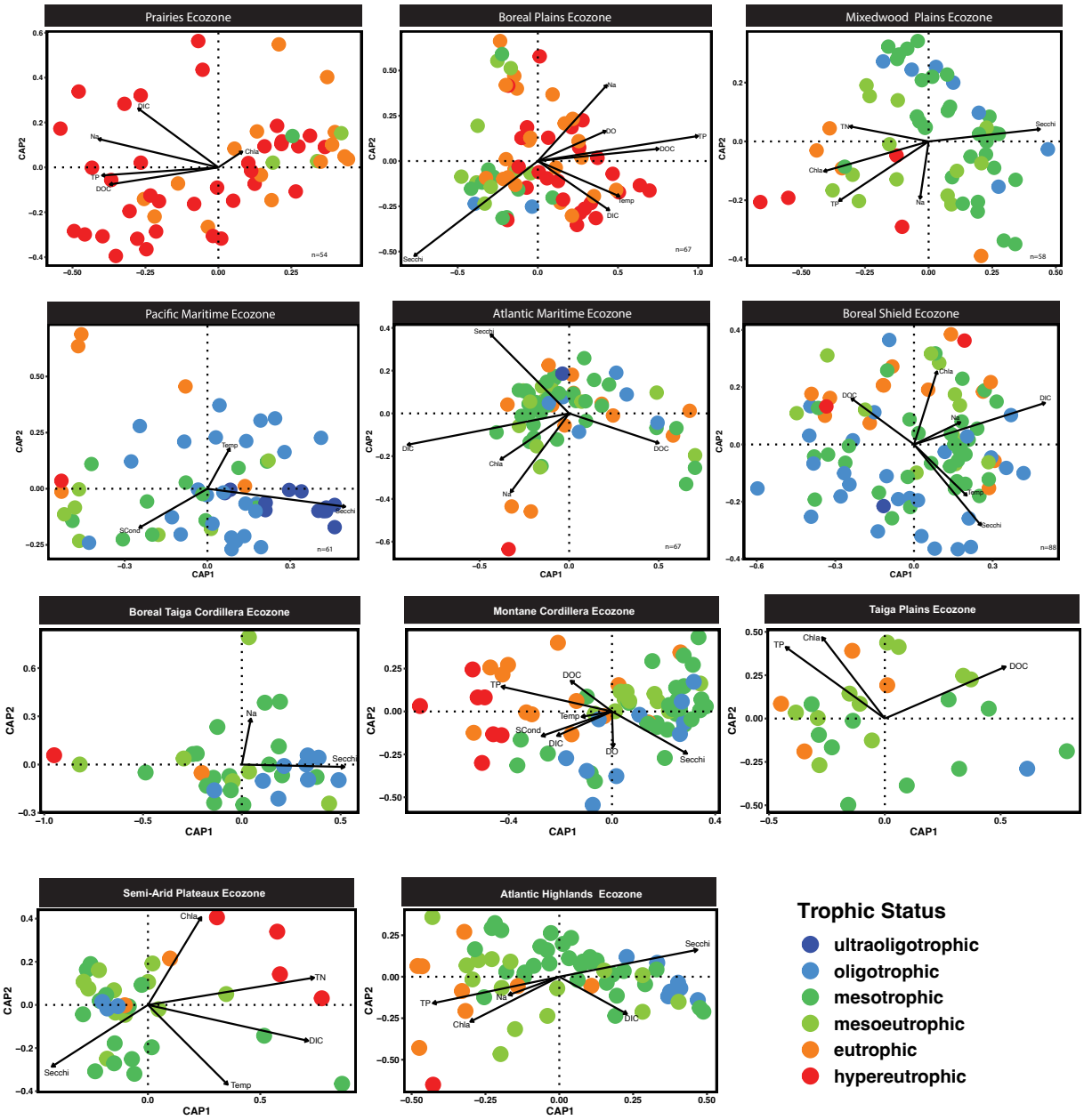
3.8 Funding and acknowledgement

This chapter was funded by the NSERC Canadian LakePulse Network (Strategic Network Grant NETGP-479720). I received scholarship from the NSERC CREATE ÉcoLac Training Program in Lake and Fluvial Ecology. I appreciate the LakePulse researchers involved in the generation, management, and quality control of the data set. Thanks to Genevieve Bourette, UQAM for support with illumina sequencing and Wentworth Brookes and Shawn Simpson, for server usage and bioinformatics technical support.

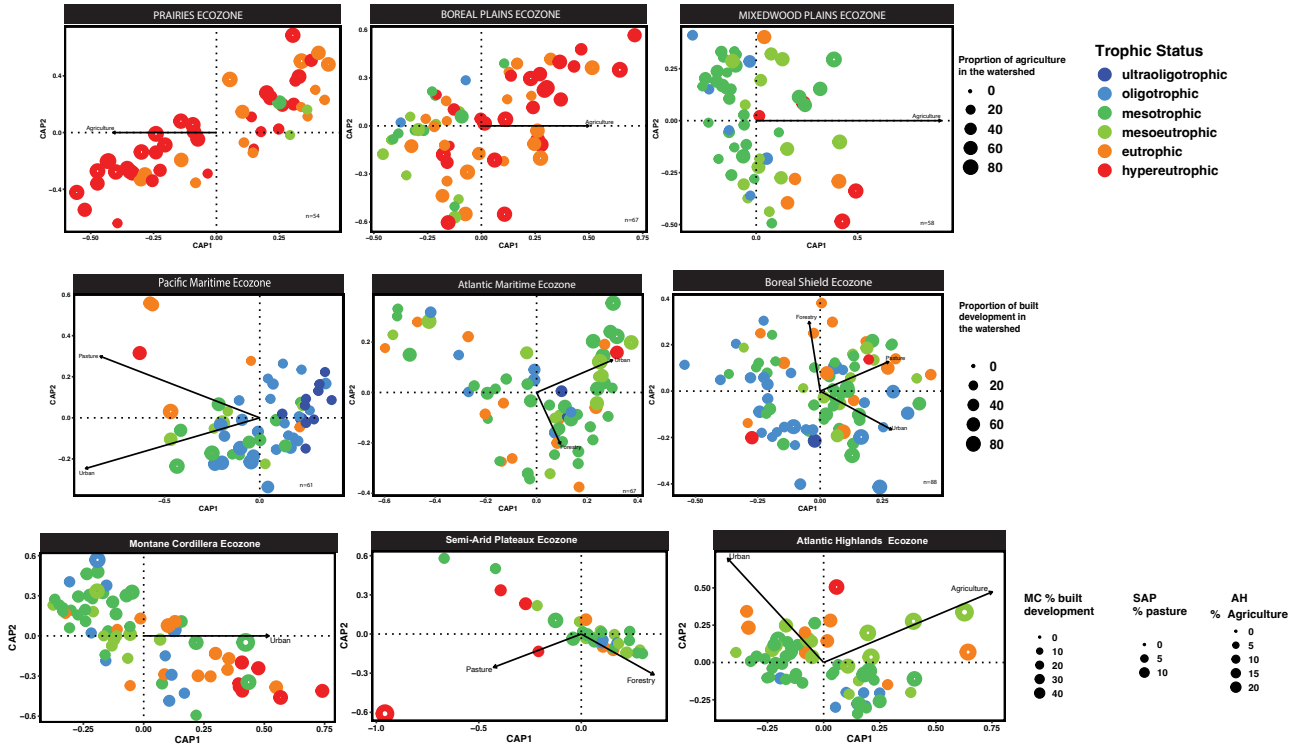
3.9 Supplementary Figures



Supplementary Figure S3.1 *Random Forest analysis showing the top predictors of bacterial richness.*



Supplementary Figure S3.2 *db-RDA analyses on the sets of lakes within each ecozone, showing the influence of water quality on bacterial community structure in different regions.*



Supplementary Figure S3.3 *db-RDA analyses on the sets of lakes within each ecozone, showing the influence of land use on bacterial community structure in different regions.*

Bibliography

Ahmed, S. F., Kumar, P. S., Kabir, M., Zuhara, F. T., Mehjabin, A., Tasannum, N., Hoang, A. T., Kabir, Z., & Mofijur, M. (2022). Threats, challenges and sustainable conservation strategies for freshwater biodiversity. *Environmental Research*, *214*, 113808. https://doi.org/10.1016/j.envres.2022.113808

Aylward, B. (2005). Freshwater ecosystem services. *Ecosystems and Human Well-Being: Policy Responses*, *3*, 213–255.

Baker M, King R, Kahle D (2023). _TITAN2: Threshold Indicator Taxa Analysis_. R package version 2.4.2, https://CRAN.R-project.org/package=TITAN2.

Baker, M. E., & King, R. S. (2010). A new method for detecting and interpreting biodiversity and ecological community thresholds. *Methods in Ecology and Evolution*, *1*(1), 25–37. https://doi.org/10.1111/j.2041-210X.2009.00007.x

Becker RA, Wilks AR, Brownrigg R, Minka TP, Deckmyn A. 2018. maps: draw geographical maps. CRAN.

Bock, C., Jensen, M., Forster, D., Marks, S., Nuy, J., Psenner, R., Beisser, D., & Boenigk, J. (2020). Factors shaping community patterns of protists and bacteria on a European scale. *Environmental Microbiology*, *22*(6), 2243–2260. https://doi.org/10.1111/1462-2920.14992

Bouchard, I & Shiab, N. (2022). A Study of Urban Sprawl and Commuting Modes in Canadian Metropolitan Areas. https://ici.radio-canada.ca/info/codesource/code-ouvert/2022/03/etalement-urbain/analysis.nb.html#Context

Brauman, K. A., Daily, G. C., Duarte, T. K., & Mooney, H. A. (2007). The Nature and Value of Ecosystem Services: An Overview Highlighting Hydrologic Services. *Annual Review of Environment and Resources*, *32*(1), 67–98. https://doi.org/10.1146/annurev.energy.32.031306.102758

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32. https://doi-org.lib-ezproxy.concordia.ca/10.1023/A:1010933404324

Bunting, L., Leavitt, P. r., Simpson, G. l., Wissel, B., Laird, K. r., Cumming, B. f., St. Amand, A., & Engstrom, D. r. (2016). Increased variability and sudden ecosystem state change in Lake Winnipeg, Canada, caused by 20th century agriculture. *Limnology and Oceanography*, *61*(6), 2090–2107. https://doi.org/10.1002/lno.10355

Cáceres, M. D., & Legendre, P. (2009). Associations between species and groups of sites: Indices and statistical inference. *Ecology*, *90*(12), 3566–3574. https://doi.org/10.1890/08-1823.1

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), Article 7. https://doi.org/10.1038/nmeth.3869

Carpenter, S. R., Benson, B. J., Biggs, R., Chipman, J. W., Foley, J. A., Golding, S. A., Hammer, R. B., Hanson, P. C., Johnson, P. T. J., Kamarainen, A. M., Kratz, T. K., Lathrop, R. C., McMahon, K. D., Provencher, B., Rusak, J. A., Solomon, C. T., Stanley, E. H., Turner, M. G., Vander Zanden, M. J., … Yuan, H. (2007). Understanding Regional Change: A Comparison of Two Lake Districts. *BioScience*, *57*(4), 323–335. https://doi.org/10.1641/B570407

Carpenter, S. R., Stanley, E. H., & Vander Zanden, M. J. (2011). State of the World's Freshwater Ecosystems: Physical, Chemical, and Biological Changes. *Annual Review of Environment and Resources*, *36*(1), 75–99. https://doi.org/10.1146/annurev-environ-021810-094524

Canadian Council of Ministers of the Environment. 2004. Canadian water quality guidelines for the protection of aquatic life: phosphorus: Canadian guidance framework for the management of

freshwater systems. *In Canadian environmental quality guidelines.* Canadian Council of Ministers of the Environment, Winnipeg, Canada

Chen, W., Wilkes, G., Khan, I. U. H., Pintar, K. D. M., Thomas, J. L., Lévesque, C. A., Chapados, J. T., Topp, E., & Lapen, D. R. (2018). Aquatic Bacterial Communities Associated With Land Use and Environmental Factors in Agricultural Landscapes Using a Metabarcoding Approach. *Frontiers in Microbiology*, *9*. https://www.frontiersin.org/articles/10.3389/fmicb.2018.02301

Cotner, J. B., & Biddanda, B. A. (2002). Small Players, Large Role: Microbial Influence on Biogeochemical Processes in Pelagic Aquatic Ecosystems. *Ecosystems*, *5*(2), 105–121. https://doi.org/10.1007/s10021-001-0059-3

Davis, T. W., Bullerjahn, G. S., Tuttle, T., McKay, R. M., & Watson, S. B. (2015). Effects of Increasing Nitrogen and Phosphorus Concentrations on Phytoplankton Community Growth and Toxicity During Planktothrix Blooms in Sandusky Bay, Lake Erie. *Environmental Science & Technology*, *49*(12), 7197–7207. https://doi.org/10.1021/acs.est.5b00799

De Cáceres, M., Legendre, P., & Moretti, M. (2010). Improving indicator species analysis by combining groups of sites. *Oikos*, *119*(10), 1674–1684. https://doi.org/10.1111/j.1600-0706.2010.18334.x

Dufrêne, M., & Legendre, P. (1997). Species Assemblages and Indicator Species:the Need for a Flexible Asymmetrical Approach. *Ecological Monographs*, *67*(3), 345–366. https://doi.org/10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2

Dugan, H. A., Bartlett, S. L., Burke, S. M., Doubek, J. P., Krivak-Tetley, F. E., Skaff, N. K., Summers, J. C., Farrell, K. J., McCullough, I. M., Morales-Williams, A. M., Roberts, D. C., Ouyang, Z., Scordo, F., Hanson, P. C., & Weathers, K. C. (2017). Salting our freshwater lakes. *Proceedings of the National Academy of Sciences*, *114*(17), 4453–4458. https://doi.org/10.1073/pnas.1620211114

Dupont, A., Botrel, M., St-Gelais, N. F., Poisot, T., & Maranger, R. (2023). A social–ecological geography of southern Canadian lakes. *FACETS*, *8*, 1–16. https://doi.org/10.1139/facets-2023-0025

Fournier, I. B., Lovejoy, C., & Vincent, W. F. (2021). Changes in the Community Structure of Under-Ice and Open-Water Microbiomes in Urban Lakes Exposed to Road Salts. *Frontiers in Microbiology*, *12*. https://www.frontiersin.org/articles/10.3389/fmicb.2021.660719

Gao, C.-H., Yu, G., & Cai, P. (2021). ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram. *Frontiers in Genetics*, *12*. https://www.frontiersin.org/articles/10.3389/fgene.2021.706907

Garner, R. E., Kraemer, S. A., Onana, V. E., Fradette, M., Varin, M.-P., Huot, Y., & Walsh, D. A. (2023). A genome catalogue of lake bacterial diversity and its drivers at continental scale. *Nature Microbiology*, *8*(10), Article 10. https://doi.org/10.1038/s41564-023-01435-6

(GOC) Government of Canada. (2015a). Land Use 2010. Open government portal. Available from https://open.canada.ca/data/en/dataset/9e1efe92-e5a3-4f70-b313-68fb1283eadf

(GOC) Government of Canada. (2015b). Canadian Digital Elevation Model. Open government portal. Available from https://open.canada.ca/data/en/dataset/7f245e4d-76c2-4caa-951a-45d1d2051333

(GOC) Government of Canada. (2017). Annual Crop Inventory 2016. Open government portal. Available from https://open.canada.ca/data/en/dataset/b8e4da73-fb5f-4e6e-93a4-8b1f40d95b51

Griffiths, K., Jeziorski, A., Paquette, C., Taranu, Z. E., Baud, A., Antoniades, D., Beisner, B., Hamilton, P. B., Smol, J. P., & Gregory-Eaves, I. (2021). Multi-trophic level responses to

environmental stressors over the past ~150 years: Insights from a lake-rich region of the world. *Ecological Indicators*, *127*, 107700. https://doi.org/10.1016/j.ecolind.2021.107700

Hengy, M. H., Horton, D. J., Uzarski, D. G., & Learman, D. R. (2017). Microbial community diversity patterns are related to physical and chemical differences among temperate lakes near Beaver Island, MI. *PeerJ*, *5*, e3937. https://doi.org/10.7717/peerj.3937

Ho, J. C., Michalak, A. M., & Pahlevan, N. (2019). Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature*, *574*(7780), Article 7780. https://doi.org/10.1038/s41586-019-1648-7

Horton, D. J., Theis, K. R., Uzarski, D. G., & Learman, D. R. (2019). Microbial community structure and microbial networks correspond to nutrient gradients within coastal wetlands of the Laurentian Great Lakes. *FEMS Microbiology Ecology*, *95*(4), fiz033. https://doi.org/10.1093/femsec/fiz033

Huot, Y., Brown, C. A., Potvin, G., Antoniades, D., Baulch, H. M., Beisner, B. E., Bélanger, S., Brazeau, S., Cabana, H., Cardille, J. A., del Giorgio, P. A., Gregory-Eaves, I., Fortin, M.-J., Lang, A. S., Laurion, I., Maranger, R., Prairie, Y. T., Rusak, J. A., Segura, P. A., … Walsh, D. A. (2019). The NSERC Canadian Lake Pulse Network: A national assessment of lake health providing science for water management in a changing climate. *Science of The Total Environment*, *695*, 133668. https://doi.org/10.1016/j.scitotenv.2019.133668

Jankowski, K., Schindler, D. E., & Horner-Devine, M. C. (2014). Resource Availability and Spatial Heterogeneity Control Bacterial Community Response to Nutrient Enrichment in Lakes. *PLOS ONE*, *9*(1), e86991. https://doi.org/10.1371/journal.pone.0086991

Jeppesen, E., Brucet, S., Naselli-Flores, L., Papastergiadou, E., Stefanidis, K., Nõges, T., Nõges, P., Attayde, J. L., Zohary, T., Coppens, J., Bucak, T., Menezes, R. F., Freitas, F. R. S., Kernan, M., Søndergaard, M., & Beklioğlu, M. (2015). Ecological impacts of global warming and water abstraction on lakes and reservoirs due to changes in water level and related changes in salinity. *Hydrobiologia*, *750*(1), 201–227. https://doi.org/10.1007/s10750-014-2169-x

Ji, B., Qin, H., Guo, S., Chen, W., Zhang, X., & Liang, J. (2018). Bacterial communities of four adjacent fresh lakes at different trophic status. *Ecotoxicology and Environmental Safety*, *157*, 388–394. https://doi.org/10.1016/j.ecoenv.2018.03.086

Keeler, B. L., Polasky, S., Brauman, K. A., Johnson, K. A., Finlay, J. C., O'Neill, A., Kovacs, K., & Dalzell, B. (2012). Linking water quality and well-being for improved assessment and valuation of ecosystem services. *Proceedings of the National Academy of Sciences*, *109*(45), 18619–18624. https://doi.org/10.1073/pnas.1215991109

Kraemer, S. A., Barbosa da Costa, N., Shapiro, B. J., Fradette, M., Huot, Y., & Walsh, D. A. (2020). A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. *The ISME Journal*, *14*(12), 3011–3023. https://doi.org/10.1038/s41396-020-0733-0

Langenheder, S., & Lindström, E. S. (2019). Factors influencing aquatic and terrestrial bacterial community assembly. *Environmental Microbiology Reports*, *11*(3), 306–315. https://doi.org/10.1111/1758-2229.12731

Linz, A. M., He, S., Stevens, S. L. R., Anantharaman, K., Rohwer, R. R., Malmstrom, R. R., Bertilsson, S., & McMahon, K. D. (2018). Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *PeerJ*, *6*, e6075. https://doi.org/10.7717/peerj.6075

Liu, K., Liu, Y., Hu, A., Wang, F., Chen, Y., Gu, Z., Anslan, S., & Hou, J. (2020). Different community assembly mechanisms underlie similar biogeography of bacteria and microeukaryotes in Tibetan lakes. *FEMS Microbiology Ecology*, *96*(6), fiaa071. https://doi.org/10.1093/femsec/fiaa071

Logue, J. B., Langenheder, S., Andersson, A. F., Bertilsson, S., Drakare, S., Lanzén, A., & Lindström, E. S. (2012). Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species–area relationships. *The ISME Journal*, *6*(6), 1127–1136. https://doi.org/10.1038/ismej.2011.184

MacKeigan, P. W., Zastepa, A., Taranu, Z. E., Westrick, J. A., Liang, A., Pick, F. R., Beisner, B. E., & Gregory-Eaves, I. (2023). Microcystin concentrations and congener composition in relation to environmental variables across 440 north-temperate and boreal lakes. *Science of The Total Environment*, *884*, 163811. https://doi.org/10.1016/j.scitotenv.2023.163811

MacLeod, J., Keller, B., Paterson, A., Dyer, R., & Gunn, J. (2017). Scale and Watershed Features Determine Lake Chemistry Patterns Across Physiographic Regions in the Far North of Ontario, Canada. *Journal of Limnology*, *76*, 211–220. https://doi.org/10.4081/jlimnol.2016.1553

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H., & Staley, J. T. (2006). Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology*, *4*(2), Article 2. https://doi.org/10.1038/nrmicro1341

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading : The Official Journal of the Society for the Scientific Study of Reading*, *20*(1), 20–33. https://doi.org/10.1080/10888438.2015.1107073

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13603

Minns, C. K., Moore, J. E., Shuter, B. J., & Mandrak, N. E. (2008). A preliminary national analysis of some key characteristics of Canadian lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(8), 1763–1778. https://doi.org/10.1139/F08-110

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews: MMBR*, *75*(1), 14–49. https://doi.org/10.1128/MMBR.00028-10

NSERC Canadian Lake Pulse Network. (2021). *NSERC Canadian Lake Pulse Network field manual 2017—2018—2019 surveys* (M.-P. Varin, M.-L. Beaulieu, & Y. Huot, Eds.). Université de Sherbrooke. https://doi.org/10.17118/11143/18662

Numberger, D., Zoccarato, L., Woodhouse, J., Ganzert, L., Sauer, S., Márquez, J. R. G., Domisch, S., Grossart, H.-P., & Greenwood, A. D. (2022). Urbanization promotes specific bacteria in freshwater microbiomes including potential pathogens. *Science of The Total Environment*, *845*, 157321. https://doi.org/10.1016/j.scitotenv.2022.157321

Obieze, C. C., Wani, G. A., Shah, M. A., Reshi, Z. A., Comeau, A. M., & Khasa, D. P. (2022). Anthropogenic activities and geographic locations regulate microbial diversity, community assembly and species sorting in Canadian and Indian freshwater lakes. *Science of The Total Environment*, *826*, 154292. https://doi.org/10.1016/j.scitotenv.2022.154292

Oksanen J., Guillaume Blanchet F., Friendly M., Kindt R., Legendre P., McGlinn D., et al. 2020. Vegan: community ecology package [software]. R package Ver. 2.5–7. Available from https://CRAN.R-project.org/package=vegan

Oliva, A., Onana, V. E., Garner, R. E., Kraemer, S. A., Fradette, M., Walsh, D. A., & Huot, Y. (2023). Geospatial analysis reveals a hotspot of fecal bacteria in Canadian prairie lakes

linked to agricultural non-point sources. *Water Research*, *231*, 119596. https://doi.org/10.1016/j.watres.2023.119596

Paver, S. F., Newton, R. J., & Coleman, M. L. (2020). Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environmental Microbiology*, *22*(1), 433–446. https://doi.org/10.1111/1462-2920.14862

Pruesse, E., Peplies, J., & Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, *28*(14), 1823–1829. https://doi.org/10.1093/bioinformatics/bts252

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590–D596. https://doi.org/10.1093/nar/gks1219

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T. J., Kidd, K. A., MacCormack, T. J., Olden, J. D., Ormerod, S. J., Smol, J. P., Taylor, W. W., Tockner, K., Vermaire, J. C., Dudgeon, D., & Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews*, *94*(3), 849–873. https://doi.org/10.1111/brv.12480

Reid, K., & Schneider, K. D. (2019). Phosphorus accumulation in Canadian agricultural soils over 30 yr. *Canadian Journal of Soil Science*, *99*(4), 520–532. https://doi.org/10.1139/cjss-2019-0023

Ribeiro P.J. Jr, Diggle P.J., Schlather M., Bivand R., Ripley B. 2020. geoR: analysis of geostatistical data [software]. R package Ver. 1.8–1. Available from https://cran.r-project.org/package=geoR

Rohwer, R. R., Hamilton, J. J., Newton, R. J., & McMahon, K. D. (2018). TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere*, *3*(5), e00327-18. https://doi.org/10.1128/mSphere.00327-18

Ryo, M., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, *8*(11), e01976. https://doi.org/10.1002/ecs2.1976

Sadeghi, J., Chaganti, S. R., Shahraki, A. H., & Heath, D. D. (2021). Microbial community and abiotic effects on aquatic bacterial communities in north temperate lakes. *Science of The Total Environment*, *781*, 146771. https://doi.org/10.1016/j.scitotenv.2021.146771

Schacht, J. R. S., MacKeigan, P. W., Taranu, Z. E., Huot, Y., & Gregory-Eaves, I. (2023). Agricultural land use and morphometry explain substantial variation in nutrient and ion concentrations in lakes across Canada. *Canadian Journal of Fisheries and Aquatic Sciences*, *80*(11), 1785–1797. https://doi.org/10.1139/cjfas-2023-0109

Shahraki, A. H., Chaganti, S. R., & Heath, D. (2021). Spatio-temporal dynamics of bacterial communities in the shoreline of Laurentian great Lake Erie and Lake St. Clair's large freshwater ecosystems. *BMC Microbiology*, *21*, 253. https://doi.org/10.1186/s12866-021-02306-y

Smith, V. H. (2003). Eutrophication of freshwater and coastal marine ecosystems: A global problem. *Environmental Science and Pollution Research International*, *10*(2), 126–139. https://doi.org/10.1065/espr2002.12.142. PMID: 12729046.

Solomon, C. T., Jones, S. E., Weidel, B. C., Buffam, I., Fork, M. L., Karlsson, J., Larsen, S., Lennon, J. T., Read, J. S., Sadro, S., & Saros, J. E. (2015). Ecosystem Consequences of Changing Inputs of Terrestrial Dissolved Organic Matter to Lakes: Current Knowledge and Future Challenges. *Ecosystems*, *18*(3), 376–389. https://doi.org/10.1007/s10021-015-9848-y

Sterner, R. W., Keeler, B., Polasky, S., Poudel, R., Rhude, K., & Rogers, M. (2020). Ecosystem services of Earth's largest freshwater lakes. *Ecosystem Services*, *41*, 101046. https://doi.org/10.1016/j.ecoser.2019.101046

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Taranu, Z. E., & Gregory-Eaves, I. (2008). Quantifying Relationships Among Phosphorus, Agriculture, and Lake Depth at an Inter-Regional Scale. *Ecosystems*, *11*(5), 715–725. https://doi.org/10.1007/s10021-008-9153-0

Ting, A. S. Y., Zoqratt, M. Z. H. M., Tan, H. S., Hermawan, A. A., Talei, A., & Khu, S. T. (2021). Bacterial and eukaryotic microbial communities in urban water systems profiled via Illumina MiSeq platform. *3 Biotech*, *11*(2), 40. https://doi.org/10.1007/s13205-020-02617-3

Tranvik, L. J., Cole, J. J., & Prairie, Y. T. (2018). The study of carbon in inland waters—From isolated ecosystems to players in the global carbon cycle. *Limnology and Oceanography Letters*, *3*(3), 41–48. https://doi.org/10.1002/lol2.10068

Verpoorter, C., Kutser, T., Seekell, D. A., & Tranvik, L. J. (2014). A global inventory of lakes based on high-resolution satellite imagery. *Geophysical Research Letters*, *41*(18), 6396–6402. https://doi.org/10.1002/2014GL060641

Vitousek, P. M., Mooney, H. A., Lubchenco, J., & Melillo, J. M. (1997). Human Domination of Earth's Ecosystems. *Science*, *277*(5325), 494–499. https://doi.org/10.1126/science.277.5325.494

Wang, Y., Guo, M., Li, X., Liu, G., Hua, Y., Zhao, J., Huguet, A., & Li, S. (2022). Shifts in microbial communities in shallow lakes depending on trophic states: Feasibility as an evaluation index for eutrophication. *Ecological Indicators*, *136*, 108691. https://doi.org/10.1016/j.ecolind.2022.108691

Wiken EB, Gauthier D, Marshall I, Lawton K, Hirvonen H. 1996. A perspective on Canada's ecosystems: an overview of the terrestrial and marine ecozones. *In Canadian Council on Ecological Areas occasional papers.* Canadian Council on Ecological Areas, Ottawa, Canada.

Williamson, C. E., Saros, J. E., Vincent, W. F., & Smol, J. P. (2009). Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnology and Oceanography*, *54*(6part2), 2273–2282. https://doi.org/10.4319/lo.2009.54.6_part_2.2273

Yang, J., Jiang, H., Liu, W., Huang, L., Huang, J., Wang, B., Dong, H., Chu, R. K., & Tolic, N. (2020). Potential utilization of terrestrially derived dissolved organic matter by aquatic microbial communities in saline lakes. *The ISME Journal*, *14*(9), 2313–2324. https://doi.org/10.1038/s41396-020-0689-0

Yannarell, A. C., & Triplett, E. W. (2005). Geographic and Environmental Sources of Variation in Lake Bacterial Community Composition. *Applied and Environmental Microbiology*, *71*(1), 227–239. https://doi.org/10.1128/AEM.71.1.227-239.2005

Zeglin, L. H. (2015). Stream microbial diversity in response to environmental changes: Review and synthesis of existing research. *Frontiers in Microbiology*, *6*. https://www.frontiersin.org/articles/10.3389/fmicb.2015.00454

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Chapter 4: Functional profiling of Canadian lake microbiomes reveals an association between agriculture and bacterial metabolism of xenobiotics

4.1 Abstract

Microbes are a significant biological component of freshwater ecosystems that respond rapidly to human-induced environmental changes. Their population dynamics and genomic traits may therefore be useful for assessing the impact of environmental stressors in lake ecosystems. In this study, we conducted a metagenomic survey of microbial communities in 357 lakes sampled across Canada that capture a wide range of environmental conditions. Our gene-centric analyses along a wide trophic gradient from ultraoligotrophic to hypereutrophic lakes revealed a unimodal relationship between bacterial richness and lake trophic status wherein highest bacterial richness was at intermediate trophic states (mesoeutrophic to mesotrophic lakes). Remarkably, we uncovered an unexpected abundance of Firmicutes, typically associated with soils, in high nutrient lakes indicating anthropogenic disturbances. Spatial variation in taxonomic and functional composition was primarily driven by lake physicochemistry, with pH, total nitrogen concentration, and chlorophyll-*a* being significant predictors. Lake metagenomes were enriched with genes coding for xenobiotics biodegradation and metabolism, particularly in lakes with high nutrient concentrations, highlighting the potential for these environments to influence pollutant breakdown. Agriculture emerged as the dominant driver of xenobiotics biodegradation and metabolism in lakes. These findings underscore the interconnectedness of lake ecosystems with their surrounding landscapes and have important implications for understanding the functioning of freshwater ecosystems across Canada.

4.2 Introduction

Covering the northern expanse of North America, Canada is a vast and lake-rich nation, holding around 20% of the world's freshwater (Huot et al., 2019). With over one million lakes distributed across its diverse landscapes, these water bodies are dispersed among various ecozones, which represent distinct climatic and ecological regions based on prevailing abiotic and biotic factors (CCEA 2016). Each ecozone contributes to the wide-ranging physicochemical heterogeneity observed in lakes across the country. For instance, lakes in the Prairies ecozone (spanning Alberta, Saskatchewan, and Manitoba), are generally shallower and nutrient-rich, reflecting the productive agricultural nature of the lands. Agricultural activities exert significant impacts on water quality by

introducing nutrients, salts, pesticides, organic matter, sediments, pathogens, heavy metals, and other emerging contaminants into surface water bodies through processes such as runoff, erosion, or leaching (Rey-Romero et al., 2022; Zia et al., 2013). Phosphorus (P) runoff from agricultural land has been reported in Canadian provinces of Saskatchewan and Manitoba (J. Liu et al., 2021) as well as organic carbon and nutrient loads (Y. Liu et al., 2023). In Alberta, watershed surveys comprising mainly agricultural lands revealed pesticide occurrence in aquatic agroecosystems reflecting runoffs from agricultural land use and effluent pollutions from urban land use (Sheedy et al., 2019).

Xenobiotics is a term that refers to chemical compounds that do not occur naturally. Xenobiotics include pesticides and other contaminants such as azodyes, phenolics, polycyclic aromatic hydrocarbons (PAHs), halogenated compounds, personal care products (PCPs), pharmaceutical active compounds (PhACs), nitroaromatic compounds, triazines, and chlorinated compounds. Many of these adversely affect aquatic ecosystems due to their slow degradation (Štefanac et al., 2021). Xenobiotic in aquatic ecosystems have been linked to agricultural and urban development runoffs into lakes (Rieger et al., 2002; Garner et al., 2023).

Since lakes are characterized by high microbial diversity (Dudgeon et al., 2006; Garcia-Moreno et al., 2014; Strayer & Dudgeon, 2010) that carry out biodegradation activities (Cotner & Biddanda, 2002), it follows that lakes may harbour organisms capable of biologically degrading xenobiotics. Due to their genetic diversity and functionality, bacterial communities exhibit remarkable metabolic potential with genes and metabolic pathways implicated in the process of biodegradation (Mishra et al., 2021); for instance, various genes participating in the aerobic degradation pathway of atrazine, a commonly researched pesticide, have been pinpointed in the *Pseudomonas* sp. strain ADP. Subsequently, these genes were employed as biomarkers, given their high conservation across different bacterial genera, for detecting atrazine-degrading capabilities (Aldas-Vargas et al., 2022; de Souza et al., 1998).

Such research conducted to identify genes responsible for lake bacteria's capability to metabolize certain xenobiotics have been tremendously facilitated by the advancement of metagenomics, a culture-independent genomic approach, which has revolutionized our understanding of microbial communities. Shotgun metagenomics involves sequencing genomic DNA isolated from an environmental sample. This approach differs from 16S rRNA amplicon sequencing and provides

more information on community diversity. For example, metagenomics allows for the identification and profiling of bacteria, fungi, viruses, and various other microorganisms. Furthermore, it enables the identification and profiling of microbial metabolic genes present in the environment, revealing valuable insights into the functional potential of the microbiome (Quince et al., 2017).

Over the past decade, metagenomics has led to significant expansion in our understand of microbial diversity across habitats. These insights include a 44% expansion of known phylogenetic diversity among bacteria and archaea in oceans, human and animal hosts, engineered environments, as well as natural and agricultural soils (Nayfach et al., 2021). Metagenomic approaches have also deepened our understanding of archaeal taxonomy, exemplified by the discovery of the Asgard superphylum, which includes the Lokiarchaeota (Spang et al., 2015). Moreover, these metagenomic techniques have found applications in various fields, ranging from agriculture, pollution control, and energy to broader aspects of biology, including medicine (Zhang et al., 2021).

Furthermore, metagenomic assessments can offer significant information on the relationships between microbial communities and environmental conditions. For instance, metagenomics has been used to provide a broad view on microbial diversity in productive urban lakes in Sweden (Rodríguez-Gijón et al., 2023). Likewise, metagenomic based ecosystem monitoring has facilitated assessment of freshwater ecosystem biodiversity and water quality in some Canadian rivers and lakes (Edge et al., 2020). Metagenomics has also been used to implement biotechnological strategies in the environment for heavy metal in soils (Feng et al., 2018; Xing et al., 2020). High resolution genome-resolved metagenomics studies have also been beneficial in linking microbial function to taxonomy and paved the way for new discoveries in microbial ecology (Grossart et al., 2020). Overall, these research avenues have contributed significantly to understanding the relationships between microbial dynamics and biogeochemical cycles.

Metagenomics can be used to investigate diversity beyond microbial communities. In a regional investigation of Canadian lakes, metagenomics has been used to describe zooplankton diversity in eastern Canada. In this study, comparative analysis between zooplankton morphological identification and metagenomic identification showed a significant correlation of zooplankton abundances (Monchamp et al., 2022). Comparative metagenomic methods have been used to probe protists diversity in contemporary and pre-industrial sediments from eastern Canada lakes (Garner et

al., 2020) lakes but only a handful of investigations have been done on prokaryotic diversity in Canadian surface waters across the continental scale.

The objective of the research presented in this chapter is to 1) provide an inventory of prokaryotic taxonomic composition from metagenomes across the continental scale, and 2) to elucidate the functional composition of lake bacteria across the continental scale and 3) to identify the impact of watershed environmental conditions and human land use on both bacterial taxonomic and functional composition. Using annotated metagenomic assemblies we resolved which microbial metabolisms are most impacted by both lake limnological conditions and human activities within the watershed. We also identify potential metabolic capabilities of lake bacteria using protein coding genes and investigated the susceptibility of these metabolisms to environmental and human land use disturbances.

4.3 Results

4.3.1 The LakePulse metagenome resource

We generated metagenomic data for 357 lakes across 12 Canadian ecozones sampled at the peak of summer stratification between 2017 to 2019 via the first pan Canadian lake health assessment scheme called the NSERC Canadian Lake Pulse Network (referred to as LakePulse) as previously described by (Huot et al., 2019; Kraemer et al., 2020). To produce this metagenome resource, we carried out whole genome shotgun sequencing on community DNA from 366 lakes representing the longitudinal and latitudinal scale of Canada. Nine hypersaline lakes were excluded from this dataset leaving a total of 357 lakes for our metagenomic resource (**Figure 4.1**). Lakes depicted a vast environmental heterogeneity as demonstrated by their variation in physicochemical conditions, productivity levels, morphometry, climatic conditions, human land use and soil properties within the watershed. Across the continental gradient (43 – 68 °N, 53 – 141 °W), we detected a wide trophic gradient from ultraoligotrophic to hypereutrophic across our lakes (**Figure 4.1**). Regional comparison of lakes revealed some spatial differentiation in lake characteristics: Prairies ecozone lakes of western Canada were shallow, nutrient-rich and highly productive compared to lakes in eastern Canada within the Mixedwood Plains ecozone where watersheds had the most extensive built development and human population density.

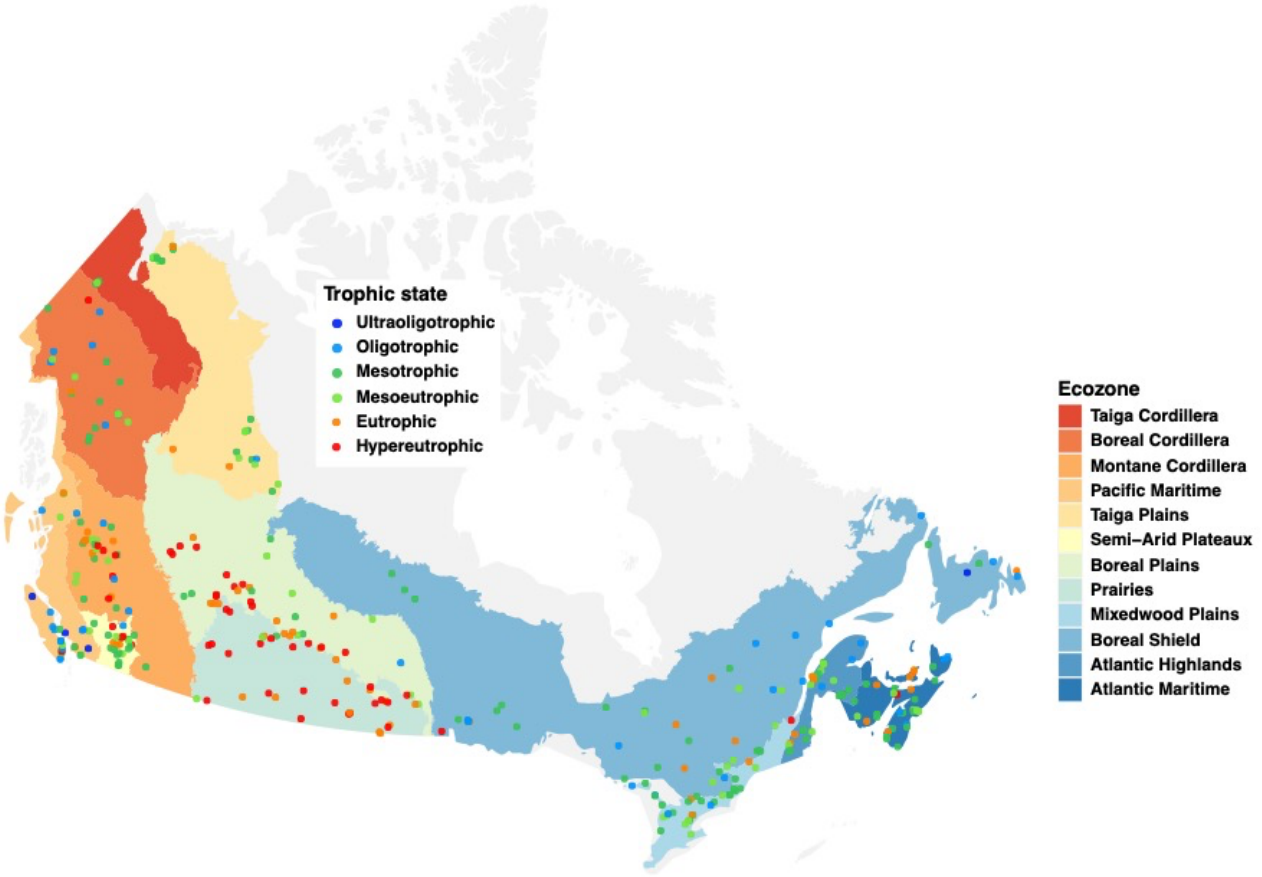


Figure 4.1 *Distribution of lakes sampled across Canada. Background colour indicates ecozone and the points indicate trophic state of sampled lakes based on total phosphorus concentration. Ultraoligotrophic (<4 µg/L), Oligotrophic (4 – 10 µg/L), Mesotrophic (10 – 20 µg/L), Mesoeutrophic (20 – 35 µg/L), Eutrophic (35 – 100 µg/L), and Hypereutrophic (>100 µg/L).*

4.3.2 Taxonomic overview of lake metagenomes

We assessed the abundance of operational taxonomic units (OTUs) directly from unassembled metagenomic data using single copy ribosomal proteins (rpL2). A total of 30,771 OTUs were identified, out of which 88% were bacterial and the rest were either archaeal or eukaryotic OTUs or were unclassified. To elucidate taxonomic diversity and composition of lake metagenomes, we grouped lakes according to trophic states and computed the relative abundance of annotated OTUs in lakes. Estimated bacterial richness was assessed using Chao1 index and greater bacterial richness was observed within the Prairies and the Boreal Plains ecozones of western Canada (**Figure 4.2a**) with a peak at intermediate (mesoeutrophic to mesotrophic) trophic levels (**Figure 4.2b**).

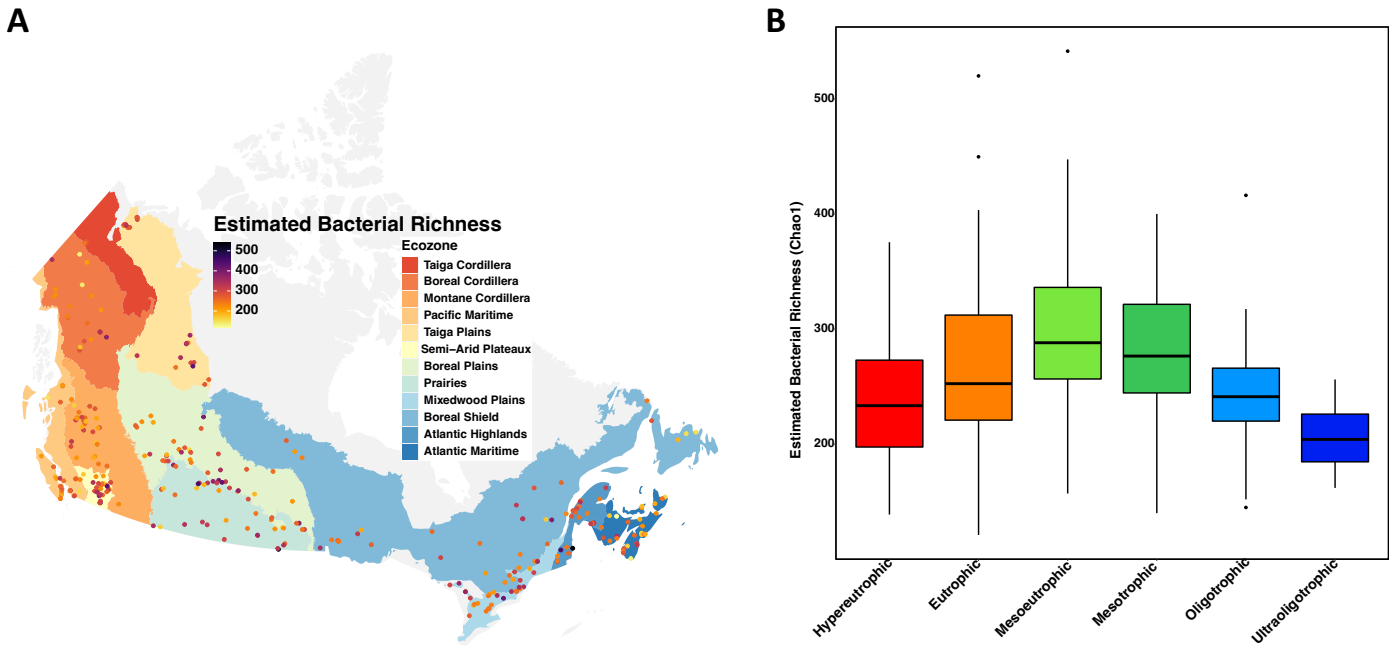


Figure 4.2 *Estimated bacterial richness across sampled ecozones and trophic states. A) Distribution of lakes sampled across Canada. The background colour indicates ecozone and the points indicate estimated bacterial richness in lakes. B) Estimated bacterial richness in lakes showing peaks of richness at intermediate levels of productivity.*

For bacterial taxonomic composition, we found that the most abundant OTUs belonged to the phyla Actinobacteria (42%), Proteobacteria (30%), Bacteriodetes (10%), Verrucomicrobia (7%) and Cyanobacteria (4%). We also observed an unexpected abundance of Firmicutes in eutrophic to hypereutrophic lakes within the Prairies and Boreal Plains (**Figure 4.3**). To explore the variation in the taxonomic distribution of lake bacterial communities, we conducted a Principal Component Analysis (PCA) on the bacterial taxonomic composition data. Principal Component axis 1 (PC1) captured 6.81% of the variance in bacterial taxonomic composition while displaying a spatial variation that revealed a dispersion of high nutrient lakes located in the Prairies and Boreal Plains from others (**Figure 4.4**).

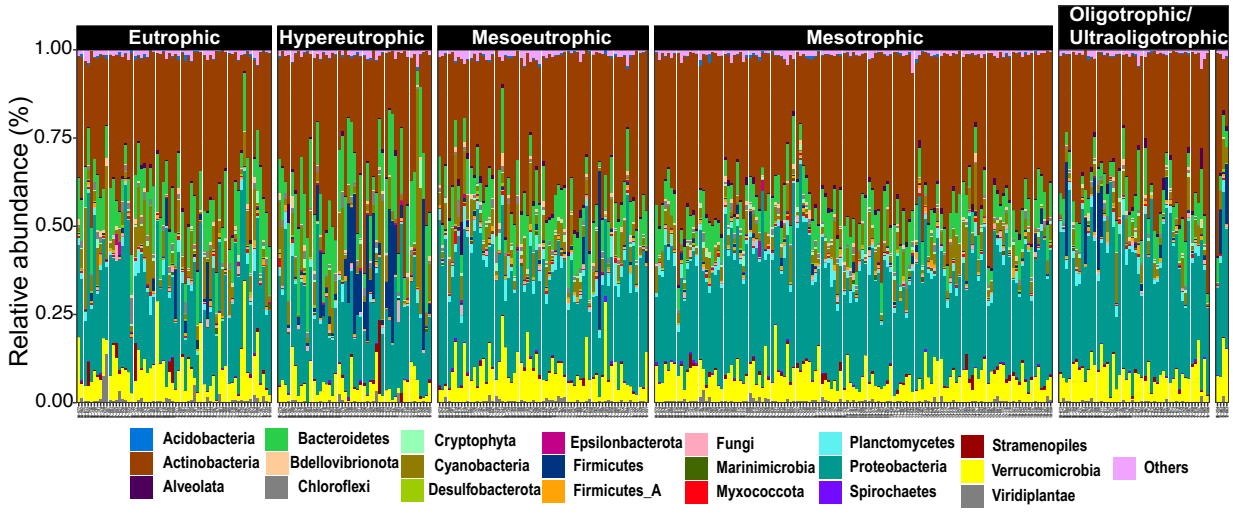


Figure 4.3 *Taxonomic composition of lake metagenomes grouped by trophic state.*

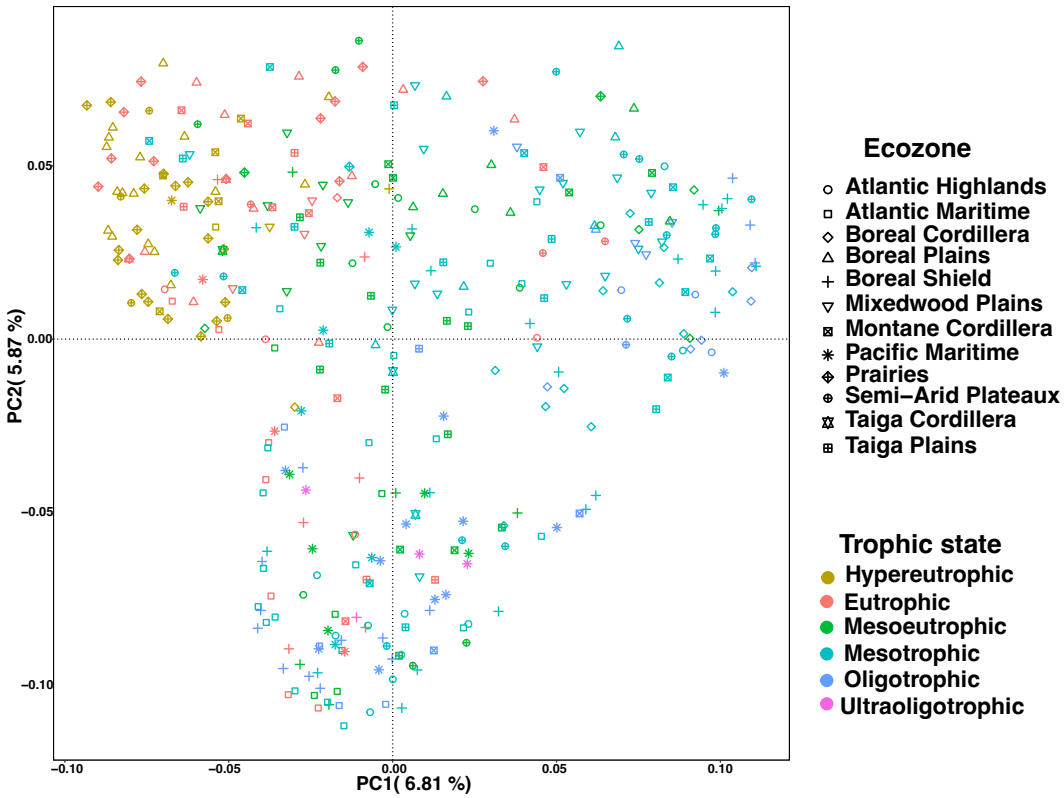


Figure 4.4 *Principal component analysis (PCA) showing the taxonomic distribution of lake bacterial communities across ordination space.*

4.3.3 Functional overview of lake metagenomes

Functional composition of lake metagenomes was performed using the KEGG database, a popular reference database that assigns proteins to orthologs group, termed KEGG orthologs (KOs) (Kanehisa et al., 2022) that are organized in to metabolic pathway maps. We identified 4,538 KOs across metagenomes that were assigned to carbohydrate metabolism (847 KOs), energy metabolism (549 KOs), lipid metabolism (371 KOs), nucleotide metabolism (224 KOs), amino acid metabolism (514 Kos), glycan biosynthesis (272 KOs), terpenoids and polyketide metabolism (396), xenobiotics biodegradation (272 KOs), biosynthesis of secondary metabolites (129 KOs), cofactor and vitamins metabolism (387 KOs) and other metabolic pathways (membrane transport and motility). We investigated the functional distribution of lake bacterial communities via a Principal Component Analysis (PCA) using the relative abundance of KOs across lakes. Functional composition was observed to vary more than taxonomic composition as Principal Component axis 1 (PC1) captured 19.6% of the variance in bacterial functional composition, more than double the variance captured in the taxonomic composition PCA. Interestingly, variance was driven by high nutrient lakes located in the Prairies and the Boreal Plains (**Figure 4.5**).

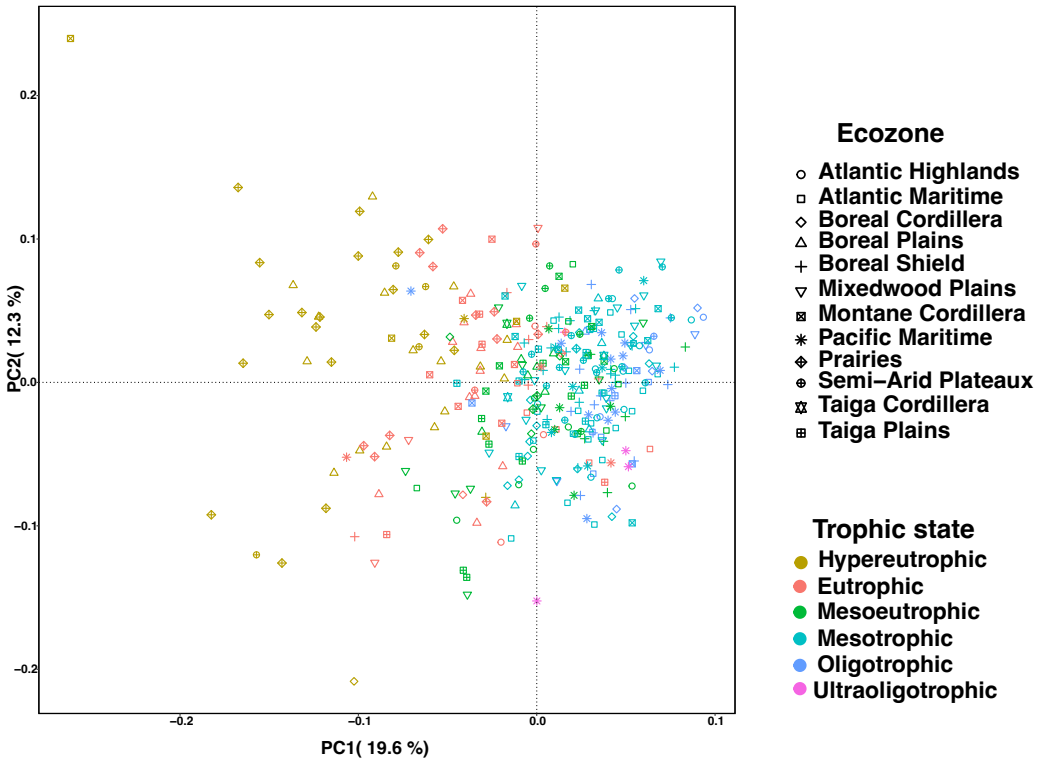
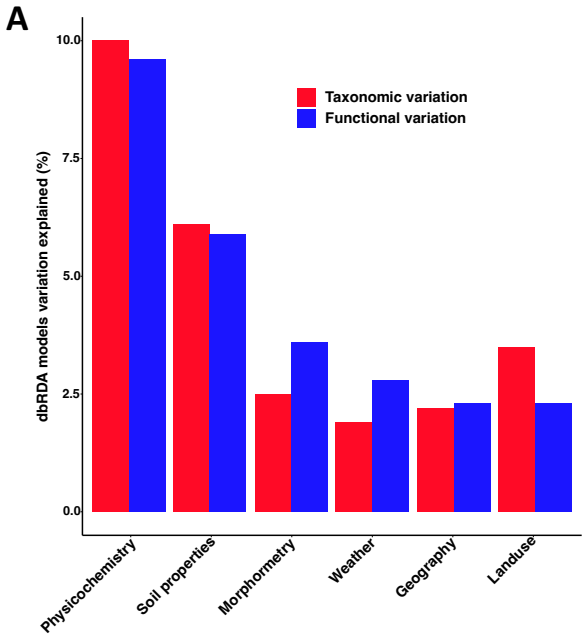


Figure 4.5 *Principal component analysis (PCA) showing the functional distribution of lake bacterial communities across ordination space.*

4.3.4 Drivers of variation in lake taxonomic and functional composition

We assessed the environmental drivers of both taxonomic and functional composition of lake bacterial communities using distance-based redundancy analysis (dbRDA), a constrained ordination method that identifies variables with the most explanatory power. Here, we compared six models, that incorporated variables that reflected either lake geography, physicochemistry, watershed land use, morphometry, soil properties or climate. All models were significant (*p value* = 0.001) for both taxonomic and functional composition. We observed that the physicochemistry model explained the greatest amount of variation for both taxonomic composition (10%) and functional composition (9%) (**Figure 4.6a**). Within the physiochemistry model, the most important predictors of taxonomic composition were pH, total nitrogen (TN), and potassium concentration in lakes (**Figure 4.6b**). For functional composition, pH, DIC, TN and chlorophyll-*a* concentrations were the most important predictors (**Figure 4.6c**). Our land use model explained a lesser, but still significant amount of variation for both taxonomic composition (3.5%) (**Figure 4.6d**) and functional composition (2.3%) (**Figure 4.6e**) and agriculture was the single most important variable explaining this observed variation in both models.



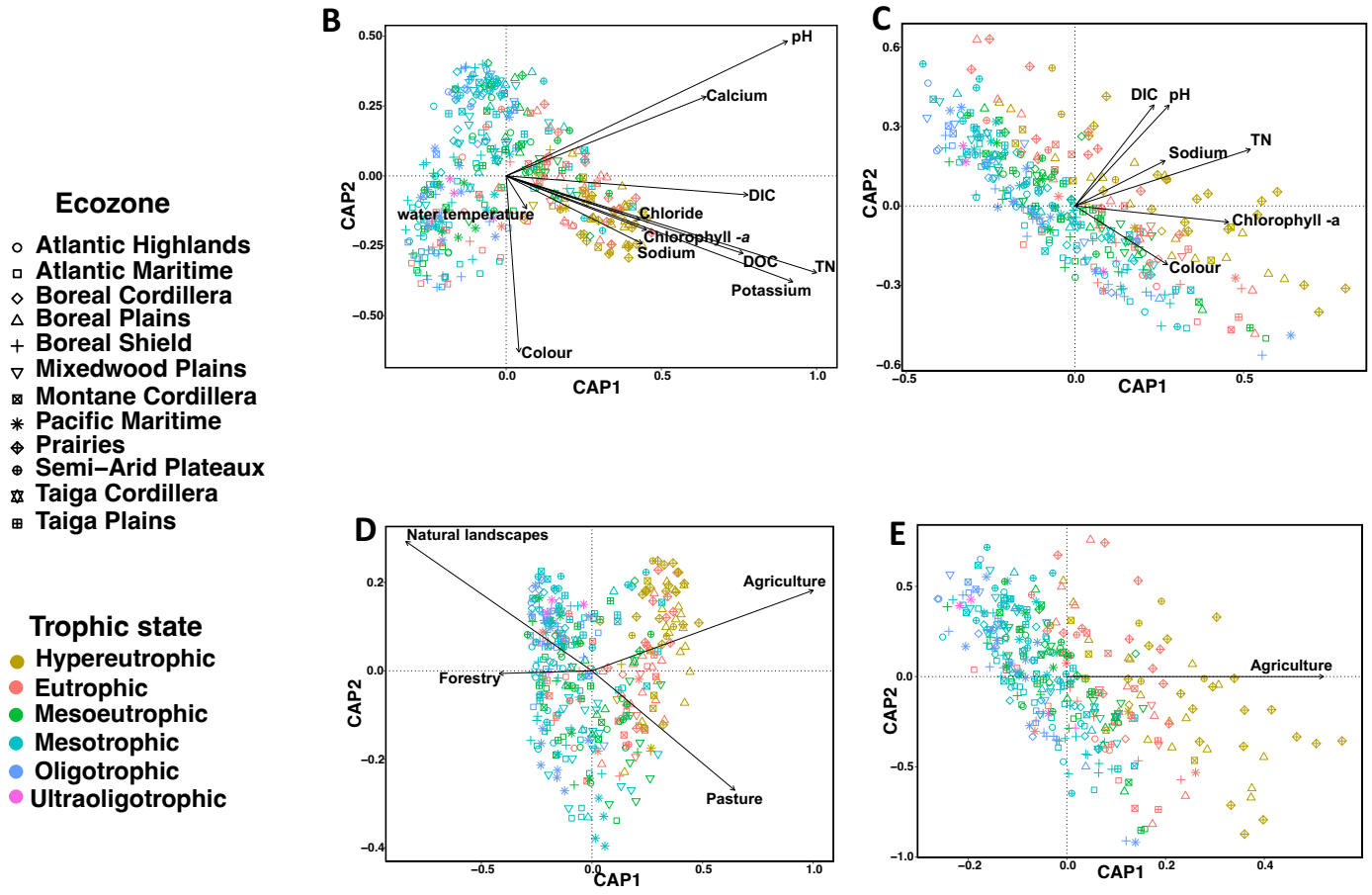


Figure 4.6 *Influence of geography, physicochemistry and land use on bacterial taxonomic composition and function. A) Summary of db-RDA analysis on six categories of environmental and land use models depicting their influence on taxonomic and functional composition. B) db-RDA showing the influence of water chemistry variables on bacterial taxonomic composition. C) db-RDA showing the influence of water chemistry variables on bacterial functional composition. D) db-RDA showing the influence of land use types on bacterial taxonomic composition. E) db-RDA showing the influence of land use types on bacterial functional composition.*

Next, we investigated how well our models explained variation within specific metabolic categories across lakes by separating KOs into those assigned to either carbohydrate metabolism, energy metabolism, lipid metabolism, nucleotide metabolism, amino acid metabolism, glycan biosynthesis, terpenoids and polyketide metabolism or xenobiotics biodegradation and metabolism. Our dbRDA analysis using six broad models of environmental variables to access drivers of functional composition showed that lake physicochemistry explained the greatest variation in functional composition across all metabolism categories. Moreover, xenobiotics biodegradation and metabolism was the most influenced by environmental conditions with 15% of variation explained by lake physicochemistry (**Table 4**). On assessing the specific variables within the physicochemistry model, it was detected that lake pH, TN, and calcium concentration were the strongest predictors of variation in xenobiotics biodegradation and metabolism (**Figure 4.7a**). In addition, while our land use model explained relatively less amount of variation in xenobiotics biodegradation and metabolism (5.1%) (**Figure 4.7b**) (**Table 4**), we detected that agricultural land use and pasture within the watershed were the strongest predictors of variation in xenobiotics biodegradation and metabolism within the land use model. Interesting, this was linked to high nutrient lakes within the Prairies and the Boreal Shield.

Table 4 Summary of percentage variation explained by db-RDA models on bacterial metabolic category.

Metabolism	% variation explained by different db-RDA model categories (dbRDA)					
	Physicochemistry	Land use	Morphometry	Soil properties	Geography	Climate
Carbohydrate metabolism	8.2	1.5	2.9	5.0	1.1	2.2
Energy metabolism	10.3	1.7	3.9	4.9	0.9	2.3
Lipid metabolism	7.6	1.6	3.0	5.0	2.5	2.2
Nucleotide metabolism	11.1	3.6	3.1	6.4	2.1	2.9

Metabolism	% variation explained by different db-RDA model categories (dbRDA)					
	Physicochemistry	Land use	Morphometry	Soil properties	Geography	Climate
Aminoacid metabolism	8.9	2.4	3.7	5.6	2.1	3.2
Glycan biosynthesis	6.3	2.0	3.3	4.5	0.9	2.9
Terpenoids and polyketide	9.7	1.6	3.2	5.9	2.0	2.9
Xenobiotics biodegradation	15.2	5.1	5.8	10.7	2.5	3.3

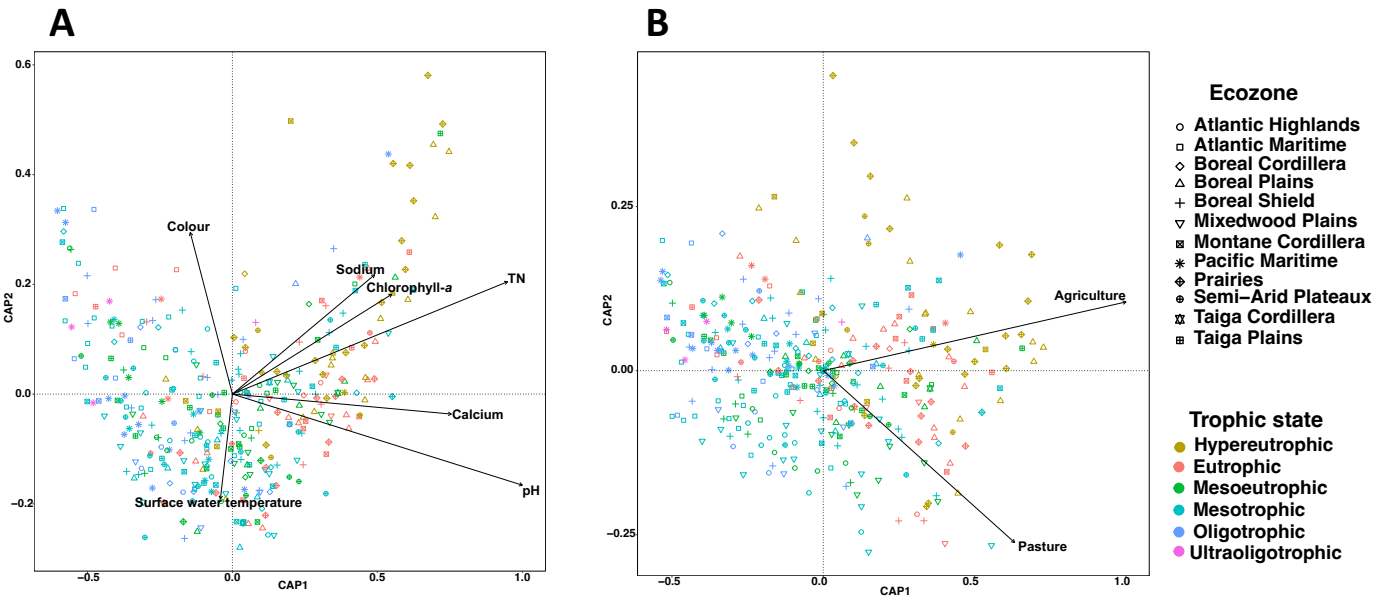


Figure 4.7 *Influence of physicochemical and land use variables on xenobiotics biodegradation and metabolism. A) db-RDA showing the influence of water chemistry parameters on xenobiotics biodegradation, most influential variables in pH, TN among others. B) db-RDA showing the influence of land use types on xenobiotic biodegradation reveals agriculture and pasture as influential predictors of changes in metabolism across lakes.*

4.3.5 Elucidating metabolic features of metagenomes

We analyzed the KO profiles from the lake metagenomic dataset using non-negative matrix factorization (NMF) analysis to reduce the matrix dimensionality. An advantage of this method is that it is capable of directly linking the overall structure of an abundance matrix to the individual elements (KO number) driving a perceived structure. This kind of analysis unlike constrained ordinations, is useful in decomposing signals in a large biological dataset, in this case, a protein coding gene matrix, into the individual elements that are summed up to make the signal (i.e. metabolic genes specifically associated with lake environmental conditions and trophic status). In reducing the dimensionality of the data, the objective of NMF on the KO matrix is to find a small number of metagenomes that depict the underlying structure in the dataset. In our analyses, NMF decomposed our KO abundances into two matrices. Matrix 1 represented a reduced number of elements that describe the overall similarities of the metagenomes based on KO number composition, while matrix 2 represented the weighted contribution of individual KO numbers on each of the elements in matrix 1. It also revealed a rank of the factorization, which represents the number of latent factors in the decomposition (in our case, a rank of 6). This NMF rank revealed that six sub-metagenomes (SMGs) optimally described the biogeography of lake microbial metabolism (**Supplementary Figure S4a-b**) as they represented the individual specific environmental conditions attributable to the metabolic structure in the dataset. SMG 1 was closely associated with lakes with high trophic states, SMG 3 was associated with lakes that have high concentrations of DOC, DIC and chlorophyll-*a*; SMG 5 was similarly associated with high DOC, DIC, chlorophyll-*a* and lake trophic state. However, SMGs 2, 4 and 6 showed a broad relationship with geographic and other environmental conditions. (**Figure 4.8)**.

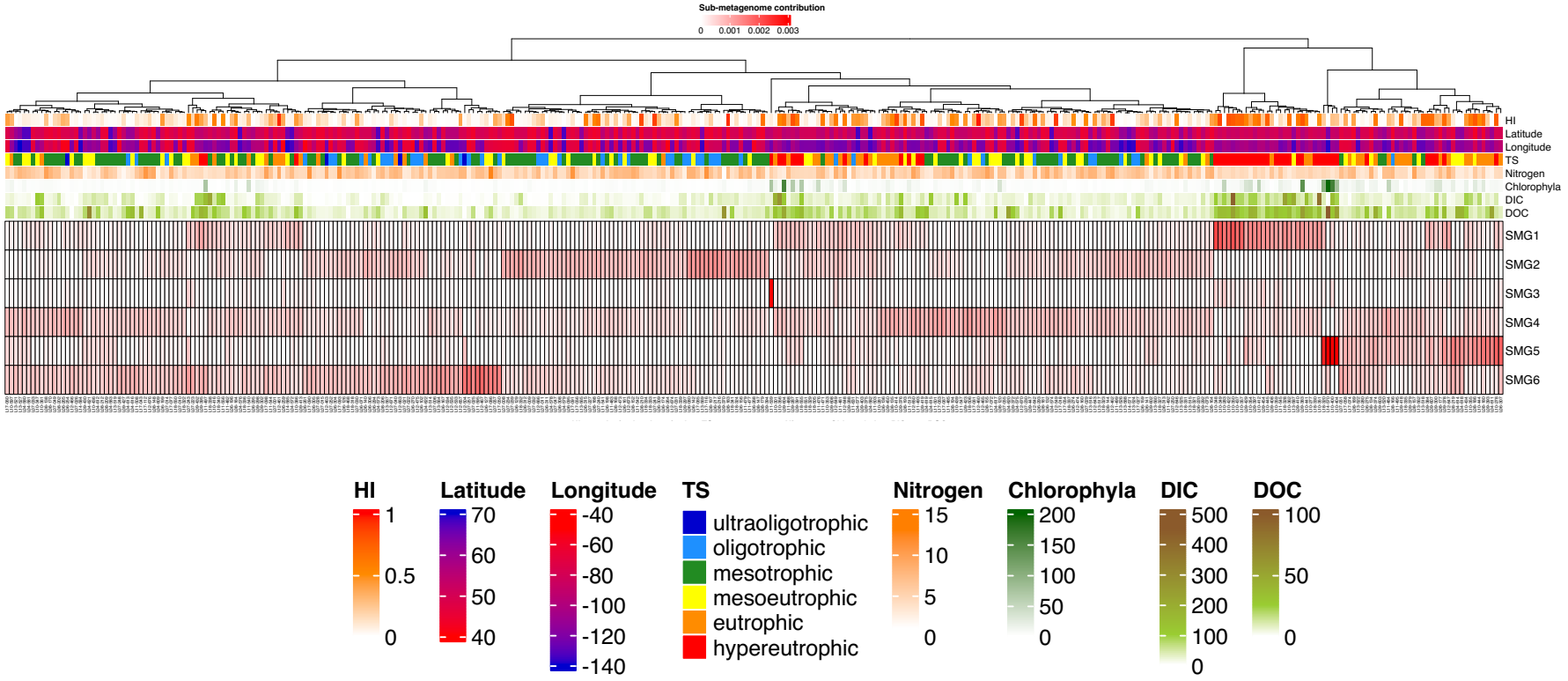


Figure 4.8 *Associating submetagenomes to geography and lake environmental conditions. NMF analysis showing decomposition of six elements (referred to as submetagenomes) best representing the overall functional composition of metagenomes and the contribution of each submetagenome to the lake metagenomes.*

Based on the partitioning of lake microbial metabolism into SMGs, we sought to determine which KOs and associated pathways differentiated the six SMGs. We calculated a KO index that quantified the specificity of a KO for each of the six SMGs. This index ranges from -1 (a KO is not found in a SMG and equally represented in all the other SMGs), to 0 (a KO is equally represented across all SMGs), to 1 (a KO is represented in only one SMG). In addition, using the KO index allowed us to calculate a median index per KO pathway and KO modules for each of the six SMGs.

The distribution of KO indices was plotted for each of the six SMGs with the KEGG pathways with the top 50 median indices for each SMG; we detected that KEGG pathways with highest indices in the water column for SMG-1 were assigned to xenobiotics biodegradation and metabolism, specifically nitrotoluene degradation (KO00633) and atrazine degradation (KO00791) (**Figure 4.9.1**) which were associated with hypereutrophic lakes located in Alberta, Saskatchewan and Manitoba.

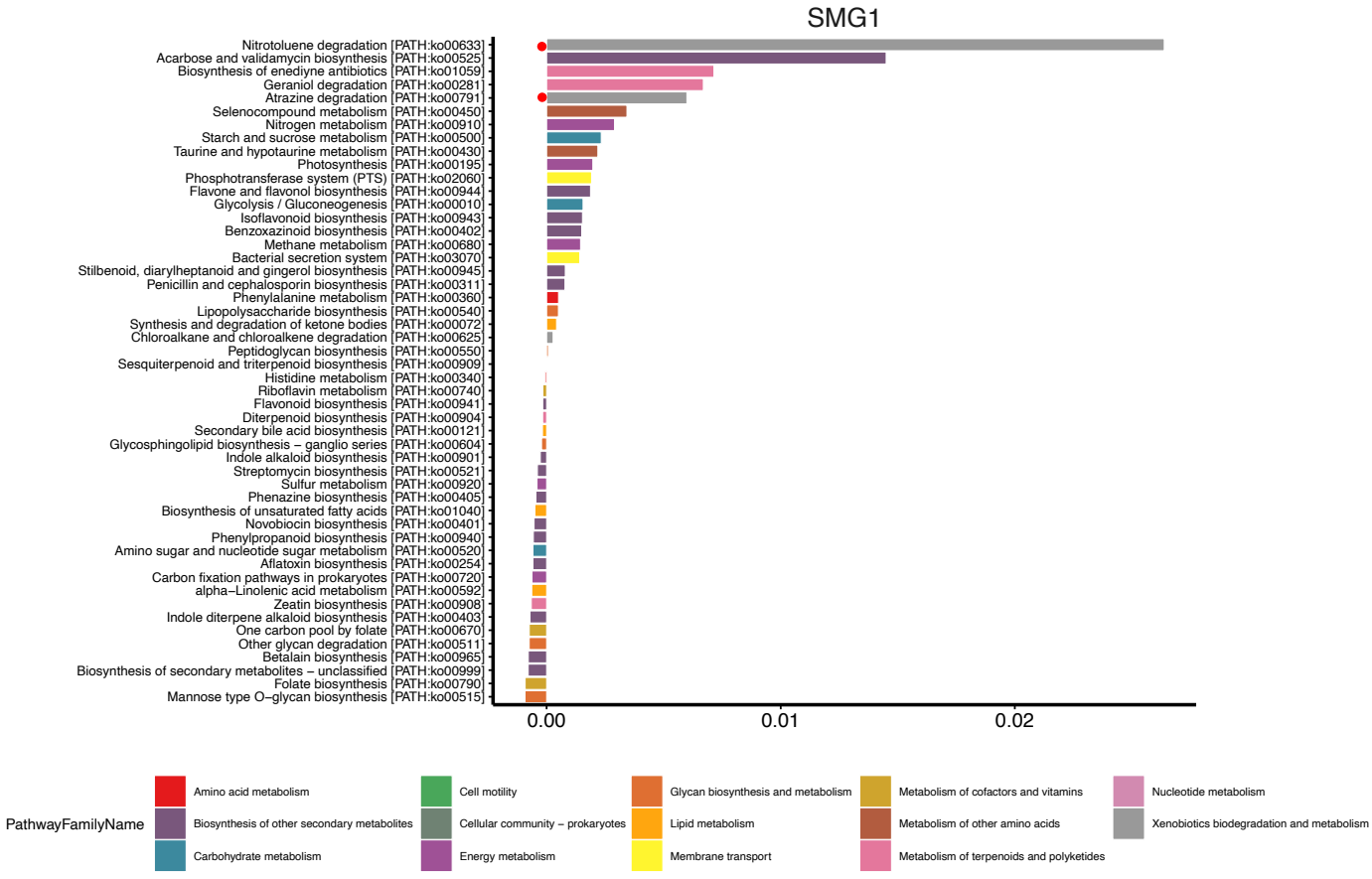


Figure 4.9.1 *Fifty highest ranking median indices for KEGG pathways for SMG-1. Red dots indicate some highest-ranking pathways.*

In SMG-2, we noticed an enrichment for amino acid metabolism (KO00472/3) and furfural degradation (KO00365). These KEGG pathways had the highest indices and lakes associated with these pathways were broadly found across ecozones, especially in the Boreal Shield within lakes of low to intermediate trophic states (oligotrophic, mesotrophic and mesoeutrophic lakes) (**Figure 4.9.2a**). The KEGG pathways with the highest indices in SMG-3 were associated with drug biosynthesis (Acarbose and Validamycin, KO00525) as well as photosynthesis (KO00195). We identified a single hypereutrophic lake in the Montane Cordillera ecozone contributing the most to SMG-3 and could be the reason for the detection of photosynthetic signal within the SMG (**Figure 4.9.2b**).

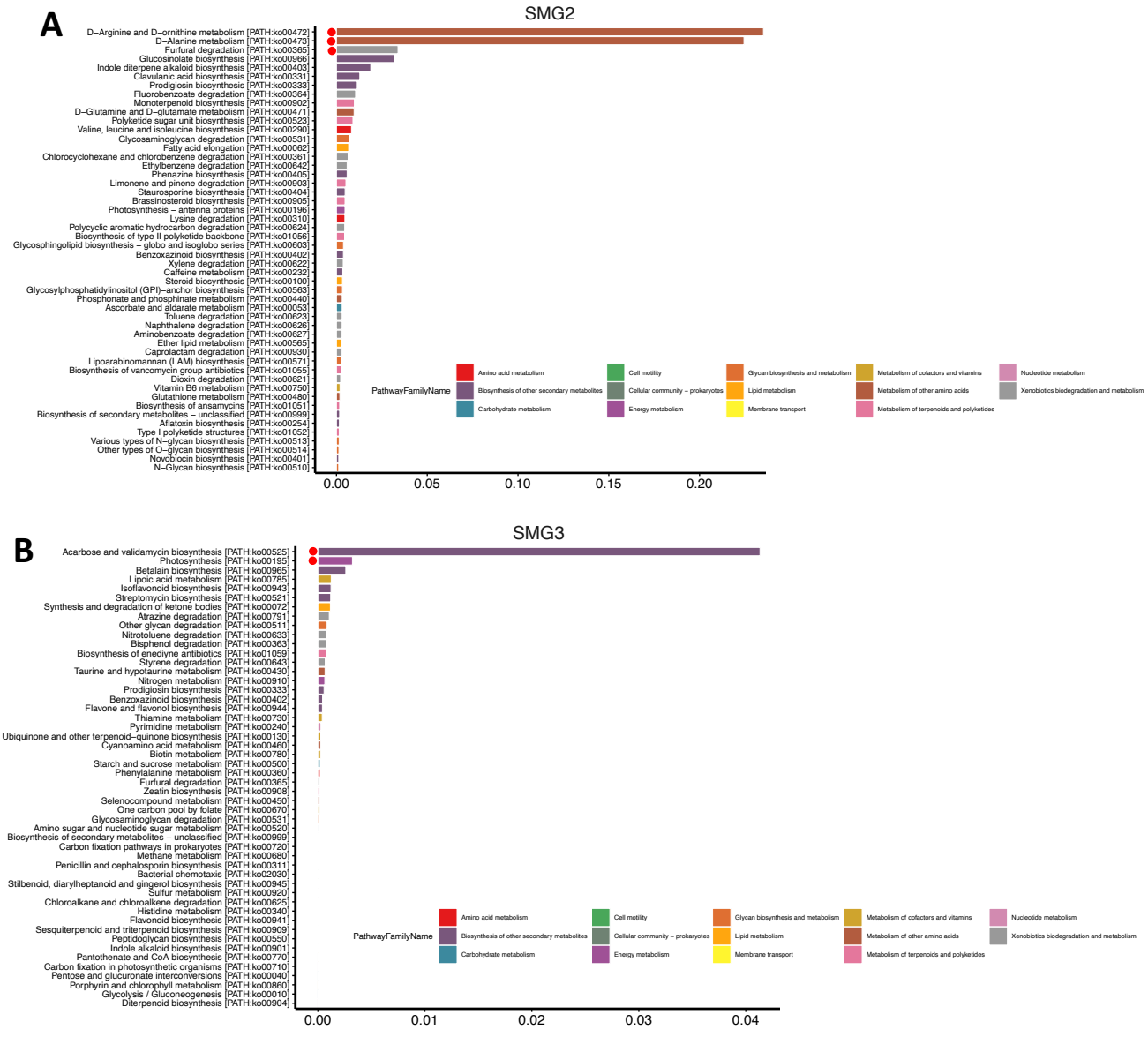


Figure 4.9.2 A) *Fifty highest ranking median indices for KEGG pathways for SMG-2. B) Fifty highest ranking median indices for KEGG pathways for SMG-3. Red dots indicate highest-ranking pathways.*

Interestingly, KEGG pathways with the highest indices in SMG-5 were involved in photosynthesis. Lakes associated with this photosynthesis signal were found across more ecozones including the Prairies, Semi-Arid Plateaux, Boreal Plains and others. These suites of lakes contributing to SMG-5 were mostly hypereutrophic or eutrophic (**Figure 4.9.3a**). SMG-6 was enriched for xenobiotics biodegradation and metabolism, including furfural degradation (KO00365); fluorobenzoate degradation (KO00364); bisphenol degradation (KO00363); atrazine degradation (KO00791) and aminobenzoate degradation (KO00627) (**Figure 4.9.3b**) but associated with mesoeutrophic to mesotrophic lakes located within Quebec, British Columbia, Nova Scotia and New Brunswick.

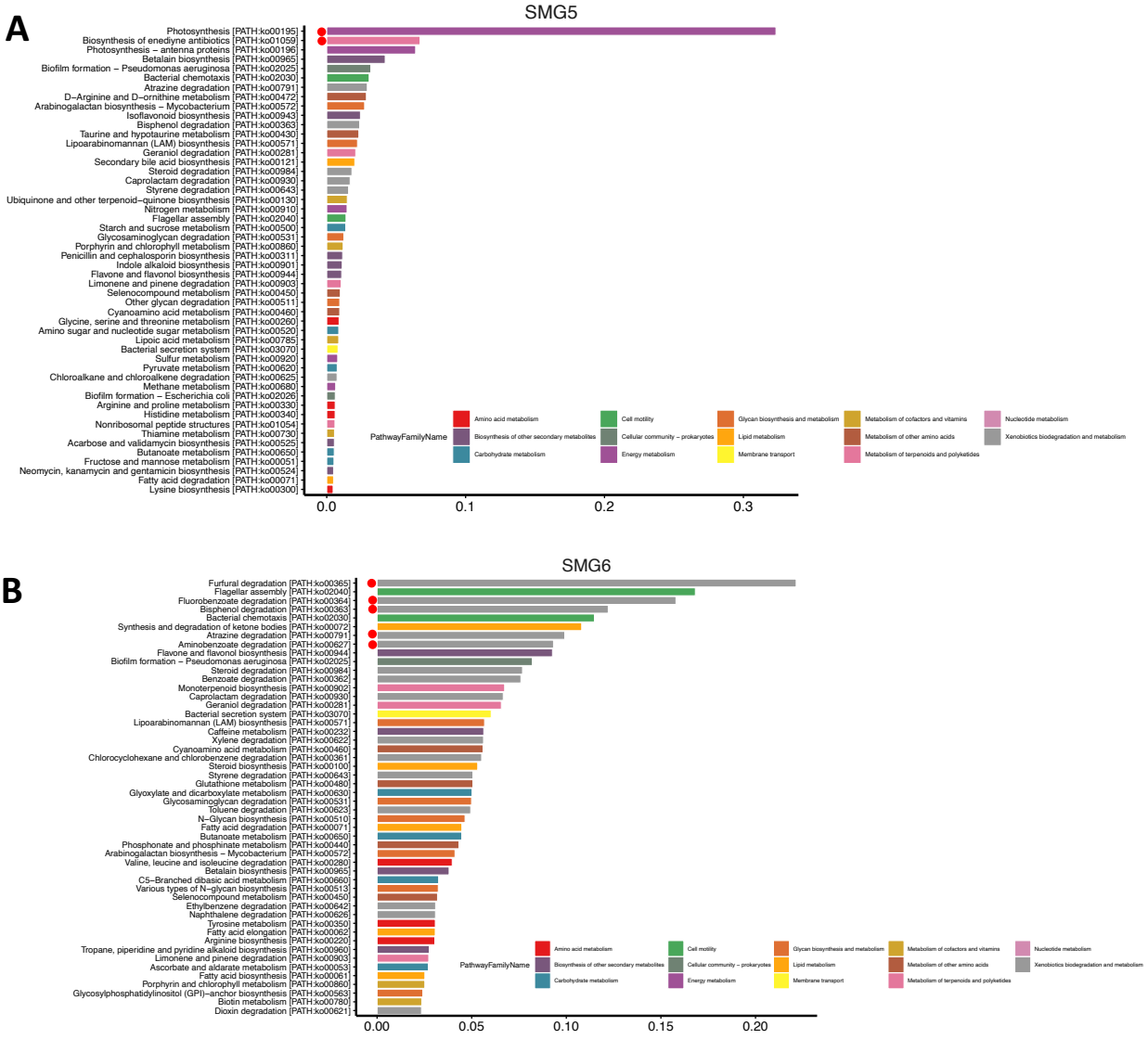


Figure 4.9.3 A) *Fifty highest ranking median indices for KEGG pathways for SMG-5. B) Fifty highest ranking median indices for KEGG pathways for SMG-6. Red dots indicate some highest ranking pathways.*

Finally, to associate the metabolic pathways enriched in SMGs to specific watershed land use conditions, we investigated land use types associated with metabolisms across SMGs using a random forest (RF) modelling approach. Overall models explained a range of variation for SMGs (40-77 % variation explained) (**Supplementary Table S4**). The strengths of land use types within models differed in their contribution to the SMGs. To explore land use type importance, we categorized variables as either strong (=>5% variation explained), intermediate (2-5%) or low (<2%) predictors of metabolism.

Our RF analyses revealed that all SMGs were relatively influenced by different land use types. We observed that all SMGs except 3 and 6 were predicted by a similar land use type – agriculture. In SMG-1, agriculture strongly predicted 5.8% of variation; 3.9% in SMG 2; 4.5% in SMG-4 (at intermediate levels); and the lowest amount of variation in SMG-5 (1.5%). In SMG-3, pasture predicted 4.6% of variation in metabolism, while for SMG-6, both natural landscapes and agriculture (potentially orthogonal) was detected, predicting 2.4 and 1.8% of variation in enriched metabolism respectively. (**Supplementary Table S4**).

4.4 Discussion

In this chapter, we expand beyond asking "which taxa are in these lakes" to elucidating functional capabilities, that is "what could these bacterial groups do". The LakePulse metagenome resource represents an exploration of bacterial diversity, taxonomic composition, and functional composition across a wide range of Canadian lakes. This chapter's breadth, covering 357 lakes across 12 ecozones, provides comprehensive insight into the spatial variation in bacterial communities in relation to environmental gradients and watershed land use influences. Overall, we detected that the influence of lake physicochemical properties on both taxonomic and functional composition of lake bacteria surpasses other environmental conditions but land use, specifically agriculture, influences both taxonomic and functional similarly. A major highlight of this chapter was the enrichment of metagenomes with genes coding for xenobiotics metabolism and the susceptibility of this bacterial metabolism to both environmental and land use influences, wherein physicochemical parameters (like lake pH, total nitrogen concentration) as well as land use types (like agriculture and pasture) strongly influenced the capability of lake bacteria to degrade xenobiotics in lakes.

4.4.1 Bacterial richness showed unimodal patterns with peaks at intermediate trophic state

This chapter revealed that bacterial richness varied across ecozones but more specifically, greater richness was observed in the Prairies and Boreal Plains ecozones; across nutrient gradients, bacterial richness peaked at intermediate trophic states (mesoeutrophic to mesotrophic). One central goal of aquatic microbial ecology is to understand the mechanisms that generate and maintain biodiversity in aquatic ecosystems. Previous research has established a link between freshwater bacterioplankton richness and nutrient availability wherein freshwater bacterioplankton richness in fourteen nutrient-poor lakes was positively influenced by nutrient availability (Logue et al., 2012). This finding, as well as ours, remains consistent with others that demonstrate the role of nutrient availability as a major driver of species richness; a pattern that may be universally valid for both micro- and macro-organisms, not only in lakes but also in soils (Das Gupta & Pinno, 2020; Garrido-Sanz et al., 2023; Long et al., 2018). The peak in bacterial richness observed at intermediate trophic states in our study aligns with the species-energy theory (SET), which suggests that population sizes of resident species increase with nutrient availability, thereby reducing the risk of extinction of rare species (Cardinale et al., 2009; Wright, 1983).

Similar positive relationships between nutrient availability and phytoplankton species richness have been demonstrated in Norwegian lakes (Cardinale et al., 2009) and relationships with bacterio-, phyto-, and zooplankton species richness in Finnish lakes (Korhonen et al., 2011), as well as in a meta-analysis of freshwater studies (Lewandowska et al., 2016). In Canadian lakes located in the Province of Quebec, unimodal relationships between zooplankton species richness and TP have been previously established (Barnett & Beisner, 2007). It is noteworthy though, that the relationship with nutrient availability may vary depending on the trophic state of aquatic ecosystems, ranging from a positive linear relation in ultraoligotrophic systems to a negative linear relation in eutrophic systems (Korhonen et al., 2011). This variability may explain the unimodal pattern in bacterial richness observed in our lakes (like findings with zooplankton species richness in Canadian lakes reported by Barnett and Beisner), where richness was highest at intermediate trophic levels (mesoeutrophic to mesotrophic) but relatively tapered at lower (ultraoligotrophic to oligotrophic) and higher (hypereutrophic to eutrophic) trophic states. These patterns in describing species richness in relation with nutrient availability align with local-scale ecological models which predict that productivity–diversity relationship should be hump-shaped (i.e. initially increase with slight increases in productivity, but then decline to low levels when productivity is highest) (Chase & Leibold, 2002).

The unimodal pattern in bacterial richness observed in this chapter could also be attributed to the higher throughput method which differed from what was used in previous chapters. Specifically, in the preceding chapter where amplicon sequencing technique was employed, peaks in bacterial richness at intermediate trophic states were not observed but bacterial richness was highest in lakes with higher nutrient concentrations within the Prairies and Boreal Plains ecozone. This could be due to the presence of multiple copies of the amplicon (16S rRNA) gene in copiotrophs, which is one of the challenges of the amplicon gene method (Crosby & Criddle, 2003). Furthermore, research has demonstrated that larger scale often shows a very different pattern from those performed at local scale wherein species diversity often monotonically increases with increasing productivity due to system heterogeneity or the use of different methodology (Chase & Leibold, 2002).

4.4.2 An unexpected abundance of Firmicutes in high-nutrient lakes

The dominance of phyla such as Actinobacteria, Proteobacteria, and Bacteroidetes found in this chapter aligns with previous findings on the taxonomic composition in the water column of Canadian freshwater environments (Kraemer et al., 2020; Paver et al., 2020; Sadeghi et al., 2021; Shahraki et al., 2021). However, the unexpected abundance of Firmicutes in eutrophic to hypereutrophic lakes within the Prairies and Boreal Plains is notable. Firmicutes have been found in diverse environments including the air (Gusareva et al., 2019; Lang-Yona et al., 2022), human gut (Hou et al., 2022; King et al., 2019; Rinninella et al., 2019) , oceans (Orcutt et al., 2011; Lang-Yona et al., 2022) but are most usually abundant in soils (Kuramae et al., 2012; W. Li et al., 2020; Mhete et al., 2020; Sheibani et al., 2013), and rarely reported in freshwater lakes (Shahraki et al., 2021). They are often associated with high nutrient environments and can indicate potential anthropogenic impact, such as agricultural runoff (Hashmi et al., 2020). Researchers have demonstrated that these group of organisms, generally known as copiotrophs (Song et al., 2016), may possess diverse metabolic capabilities including high-nutrient organic matter degradation, and have been implicated as bacterial keystone biomarkers in agricultural soils (Wongkiew et al., 2022). The remarkable discovery of Firmicutes in nutrient-rich lakes of the Prairies and Boreal Plains could be attributed to the fact that bacterial composition in these lakes may have shifted from oligotrophic taxa to high nutrient-tolerant taxa. A metagenomic exploration of bacterial communities in mesotrophic Washademoak Lake (median TP of 17 $\mu g/L^{-1}$) in Atlantic Canada revealed findings similar to our continental study, with bacterial communities dominated by Proteobacteria, Actinobacteria and Cyanobacteria but with other lineages including Firmicutes at a relative abundance of about 19 % (Valadez-Cano et al., 2022). It is

noteworthy that while Washademoak Lake may be classified as mesotrophic, it has experienced cyanobacterial blooms, and when assessed for the presence of potential cyanotoxin, well-known bloom-forming genera (*Microcystis*, *Nostoc*, and *Dolichospermum*) were prevalent in Washademoak. Also, assembly and binning strategies recovered the genome of *Microcystis aeruginosa* WS75, a toxin-producing cyanobacterium from Washademoak Lake. Based on this, and the fact that environmental conditions in Atlantic Canada differ from those in the Prairies (the latter generally having much higher phosphorus concentrations), the presence of Firmicutes in our lakes may be thus comparable with findings in Washademoak Lake.

4.4.3 Spatial variation in taxonomic and functional composition is driven by lake physicochemistry

Research has demonstrated the importance of lake physicochemistry in driving both bacterial community composition and function in hypereutrophic lakes (Díaz-Torres et al., 2022). Both our taxonomic and functional PCA analysis highlighted spatial dispersion of high nutrient lakes in the Prairies and Boreal Plains, suggesting that these regions may have distinct bacterial communities compared to others. The overall variation in functional composition was higher than taxonomic composition (PC1 taxonomy, 6.81%; PC1 function, 19.6% variation captured), which may suggest that environmental conditions influence the functional capacity of bacterial communities more strongly than their taxonomic makeup. However, the PCAs could only tell us about the underlying structure in both datasets which may be skewed by the influence of lakes in the Prairies and Boreal Plains ecozones. Upon further investigation of this trend using constrained ordinations (db-RDA) with a forward selection step, we found that when we disentangled environmental conditions into six broad categories (namely physicochemistry, lake morphometry, land use, geography, climate and watershed soil properties), both taxonomic composition and function exhibited some similarity with the overlying structure unveiled by our PCA analysis. However, they differed from model to model. In half (3 out of 6) of the models investigated, environmental conditions influenced the functional capacity of bacterial communities more strongly than their taxonomic composition. In particular, lake morphometry, weather, and geography explained more variation in functional composition than taxonomy. For the other half of the db-RDAs, more variation in taxonomic composition was explained by environment (physicochemistry, watershed soil properties and land use) than for function. Overall, our db-RDA analysis indicated that physicochemistry was the primary driver of variation in both taxonomic composition and function. In particular, pH, TN, chlorophyll-*a* and potassium

concentration were significant predictors, emphasizing the importance of water chemistry in shaping bacterial communities and their metabolic potential. In terms of watershed land use, agriculture was the most important predictor of both taxonomic composition and function. Interestingly, more broad-scale geographical factors were related to function, while more local lake factors affected taxonomic composition. This is what might be expected given that function transcends taxonomy and provides a "common currency" with which to evaluate community structure across regions where different taxa are present owing to biogeography (McGill et al., 2006).

4.4.4 Enrichment of xenobiotics biodegradation and metabolism revealed in sub-metagenomes

Metagenomic studies have shown functional differences in communities across land use (Garner et al., 2023), eutrophication or pollution gradients. The enrichment of xenobiotics degradation and metabolism pathways in specific sub-metagenomes, particularly in lakes with high nutrient levels, highlights the potential for these environments to influence the breakdown of pollutants. This may have implications for water quality and ecosystem health, particularly in areas with significant agricultural activity. Xenobiotics are synthetic chemicals from anthropogenic sources that do not or rarely exist as natural products (Rieger et al., 2002). In a previous study using metagenome assembled genomes (MAGs) across Canadian lakes, an exploration of specific metabolic categories revealed that xenobiotics metabolism was most strongly explained by land use and its turnover within xenobiotic metabolism was mostly impacted by human population density and agriculture (Garner et al., 2023).

Similarly, in hypereutrophic lakes located in Europe, metagenomics was used to investigate the microbial communities and their functional potential in surface sediments collected from three lakes of differing trophic states (mesotrophic, eutrophic and supertrophic). The researchers reported that xenobiotic pathways, such as those involving polycyclic aromatic hydrocarbons, were highest in the lakes with the greatest agricultural land use in their catchment (Biessy et al., 2022). In our study, we also found that agriculture was the dominant driver of xenobiotic metabolism in lakes. Though Biessy and colleagues' investigation was based on surface sediment of freshwater lakes and not the water column, interactions between bed sediment and the water column in shallow lakes may have been demonstrated to influence the availability and transport of nutrients (Julian II et al., 2023). Our random forest analysis provided insight into how specific land use variables predict the metabolic variation in bacterial communities; for instance, agriculture was a significant predictor for SMGs

enriched with xenobiotic metabolism. This finding underscores the interconnectedness of lake ecosystems with their surrounding landscapes.

4.5 Conclusion

The LakePulse metagenome resource provides a comprehensive view of the bacterial communities in Canadian lakes, highlighting the complex interplay between environmental factors, land use, and bacterial diversity, community composition and function. Our findings underscore the critical importance of understanding how watershed land use practices and environmental conditions influence the functioning of freshwater ecosystems. Specifically, the identification of links between bacterial metabolic pathways, such as xenobiotics biodegradation, with lake physicochemistry and agricultural activities within the watershed suggests a pervasive impact of land use on the functional diversity of lake bacterial communities.

However, it's important to note that our study did not include metatranscriptomic investigations, which would have provided insights into the active expression of genes associated with xenobiotics biodegradation or other metabolisms within our lakes. Therefore, while we can speculate on the potential of lake bacteria to carry out these metabolic processes, we cannot conclude on whether these genes are actively expressed in the studied ecosystems. Addressing this gap in knowledge would be crucial for gaining a more comprehensive understanding of freshwater ecosystems across Canada and promoting their long-term sustainability. Therefore, future research endeavors could focus on investigating gene expression within these lake environments, allowing for a more direct assessment of how bacterial communities respond to changing environmental conditions and land use influence.

4.6 Methods

4.6.1 Lake selection and sampling

Over 600 lakes were samples within three summers from 2017 to 2019 by the Natural Sciences and Engineering Research Council (NSERC) Canadian Lake Pulse Network (Huot et al., 2019). Lakes were randomly selected across three lake area categories (small 0.1-1, medium 1- 10, and large 10-100 km^2) and varying human impact categories (low, moderate, and high characterised by land use types and coverage of the watershed) across 12 ecozones of Canada in order to adequately capture the wide environmental heterogeneity in lake and watershed conditions at the continental scale.

Water for assessing bacterial communities was collected from the euphotic zone (estimated as twice the Secchi disk depth) over a depth of up to 2 m below the surface using an integrated tube sampler. Carboys were stored in ice-pack-chilled coolers until water could be filtered on the lakeshore. Water was prefiltered through 100 μm synthetic nylon mesh and vacuum-filtered on 47 mm-diameter 0.22 μm Durapore membranes through a glass funnel apparatus at a maximum pressure of 8 inHg. Filtration concluded either at 500 mL or upon clogging of the filter. Filters were stored in sterile cryovials at -80 °C. Details for environmental sampling and field protocols can be found in the NSERC Canadian Lake Pulse Network field manual 2017 - 2018 - 2019 surveys prepared by Varin and colleagues (NSERC Canadian Lake Pulse Network, 2021). 366 lakes were selected for metagenomic data generation. Nine Saline lakes, identified as having conductivity ≥ 8,000 μS/cm or total major ions ≥4,000 mg/L, were removed leaving a total of 357 lakes for taxonomic analysis. Nine lakes with the lowest numbers of protein coding genes and a montane cordillera lake (11-538) without functional annotations were further excluded for functional analysis leaving a total of 347.

4.6.2 Environmental data

Six categories of environmental explanatory variables were selected for analysis: (1) geography, (2) lake morphometry, (3) physicochemical parameters, (4) watershed surface soil properties, (5) land use, and (6) climate. Geography variables included latitude, longitude, and altitude. Lake morphometry parameters included lake area, circularity, volume, maximum depth, discharge, water residence time, watershed area, lake-to-watershed area ratio, and watershed slope within 100 m of the shoreline (data on volume, discharge, residence time, and slope were accessed from HydroLakes v. 1.0) (Messager et al., 2016). Physicochemical parameters included surface water

temperature, pH, colour, and concentrations of Chl-*a*, DIC, DOC, TN, TP, calcium, chloride, magnesium, potassium, sodium, and sulfate. Watershed soil properties estimated for the top 0 – 5 cm soil depth interval were accessed from SoilGrids250m (Hengl et al., 2017) and included bulk density of the fine earth fraction, cation exchange capacity, volumetric fraction of coarse fragments, proportions of clay, sand, and silt particles in the fine earth fraction, total nitrogen, pH, soil organic carbon content in the fine earth fraction, and organic carbon density. Land use variables were calculated as fractions of watershed area not covered by water and included crop agriculture, pasture, forestry, built development, human population density, livestock density, and poultry density. Climate variables measured over the seven days prior to lake sampling were accessed from ERA5-Land hourly reanalysis (Muñoz-Sabater et al., 2021) and included mean air temperature, total precipitations, mean net solar radiation, mean wind speed, and ice disappearance day for the year of sampling. Lake trophic state was categorized by TP concentrations according to the Canadian Water Quality guidelines: ultraoligotrophic (<4 μg/L), oligotrophic (4 – 10 μg/L), mesotrophic (10 – 20 μg/L), mesoeutrophic (20 – 35 μg/L), eutrophic (35 – 100 μg/L), and hypereutrophic (>100 μg/L) (Canadian Council of Ministers of the Environment, 2004).

4.6.3 DNA extraction, metagenome sequencing, assembly, and annotation

DNA was extracted from filters using the DNeasy PowerWater kit (QIAGEN) according to the manufacturer's instructions supplemented by the optional addition of 1 μL ribonuclease A and 30 min incubation at 37 °C. DNA was submitted to Genome Quebec for library preparation using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) and 150 bp paired-end shotgun sequencing on an Illumina NovaSeq 6000 platform. Adapter-clipping and quality-trimming of raw reads were performed in Trimmomatic v. 0.38 using default settings (Bolger et al., 2014). Single metagenome assemblies were created for 100 individual samples, collected in 2017, using MEGAHIT v.1.2.7 (D. Li et al., 2016) with kmer lengths 27, 37, 47, 57, 67, 77, 87 and a minimum count of two while single assemblies for the other 266 samples collected in years 2018, 2019 were created using metaSPAdes v.3.13.0 with kmers 21, 33, 55, 77, 99, 127 and the assembler_only option (Nurk et al., 2017). Fasta files, representing contigs/scaffolds for each metagenome along with coverage files were deposited at the DOE Joint Genome Institute Integrated Microbial Genomes (JGI/IMG) annotation site for functional annotation. Gene prediction and annotation was performed using the DOE Joint Genome Institute Integrated Microbial Genomes Annotation Pipeline v.4.16.6 (Huntemann et al., 2016) and v.5.0.20/23 (Chen et al., 2019) respectively.

4.6.4 Computation of gene abundance matrix

Metagenomics data files containing genes IDs, gene annotations, gene depth of coverage and other gene information were retrieved from the DOE Joint Genome Institute Integrated Microbial Genomes (JGI/IMG, https://img.jgi.doe.gov) repository. The abundance of a KEGG ortholog number (KO, gene family) in a metagenome was calculated by summing the depth of coverage of all genes annotated with that KO. KO abundance matrices therefore represent the metagenomic profiles across the samples.

4.6.5 Taxonomic annotation and estimation of bacterial richness

Taxonomic annotation was performed using SingleM, a tool used to find the abundances of discrete operational taxonomic units (OTUs) directly from shotgun metagenome data, without heavy reliance on reference sequence databases (Woodcroft et al., 2024). Raw metagenome reads, not quality trimmed reads were used. SingleM concentrated on 14 single copy marker genes (rpL2, rpL3, rpL5, rpL6, rpL11, rpL14b_L23e, rpL16_L10E, rpS2, rpS5, rpS7, rpS10, rpS12_S23, rpS15P_S13e, rpS19) to provide fine-grained differentiation of species that is independent of the copy-number variation issues that hamper 16S analyses. The ribosomal protein L2_rplB was used for taxonomic profiling. Annotated bacterial sequences were rarefied to an equal depth of 600 sequences. Bacterial richness (Chao1 index) was estimated using the *vegan* package in R (Oksanen et al., 2020).

4.6.6 Unconstrained and constrained ordinations

Principal component analysis (PCA) was conducted using the *rda* function of the R package *vegan* (Oksanen et al., 2020). Distance-based redundancy analysis (dbRDA) allows constrained ordinations on community data using non-Euclidean distance measures. Six categories of db-RDA models were established as detailed in environmental data section. A synthetic distance matrix was computed from both bacterial community and function matrices using the *vegdist* function in R, employing Bray-Curtis as the distance method to capture dissimilarity in community composition and function. Subsequently, a forward selection distance-based redundancy analysis was performed to evaluate how environmental conditions and land use influence community composition and function across the dataset. The *capscale* function of the R vegan package (Oksanen et al., 2020) was utilized for this analysis.

4.6.7 Non-Negative matrix factorization (NMF)

NMF was performed with the *nmf* function from the NMF package in R (Gaujoux & Seoighe, 2010). NMF decomposes the abundance matrix into two matrices: a coefficient matrix that describes the overall structure of the abundance matrix with a limited number of descriptors (herein referred to as sub-metagenomes, their number being the rank), and a basis matrix that provides the weights of each original descriptors (KO number) on the new descriptors (sub-metagenomes). The advantage of NMF is that it directly links the overall structure of the abundance matrix to the individual elements (KO number) driving this structure. NMF analysis was first performed with rank values ranging from 3 to 7, 100 runs, and various algorithms ("nsnmf", "Brunet", "KL"). We obtained the optimal results for the nsNMF algorithm, random seed of the factorized matrices, and an optimal rank value of 6. Final analysis was performed with 200 runs, a rank of 6, random seed and nsNMF algorithm.

4.6.8 Random Forest Analysis

RF analysis was employed to assess the impact of variables on six sub-metagenomes identified by NMF analysis. RF presents advantages over traditional regression techniques, particularly in mitigating the risk of overfitting when dealing with a large number of predictor variables, as is the case in our study (Matsuki et al., 2016; Ryo & Rillig, 2017). In this context, we utilized an RF technique based on conditional inference regression trees (Strobl et al., 2009), as developed by Ryo and Rillig (2017). Importance measures were computed for each predictor variable through cross-validation, utilizing data not employed in the tree construction, known as the out-of-bag (OOB) data (Breiman, 2001). The analysis involved the utilization of 5000 regression trees to ensure a robust prediction, implemented using the *party* package in R (Horton et al., 2019; Ryo & Rillig, 2017; Strobl et al., 2007; Zeileis et al., 2008)

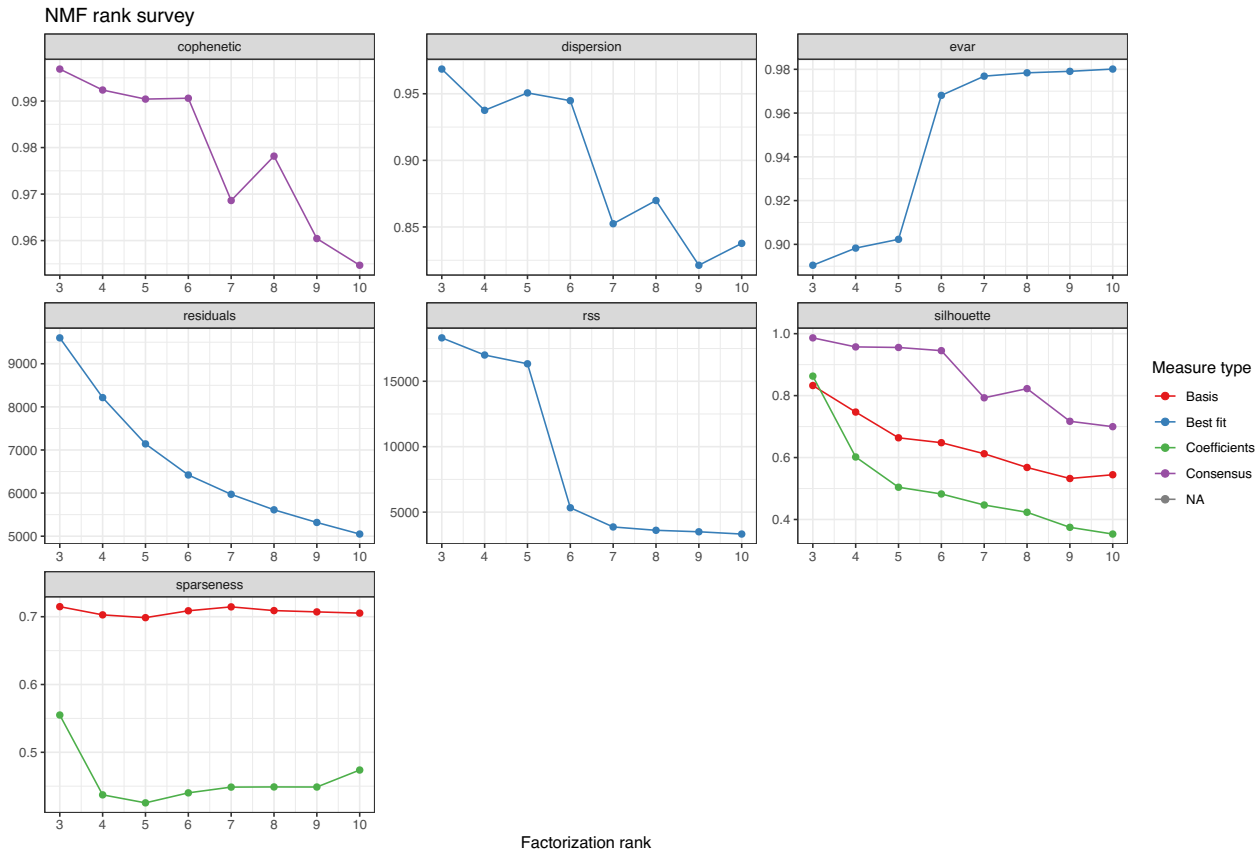
4.7 Data avaliability

Raw metagenome reads were archived in the European Nucleotide Archive under study accession PRJEB29238 (https://www.ebi.ac.uk/ena/browser/view/PRJEB29238). Single Metagenome assemblies were deposited and annotated at the Joint Genome Institute Genomes OnLine Database under study accession Gs0136026.

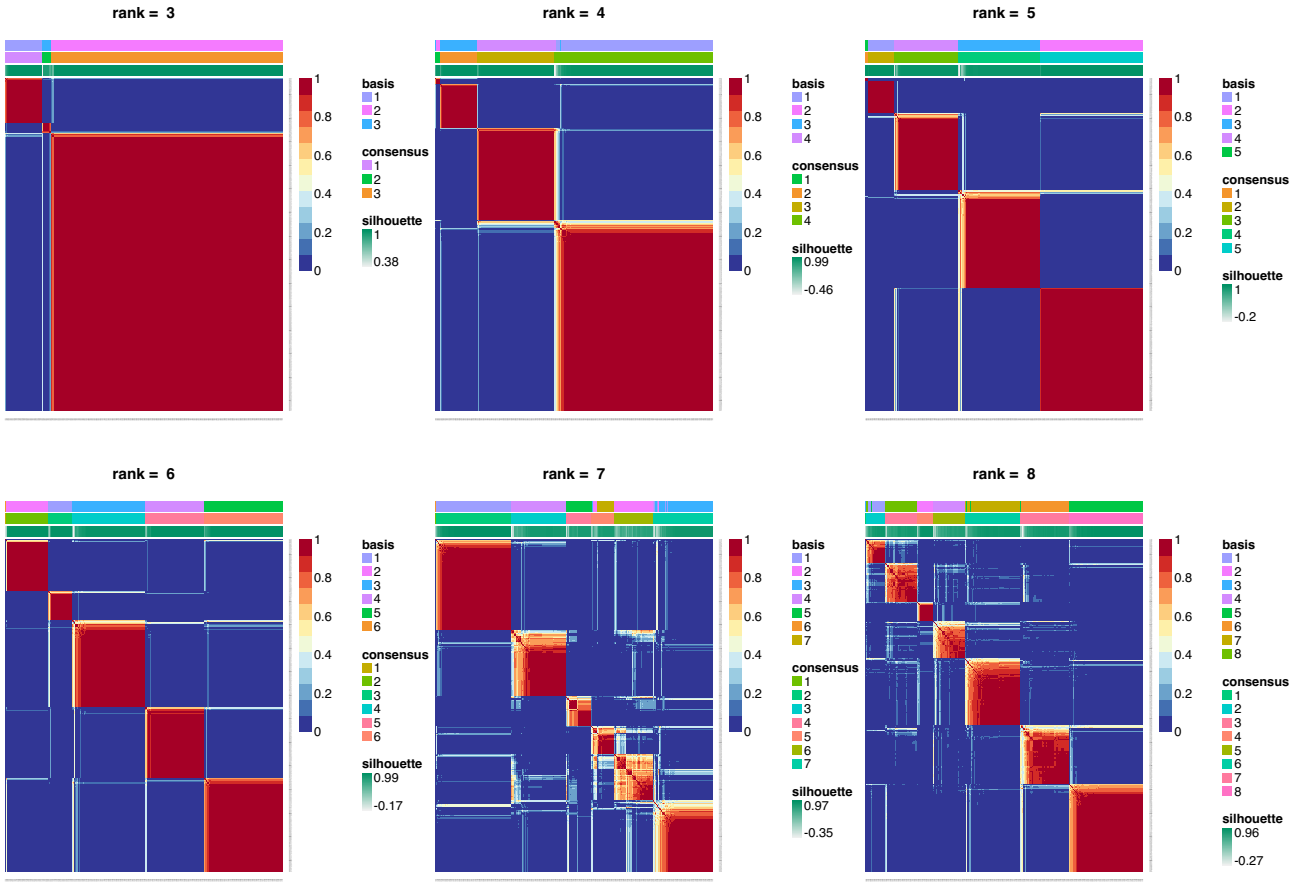
4.8 Funding and acknowledgement

This chapter was funded by the NSERC Canadian LakePulse Network (Strategic Network Grant NETGP-479720). I received scholarship from the NSERC CREATE ÉcoLac Training Program in Lake and Fluvial Ecology. I appreciate the LakePulse researchers involved in the generation, management, and quality control of the data set. Thanks to Wentworth Brookes, Shawn Simpson and Compute Canada for bioinformatics technical support.

4.9 Supplementary Figures and Tables



Supplementary Figure S4a *Evolution of various metrics used to quantify the optimal rank to be used for the non-negative matrix factorization analysis. Cophenetic correlation represents the correlation between the sample distances from the consensus matrix and the cophenetic distance between these samples when they are clustered. The dispersion is defined as 1-rss/^i,j(Vi,j)2 (Vi,j are the entries of the KO abundance matrix). Evar estimates the fraction of variance of the KO abundance matrix explained by the NMF results. Residuals is the sum of residuals between the original KO abundance matrix and the matrix estimated using the NMF. The rss is the residual sum of squares between the original KO abundance matrix and its estimate using the NMF algorithm. Sparseness is equal to 1 if all the elements of a vector are null but for 1. Oppositely, the sparseness is equal to 0 is all the element of a vector are equals. The sparseness of the basis and coefficient matrices are calculated as the mean sparseness of its element vectors.*



Supplementary Figure S4b *Consensus matrices obtained from non-negative matrix factorization of lake KO abundance matrix using various rank values and 100 runs.*

Supplementary Table S4 *Summary of random forest result on six submetagenomes. The most influential predictors are highlighted as well as the overall RF model fit and out of bag error. Strong (=>5%), intermediate (2-5%) or low (<2%) predictors of bacterial metabolism.*

Predictor	SMG-1	SMG-2	SMG-3	SMG-4	SMG-5	SMG-6
R^2 fitted	**0.77**	**0.65**	**0.40**	**0.66**	**0.52**	**0.64**
R^2 OOB	**0.63**	**0.30**	**8.07E-05**	**0.12**	**0.16**	**0.35**
Agriculture	**5.82**	**3.89**	0.59	**4.46**	**1.53**	**1.80**
Built development	0.30	0.47	1.03	1.05	0.07	0.61
Forestry	0.02	0.60	0.00	0.21	0.21	0.34
Natural landscapes	2.06	1.57	0.00	1.85	0.36	**2.36**
Pasture	0.09	1.10	**4.58**	**2.24**	0.88	1.16
Livestock density	0.92	0.28	0.11	0.61	0.42	0.43
Population density	0.06	0.82	0.17	1.03	0.34	0.14
Poultry density	0.45	0.04	0.26	0.01	0.01	0.16

Bibliography

Aldas-Vargas, A., Poursat, B. A. J., & Sutton, N. B. (2022). Potential and limitations for monitoring of pesticide biodegradation at trace concentrations in water and soil. *World Journal of Microbiology and Biotechnology*, *38*(12), 240. https://doi.org/10.1007/s11274-022-03426-x

Barnett, Allain., & Beisner, B. E. (2007). Zooplankton Biodiversity and Lake Trophic State: Explanations Invoking Resource Abundance and Distribution. *Ecology*, *88*(7), 1675–1686. https://doi.org/10.1890/06-1056.1

Biessy, L., Pearman, J. K., Waters, S., Vandergoes, M. J., & Wood, S. A. (2022). Metagenomic insights to the functional potential of sediment microbial communities in freshwater lakes. *Metabarcoding and Metagenomics*, *6*, e79265. https://doi.org/10.3897/mbmg.6.79265

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32. https://doi-org.lib-ezproxy.concordia.ca/10.1023/A:1010933404324

CCEA. (2016). Ecozone introduction. de http://www.ccea.org/ecozones-introduction/

Canadian Council of Ministers of the Environment. (2004). Canadian water quality guidelines for the protection of aquatic life: Phosphorus: Canadian Guidance Framework for the Management of Freshwater Systems. In *Canadian environmental quality guidelines*.

Cardinale, B. J., Hillebrand, H., Harpole, W. S., Gross, K., & Ptacnik, R. (2009). Separating the influence of resource 'availability' from resource 'imbalance' on productivity–diversity relationships. *Ecology Letters*, *12*(6), 475–487. https://doi.org/10.1111/j.1461-0248.2009.01317.x

Chase, J. M., & Leibold, M. A. (2002). Spatial scale dictates the productivity–biodiversity relationship. *Nature*, *416*(6879), 427–430. https://doi.org/10.1038/416427a

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., Seshadri, R., Smirnova, T., Kirton, E., Jungbluth, S. P., Woyke, T., Eloe-Fadrosh, E. A., Ivanova, N. N., & Kyrpides, N. C. (2019). IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research*, *47*(D1), D666–D677. https://doi.org/10.1093/nar/gky901

Cotner, J. B., & Biddanda, B. A. (2002). Small Players, Large Role: Microbial Influence on Biogeochemical Processes in Pelagic Aquatic Ecosystems. *Ecosystems*, *5*(2), 105–121. https://doi.org/10.1007/s10021-001-0059-3

Crosby, L. D., & Criddle, C. S. (2003). Understanding Bias in Microbial Community Analysis Techniques due to rrn Operon Copy Number Heterogeneity. *BioTechniques*, *34*(4), 790–802. https://doi.org/10.2144/03344rr01

Das Gupta, S., & Pinno, B. D. (2020). Drivers of understory species richness in reconstructed boreal ecosystems: A structural equation modeling analysis. *Scientific Reports*, *10*(1), 11555. https://doi.org/10.1038/s41598-020-68353-z

de Souza, M. L., Seffernick, J., Martinez, B., Sadowsky, M. J., & Wackett, L. P. (1998). The Atrazine Catabolism Genes atzABC Are Widespread and Highly Conserved. *Journal of Bacteriology*, *180*(7), 1951–1954.

Díaz-Torres, O., Lugo-Melchor, O. Y., de Anda, J., Orozco-Nunnelly, D. A., Gradilla-Hernández, M. S., & Senés-Guerrero, C. (2022). Characterizing a subtropical hypereutrophic lake: From physicochemical variables to shotgun metagenomic data. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.1037626

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z.-I., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A.-H., Soto, D., Stiassny, M. L. J., & Sullivan, C. A. (2006). Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biological Reviews*, *81*(2), 163–182. https://doi.org/10.1017/S1464793105006950

Edge, T. A., Baird, D. J., Bilodeau, G., Gagné, N., Greer, C., Konkin, D., Newton, G., Séguin, A., Beaudette, L., Bilkhu, S., Bush, A., Chen, W., Comte, J., Condie, J., Crevecoeur, S., El-Kayssi, N., Emilson, E. J. S., Fancy, D.-L., Kandalaft, I., … Macklin, J. (2020). The Ecobiomics project: Advancing metagenomics assessment of soil health and freshwater quality in Canada. *Science of The Total Environment*, *710*, 135906. https://doi.org/10.1016/j.scitotenv.2019.135906

Feng, G., Xie, T., Wang, X., Bai, J., Tang, L., Zhao, H., Wei, W., Wang, M., & Zhao, Y. (2018). Metagenomic analysis of microbial community and function involved in cd-contaminated soil. *BMC Microbiology*, *18*(1), 11. https://doi.org/10.1186/s12866-018-1152-5

Garcia-Moreno, J., Harrison, I. J., Dudgeon, D., Clausnitzer, V., Darwall, W., Farrell, T., Savy, C., Tockner, K., & Tubbs, N. (2014). Sustaining Freshwater Biodiversity in the Anthropocene. In A. Bhaduri, J. Bogardi, J. Leentvaar, & S. Marx (Eds.), *The Global Water System in the Anthropocene: Challenges for Science and Governance* (pp. 247–270). Springer International Publishing. https://doi.org/10.1007/978-3-319-07548-8_17

Garner, R. E., Gregory-Eaves, I., & Walsh, D. A. (2020). Sediment Metagenomes as Time Capsules of Lake Microbiomes. *mSphere*, *5*(6), e00512-20. https://doi.org/10.1128/mSphere.00512-20

Garner, R. E., Kraemer, S. A., Onana, V. E., Fradette, M., Varin, M.-P., Huot, Y., & Walsh, D. A. (2023). A genome catalogue of lake bacterial diversity and its drivers at continental scale. *Nature Microbiology*, *8*(10), Article 10. https://doi.org/10.1038/s41564-023-01435-6

Garrido-Sanz, D., Čaušević, S., Vacheron, J., Heiman, C. M., Sentchilo, V., van der Meer, J. R., & Keel, C. (2023). Changes in structure and assembly of a species-rich soil natural community with contrasting nutrient availability upon establishment of a plant-beneficial Pseudomonas in the wheat rhizosphere. *Microbiome*, *11*(1), 214. https://doi.org/10.1186/s40168-023-01660-5

Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, *11*(1), 367. https://doi.org/10.1186/1471-2105-11-367

Grossart, H.-P., Massana, R., McMahon, K. D., & Walsh, D. A. (2020). Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnology and Oceanography*, *65*(S1), S2–S20. https://doi.org/10.1002/lno.11382

Gusareva, E. S., Acerbi, E., Lau, K. J. X., Luhung, I., Premkrishnan, B. N. V., Kolundžija, S., Purbojati, R. W., Wong, A., Houghton, J. N. I., Miller, D., Gaultier, N. E., Heinle, C. E., Clare, M. E., Vettath, V. K., Kee, C., Lim, S. B. Y., Chénard, C., Phung, W. J., Kushwaha, K. K., … Schuster, S. C. (2019). Microbial communities in the tropical air ecosystem follow a precise diel cycle. *Proceedings of the National Academy of Sciences*, *116*(46), 23299–23308. https://doi.org/10.1073/pnas.1908493116

Hashmi, I., Bindschedler, S., & Junier, P. (2020). Chapter 18—Firmicutes. In N. Amaresan, M. Senthil Kumar, K. Annapurna, K. Kumar, & A. Sankaranarayanan (Eds.), *Beneficial Microbes in Agro-Ecology* (pp. 363–396). Academic Press. https://doi.org/10.1016/B978-0-12-823414-3.00018-6

Hengl, T., Jesus, J. M. de, Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on

machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Horton, D. J., Theis, K. R., Uzarski, D. G., & Learman, D. R. (2019). Microbial community structure and microbial networks correspond to nutrient gradients within coastal wetlands of the Laurentian Great Lakes. *FEMS Microbiology Ecology*, *95*(4), fiz033. https://doi.org/10.1093/femsec/fiz033

Hou, K., Wu, Z.-X., Chen, X.-Y., Wang, J.-Q., Zhang, D., Xiao, C., Zhu, D., Koya, J. B., Wei, L., Li, J., & Chen, Z.-S. (2022). Microbiota in health and diseases. *Signal Transduction and Targeted Therapy*, *7*(1), 1–28. https://doi.org/10.1038/s41392-022-00974-4

Huntemann, M., Ivanova, N. N., Mavromatis, K., Tripp, H. J., Paez-Espino, D., Tennessen, K., Palaniappan, K., Szeto, E., Pillay, M., Chen, I.-M. A., Pati, A., Nielsen, T., Markowitz, V. M., & Kyrpides, N. C. (2016). The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Standards in Genomic Sciences*, *11*, 17. https://doi.org/10.1186/s40793-016-0138-x

Huot, Y., Brown, C. A., Potvin, G., Antoniades, D., Baulch, H. M., Beisner, B. E., Bélanger, S., Brazeau, S., Cabana, H., Cardille, J. A., del Giorgio, P. A., Gregory-Eaves, I., Fortin, M.-J., Lang, A. S., Laurion, I., Maranger, R., Prairie, Y. T., Rusak, J. A., Segura, P. A., … Walsh, D. A. (2019). The NSERC Canadian Lake Pulse Network: A national assessment of lake health providing science for water management in a changing climate. *Science of The Total Environment*, *695*, 133668. https://doi.org/10.1016/j.scitotenv.2019.133668

Julian II, P., Schafer, T., Cohen, M. J., Jones, P., & Osborne, T. Z. (2023). Changes in the spatial distribution of total phosphorus in sediment and water column of a shallow subtropical lake. *Lake and Reservoir Management*, *39*(3), 213–230. https://doi.org/10.1080/10402381.2023.2243606

Kanehisa, M., Sato, Y., & Kawashima, M. (2022). KEGG mapping tools for uncovering hidden features in biological data. *Protein Science: A Publication of the Protein Society*, *31*(1), 47–53. https://doi.org/10.1002/pro.4172

King, C. H., Desai, H., Sylvetsky, A. C., LoTempio, J., Ayanyan, S., Carrie, J., Crandall, K. A., Fochtman, B. C., Gasparyan, L., Gulzar, N., Howell, P., Issa, N., Krampis, K., Mishra, L., Morizono, H., Pisegna, J. R., Rao, S., Ren, Y., Simonyan, V., … Mazumder, R. (2019). Baseline human gut microbiota profile in healthy people and standard reporting template. *PLOS ONE*, *14*(9), e0206484. https://doi.org/10.1371/journal.pone.0206484

Korhonen, J. J., Wang, J., & Soininen, J. (2011). Productivity-Diversity Relationships in Lake Plankton Communities. *PLOS ONE*, *6*(8), e22041. https://doi.org/10.1371/journal.pone.0022041

Kraemer, S. A., Barbosa da Costa, N., Shapiro, B. J., Fradette, M., Huot, Y., & Walsh, D. A. (2020). A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. *The ISME Journal*, *14*(12), Article 12. https://doi.org/10.1038/s41396-020-0733-0

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 Microbiology Ecology*, *79*(1), 12–24. https://doi.org/10.1111/j.1574-6941.2011.01192.x

Lang-Yona, N., Flores, J. M., Haviv, R., Alberti, A., Poulain, J., Belser, C., Trainic, M., Gat, D., Ruscheweyh, H.-J., Wincker, P., Sunagawa, S., Rudich, Y., Koren, I., & Vardi, A. (2022). Terrestrial and marine influence on atmospheric bacterial diversity over the north Atlantic and Pacific Oceans. *Communications Earth & Environment*, *3*(1), 1–10. https://doi.org/10.1038/s43247-022-00441-6

Lewandowska, A. M., Biermann, A., Borer, E. T., Cebrián-Piqueras, M. A., Declerck, S. A. J., De Meester, L., Van Donk, E., Gamfeldt, L., Gruner, D. S., Hagenah, N., Harpole, W. S., Kirkman, K. P., Klausmeier, C. A., Kleyer, M., Knops, J. M. H., Lemmens, P., Lind, E. M., Litchman, E., Mantilla-Contreras, J., … Hillebrand, H. (2016). The influence of balanced and imbalanced resource supply on biodiversity–functioning relationship across ecosystems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1694), 20150283. https://doi.org/10.1098/rstb.2015.0283

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, *102*, 3–11. https://doi.org/10.1016/j.ymeth.2016.02.020

Li, W., Zhang, Y., Mao, W., Wang, C., & Yin, S. (2020). Functional potential differences between Firmicutes and Proteobacteria in response to manure amendment in a reclaimed soil. *Canadian Journal of Microbiology*, *66*(12), 689–697. https://doi.org/10.1139/cjm-2020-0143

Liu, J., Elliott, J. A., Wilson, H. F., Macrae, M. L., Baulch, H. M., & Lobb, D. A. (2021). Phosphorus runoff from Canadian agricultural land: A dataset for 30 experimental fields. *Data in Brief*, *38*, 107405. https://doi.org/10.1016/j.dib.2021.107405

Liu, Y., Van Nieuwenhuizen, N., Elliott, J., Shrestha, R. R., & Yerubandi, R. (2023). Runoff, sediment, organic carbon, and nutrient loads from a Canadian prairie micro-watershed under climate variability and land management practices. *Environmental Monitoring and Assessment*, *195*(11), 1285. https://doi.org/10.1007/s10661-023-11913-3

Logue, J. B., Langenheder, S., Andersson, A. F., Bertilsson, S., Drakare, S., Lanzén, A., & Lindström, E. S. (2012). Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species–area relationships. *The ISME Journal*, *6*(6), 1127–1136. https://doi.org/10.1038/ismej.2011.184

Long, C., Yang, X., Long, W., Li, D., Zhou, W., & Zhang, H. (2018). Soil Nutrients Influence Plant Community Assembly in Two Tropical Coastal Secondary Forests. *Tropical Conservation Science*, *11*, 1940082918817956. https://doi.org/10.1177/1940082918817956

Matsuki, K., Kuperman, V., & Van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading : The Official Journal of the Society for the Scientific Study of Reading*, *20*(1), 20–33. https://doi.org/10.1080/10888438.2015.1107073

McGill, B. J., Enquist, B. J., Weiher, E., & Westoby, M. (2006). Rebuilding community ecology from functional traits. *Trends in Ecology & Evolution*, *21*(4), 178–185. https://doi.org/10.1016/j.tree.2006.02.002

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms13603

Mhete, M., Eze, P. N., Rahube, T. O., & Akinyemi, F. O. (2020). Soil properties influence bacterial abundance and diversity under different land-use regimes in semi-arid environments. *Scientific African*, *7*, e00246. https://doi.org/10.1016/j.sciaf.2019.e00246

Mishra, S., Lin, Z., Pang, S., Zhang, W., Bhatt, P., & Chen, S. (2021). Recent Advanced Technologies for the Characterization of Xenobiotic-Degrading Microorganisms and Microbial Communities. *Frontiers in Bioengineering and Biotechnology*, *9*. https://www.frontiersin.org/articles/10.3389/fbioe.2021.632059

Monchamp, M.-È., Walsh, D. A., Garner, R. E., Kraemer, S. A., Beisner, B. E., Cristescu, M. E., & Gregory-Eaves, I. (2022). Comparative analysis of zooplankton diversity in freshwaters:

What can we gain from metagenomic analysis? *Environmental DNA*, *4*(6), 1250–1264. https://doi.org/10.1002/edn3.335

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), Article 4. https://doi.org/10.1038/s41587-020-0718-6

NSERC Canadian Lake Pulse Network. (2021). *NSERC Canadian Lake Pulse Network field manual 2017—2018—2019 surveys* (M.-P. Varin, M.-L. Beaulieu, & Y. Huot, Eds.). Université de Sherbrooke. https://doi.org/10.17118/11143/18662

Oksanen J., Guillaume Blanchet F., Friendly M., Kindt R., Legendre P., McGlinn D., et al. 2020. Vegan: community ecology package [software]. R package Ver. 2.5–7. Available from https://CRAN.R-project.org/package=vegan

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Orcutt, B. N., Sylvan, J. B., Knab, N. J., & Edwards, K. J. (2011). Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. *Microbiology and Molecular Biology Reviews*, *75*(2), 361–422. https://doi.org/10.1128/mmbr.00039-10

Paver, S. F., Newton, R. J., & Coleman, M. L. (2020). Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environmental Microbiology*, *22*(1), 433–446. https://doi.org/10.1111/1462-2920.14862

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), Article 9. https://doi.org/10.1038/nbt.3935

Rey-Romero, D. C., Domínguez, I., & Oviedo-Ocaña, E. R. (2022). Effect of agricultural activities on surface water quality from páramo ecosystems. *Environmental Science and Pollution Research International*, *29*(55), 83169–83190. https://doi.org/10.1007/s11356-022-21709-6

Rieger, P.-G., Meier, H.-M., Gerle, M., Vogt, U., Groth, T., & Knackmuss, H.-J. (2002). Xenobiotics in the environment: Present and future strategies to obviate the problem of biological persistence. *Journal of Biotechnology*, *94*(1), 101–123. https://doi.org/10.1016/S0168-1656(01)00422-9

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms*, *7*(1), 14. https://doi.org/10.3390/microorganisms7010014

Rodríguez-Gijón, A., Hampel, J. J., Dharamshi, J., Bertilsson, S., & Garcia, S. L. (2023). Shotgun metagenomes from productive lakes in an urban region of Sweden. *Scientific Data*, *10*(1), 810. https://doi.org/10.1038/s41597-023-02722-x

Ryo, M., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, *8*(11), e01976. https://doi.org/10.1002/ecs2.1976

Sadeghi, J., Chaganti, S. R., Shahraki, A. H., & Heath, D. D. (2021). Microbial community and abiotic effects on aquatic bacterial communities in north temperate lakes. *Science of The Total Environment*, *781*, 146771. https://doi.org/10.1016/j.scitotenv.2021.146771

Shahraki, A. H., Chaganti, S. R., & Heath, D. (2021). Spatio-temporal dynamics of bacterial communities in the shoreline of Laurentian great Lake Erie and Lake St. Clair's large freshwater ecosystems. *BMC Microbiology*, *21*, 253. https://doi.org/10.1186/s12866-021-02306-y

Sheedy, C., Kromrey, N., Nilsson, D., & Armitage, T. (2019). From peaks to prairies: A time-of-travel synoptic survey of pesticides in watersheds of southern Alberta, Canada. *Inland Waters*, *9*(4), 438–452. https://doi.org/10.1080/20442041.2019.1634947

Sheibani, S., Yanni, S. F., Wilhelm, R., Whalen, J. K., Whyte, L. G., Greer, C. W., & Madramootoo, C. A. (2013). Soil bacteria and archaea found in long-term corn (Zea mays L.) agroecosystems in Quebec, Canada. *Canadian Journal of Soil Science*, *93*(1), 45–57. https://doi.org/10.4141/cjss2012-040

Song, W., Kim, M., Tripathi, B. M., Kim, H., & Adams, J. M. (2016). Predictable communities of soil bacteria in relation to nutrient concentration and successional stage in a laboratory culture experiment. *Environmental Microbiology*, *18*(6), 1740–1753. https://doi.org/10.1111/1462-2920.12879

Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., van Eijk, R., Schleper, C., Guy, L., & Ettema, T. J. G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), Article 7551. https://doi.org/10.1038/nature14447

Štefanac, T., Grgas, D., & Landeka Dragičević, T. (2021). Xenobiotics—Division and Methods of Detection: A Review. *Journal of Xenobiotics*, *11*(4), 130–141. https://doi.org/10.3390/jox11040009

Strayer, D. L., & Dudgeon, D. (2010). Freshwater biodiversity conservation: Recent progress and future challenges. *Journal of the North American Benthological Society*, *29*(1), 344–358. https://doi.org/10.1899/08-171.1

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Valadez-Cano, C., Hawkes, K., Calvaruso, R., Reyes-Prieto, A., & Lawrence, J. (2022). Amplicon-based and metagenomic approaches provide insights into toxigenic potential in understudied Atlantic Canadian lakes. *FACETS*, *7*, 194–214. https://doi.org/10.1139/facets-2021-0109

Wongkiew, S., Chaikaew, P., Takrattanasaran, N., & Khamkajorn, T. (2022). Evaluation of nutrient characteristics and bacterial community in agricultural soil groups for sustainable land management. *Scientific Reports*, *12*(1), 7368. https://doi.org/10.1038/s41598-022-09818-1

Woodcroft, B. J., Aroney, S. T. N., Zhao, R., Cunningham, M., Mitchell, J. A. M., Blackall, L., & Tyson, G. W. (2024). *SingleM and Sandpiper: Robust microbial taxonomic profiles from metagenomic data* (p. 2024.01.30.578060). bioRxiv. https://doi.org/10.1101/2024.01.30.578060

Wright, D. H. (1983). Species-Energy Theory: An Extension of Species-Area Theory. *Oikos*, *41*(3), 496–506. https://doi.org/10.2307/3544109

Xing, C., Chen, J., Zheng, X., Chen, L., Chen, M., Wang, L., & Li, X. (2020). Functional metagenomic exploration identifies novel prokaryotic copper resistance genes from the soil microbiome†. *Metallomics*, *12*(3), 387–395. https://doi.org/10.1039/c9mt00273a

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Zhang, L., Chen, F., Zeng, Z., Xu, M., Sun, F., Yang, L., Bi, X., Lin, Y., Gao, Y., Hao, H., Yi, W., Li, M., & Xie, Y. (2021). Advances in Metagenomics and Its Application in Environmental Microorganisms. *Frontiers in Microbiology*, *12*. https://www.frontiersin.org/articles/10.3389/fmicb.2021.766364

Zia, H., Harris, N. R., Merrett, G. V., Rivers, M., & Coles, N. (2013). The impact of agricultural activities on water quality: A case for collaborative catchment-scale management using integrated wireless sensor networks. *Computers and Electronics in Agriculture*, *96*, 126–138. https://doi.org/10.1016/j.compag.2013.05.001

Chapter 5: Conclusions and future direction

5.1 Bacterial communities, a bedrock of aquatic ecosystems

Bacterial communities are crucial to the functioning of aquatic ecosystems. They are found everywhere, capable of thriving in a wide range of environments, even the most extreme ones—such as beneath ice (Tran et al., 2018), in hydrothermal vents (Zeng et al., 2021), hot springs (Massello et al., 2020), and acid mines (Méndez-García et al., 2015). These communities exhibit vast metabolic and physiological versatility which makes them vital drivers of several biogeochemical processes on Earth (Falkowski et al., 2008). Remarkably, the total amount of prokaryotic carbon is estimated to be 60–100% of the total carbon in plants (Prosser et al., 2007). In aquatic ecosystems, including lakes, the average cellular carbon content is estimated to be between 5–20 fg of C per cell (Whitman et al., 1998). Understanding the ecology of aquatic bacteria represents one of the most significant intellectual challenges in contemporary ecology. Gaining this understanding is essential for addressing many of the critical challenges facing humanity today, such as the management of natural ecosystems and the mitigation of environmental degradation to maintain a sustainable environment.

5.2 Lake physicochemistry as a driver of bacterial diversity and community composition

This thesis, in its first research chapter, studied bacterial diversity and community composition across a wide environmental and spatial gradient spanning seven Canadian ecozones and 403 freshwater lakes. Along the wide trophic gradient investigated, we were able to detect the presence of not only cosmopolitan lake bacteria, but groups associated with high nutrient conditions. We identified a broad scale pattern in diversity, where lakes located in the more northwestern ecozones exhibited higher richness than those in the southeastern ecozones. These change in diversity were linked to lake productivity, ion composition, and lake depth. Cyanobacteria, Bacteroidetes, and Firmicutes were enriched in nutrient-rich lakes, while Verrucomicrobia were enriched in nutrient-poor lakes. We also found that bacterial community composition varied across the seven ecozones and that the variation in the observed community structure was most strongly related to lake physico-chemistry, particularly lake pH and trophic state. Moreover, nutrient rich lakes in the Prairies and Boreal Plains often exhibited the most taxonomically distinct communities.

Our findings in this chapter included linking distinct bacterial groups to lake physicochemical conditions and varying nutrient concentration level. This suggested that different taxonomic groups played varying ecological roles in response to environmental and spatial gradients.

5.3 Land use impacts bacterial community composition differently at varying spatial scales

These findings led us to expand our investigation to include five more ecozones and 218 additional freshwater lakes. We then focused on investigating the influence of water quality and land use variables on bacterial communities across 12 ecozones that represent wide regional variation of the Canadian landscape. Thus, in addition to investigating a pan Canadian scale response of bacterial communities to water quality and specific watershed land use types, we were also able to investigate regional responses specific to regions where certain land use types were prevalent. We took this approach because while in-lake conditions like pH, trophic status, productivity, ion concentration and depth are important in influencing bacterial diversity and community composition (like shown in the first research chapter of this thesis), such in-lakes conditions are typically a depiction of the lake's surrounding terrestrial watershed. This makes lake ecosystems excellent sentinels and integrators of environmental change at scales ranging from regional to continental (Williamson et al., 2008) which we went on to capture in this thesis.

On incorporating spatial scale (continental and regional) as well as land use variations in our investigations, we were able to detect that total phosphorous (TP) was the most significant water quality variable exerting a strong influence on bacterial community structure at the continental scale. In fact, we found a profound shift that correlated with the transition from eutrophic to hypereutrophic conditions. At the regional scale, and despite the presence of significant regional differences in environmental conditions, water quality strongly influenced bacterial community structure in all ecozones. Intriguingly, for land use, at the pan Canadian scale, we found that agriculture and, to a lesser extent, urbanization were significant land use types influencing community structure. However, at the regional scale, we encountered a clear dichotomy wherein in ecozones where agriculture was prevalent, agriculture was consistently significant in explaining community structure. Likewise, in extensively urbanized ecozones, urbanization was consistently significant in explaining community structure.

Our spatial scale variation was integral in revealing the influence of surrounding terrestrial land use type on bacterial community structure. We then used metagenomics to analyse functional diversity in the lakes. Bacterial communities are metabolically versatile, perpetuating a wide range of complex biogeochemical cycles. Could we pinpoint potential metabolic activities occurring in these lakes across the continental scale?

5.4 Lake bacterial communities display propensity for diverse metabolisms, including xenobiotics biodegradation across Canada

Findings from the previous chapters led us to explore both taxonomic and functional diversity in a subset of 357 lakes via metagenomic approaches. In the final research chapter of this thesis, we employed metagenomic sequencing techniques to unravel taxonomic composition and functional diversity in lakes while investigating the influence of environmental conditions and land use on both biological indices. Our metagenomic analyses along a wide trophic gradient from ultraoligotrophic to hypereutrophic lakes revealed a unimodal relationship between bacterial richness and lake trophic status wherein richness peaked at intermediate trophic states (mesoeutrophic to mesotrophic lakes), consistent with established scientific reports on microbial diversity-productivity relationships (Smith, 2007). Overall, we detected that the influence of lake physicochemical properties on both taxonomic and functional composition of lake bacteria surpasses other environmental conditions. However, land use, specifically agriculture, influenced both taxonomic composition and function similarly.

While a number of protein coding genes associated with different kinds of metabolic processes (including carbohydrate and energy metabolisms) were found in the lake metagenomes, metagenomes were particularly enriched with genes coding for xenobiotics biodegradation and metabolism. This metabolic pathway was particularly related to both environmental and land use drivers, wherein physicochemical parameters like lake pH, total nitrogen concentration strongly influenced the capability of lake bacteria to degrade xenobiotics in lakes as well as land use types like agriculture and pasture.

5.5 Future Directions

As environmental changes continue to accelerate, there is an urgent need for studies that take a comprehensive approach to understanding freshwater ecosystems. This is particularly important for ensuring freshwater sustainability in a lake-rich country like Canada. This thesis resolved bacterial diversity patterns, community composition, and function across both continental and regional scales. We were able to link changes in these biological components to variations in lake environmental conditions and the influences exerted by watershed land use in the surrounding terrestrial ecosystems. We identified the potential of lake bacteria to carry out a crucial metabolism that could facilitate bioremediation in polluted aquatic ecosystems, a mechanism tremendously advantageous in mitigating the negative effects of agricultural activities exacerbated by increasing human populations.

However, this thesis did not determine if these environmentally important genes are actively expressed, which is a limitation of metagenomic analysis. Future metatranscriptomic work, capable of revealing gene expression would be a valuable continuation of this study. By focusing on what genes are expressed by the entire microbial community, metatranscriptomics can shed light on the active functional profile of a microbial community (Aguiar-Pulido et al., 2016). Metatranscriptomics represents a deeper layer of analysis, complementary to metagenomics. It complements metagenomics by revealing which genes from the metagenome are actively transcribed and to what extent. This allows for the study of gene expression in complex microbiomes at specific times and under defined environmental conditions. In the Arctic Ocean, researchers employed both metagenomics and metatranscriptomics to evaluate the prevalence and diversity of metabolic pathways and bacterial taxa involved in the degradation of aromatic compounds. The study found that these pathways were not only widespread but also actively expressed (Grevesse et al., 2022). Conducting similar analyses would yield significant insights into the ecological roles of freshwater lake bacteria across Canada.

Bibliography

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., & Narasimhan, G. (2016). Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis: Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data. *Evolutionary Bioinformatics*, *12s1*, EBO.S36436. https://doi.org/10.4137/EBO.S36436

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science (New York, N.Y.)*, *320*(5879), 1034–1039. https://doi.org/10.1126/science.1153213

Grevesse, T., Guéguen, C., Onana, V. E., & Walsh, D. A. (2022). Degradation pathways for organic matter of terrestrial origin are widespread and expressed in Arctic Ocean microbiomes. *Microbiome*, *10*(1), 237. https://doi.org/10.1186/s40168-022-01417-6

Massello, F. L., Chan, C. S., Chan, K.-G., Goh, K. M., Donati, E., & Urbieta, M. S. (2020). Meta-Analysis of Microbial Communities in Hot Springs: Recurrent Taxa and Complex Shaping Factors beyond pH and Temperature. *Microorganisms*, *8*(6), 906. https://doi.org/10.3390/microorganisms8060906

Méndez-García, C., Peláez, A. I., Mesa, V., Sánchez, J., Golyshina, O. V., & Ferrer, M. (2015). Microbial diversity and metabolic networks in acid mine drainage habitats. *Frontiers in Microbiology*, *6*, 475. https://doi.org/10.3389/fmicb.2015.00475

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., Green, J. L., Green, L. E., Killham, K., Lennon, J. J., Osborn, A. M., Solan, M., van der Gast, C. J., & Young, J. P. W. (2007). The role of ecological theory in microbial ecology. *Nature Reviews Microbiology*, *5*(5), 384–392. https://doi.org/10.1038/nrmicro1643

Smith, V. H. (2007). Microbial diversity–productivity relationships in aquatic ecosystems. *FEMS Microbiology Ecology*, *62*(2), 181–186. https://doi.org/10.1111/j.1574-6941.2007.00381.x

Tran, P., Ramachandran, A., Khawasik, O., Beisner, B. E., Rautio, M., Huot, Y., & Walsh, D. A. (2018). Microbial life under ice: Metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes. *Environmental Microbiology*, *20*(7), 2568–2584. https://doi.org/10.1111/1462-2920.14283

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, *95*(12), 6578–6583. https://doi.org/10.1073/pnas.95.12.6578

Williamson, C. E., Dodds, W., Kratz, T. K., & Palmer, M. A. (2008). Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Frontiers in Ecology and the Environment*, *6*(5), 247–254. https://doi.org/10.1890/070140

Zeng, X., Alain, K., & Shao, Z. (2021). Microorganisms from deep-sea hydrothermal vents. *Marine Life Science & Technology*, *3*(2), 204–230. https://doi.org/10.1007/s42995-020-00086-4