

Spatial and Temporal Perspectives on the Canadian Lake Microbiome

Rebecca Esther Garner

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

October 2023

© Rebecca E. Garner, 2023

**CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES**

This is to certify that the thesis prepared

By: Rebecca Garner

Entitled: Spatial and Temporal Perspectives on the Canadian Lake Microbiome

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. Brandon Findlay

External Examiner

Dr. Karthik Anantharaman

Examiner

Dr. Rassim Khelifa

Examiner

Dr. Dylan Fraser

Examiner

Dr. Pedro Peres-Neto

Thesis Supervisor (s)

Dr. David Walsh

Dr. Irene Gregory-Eaves

Approved by

Dr. Robert Weladji

Chair of Department or Graduate Program Director

November 15, 2023

Date of Defence

Dr. Pascale Sicotte

Dean, Faculty of Arts and Science

Abstract

Spatial and Temporal Perspectives on the Canadian Lake Microbiome

Rebecca E. Garner, Ph.D.

Concordia University, 2023

Lakes are critical freshwater resources and occupy an important place in Canadian culture and recreation. Underpinning lake ecosystems are a broad diversity of microorganisms, which play central roles in regulating water quality, elemental cycles, and food webs. Lakes are increasingly menaced by a range of human pressures which are changing the composition of microbial communities in ways that are challenging to predict, largely because of the immense environmental heterogeneity among lakes at multiple spatial and temporal scales. This dissertation unites spatial and temporal perspectives on the lake microbiome to elucidate its structure, function, and environmental drivers, including human impact. To begin, two studies investigated the contemporary distributions of bacteria and protists using genome-resolved metagenomics and amplicon analyses in over 300 lakes across Canada based on the standardized sampling conducted by the LakePulse Network. For bacteria and protists, lake trophic state and the terrestrial influences of soils and land use – namely, agriculture and human population density – were drivers of taxonomic diversity and various components of microbiome function. The next two studies leveraged sediment DNA archives and paleogenetic techniques to reconstruct the historical diversity and dynamics contextualizing the contemporary lake microbiome. First, metagenome fragment recruitment was used to probe the taxonomic breadth of microorganisms preserved in preindustrial-age sediment records. Small fractions of sediment metagenomes were associated with the preferential preservation of previously undiscovered viruses and bacterioplankton, including lineages with well-characterized ecotypic diversity of paleoindicator potential. Next, multidecadal monitoring and sediment DNA records were compared from the IISD Experimental Lakes Area to evaluate how each of these time series detects eutrophication and climate change signals in lakes with different trophic state histories. Strong congruence was found between algal community turnover in monitoring and sediment records, validating the paleogenetic approach to reconstructing past ecosystem dynamics. While both time series tracked nutrient drivers in fertilized lakes, the 120-year perspective opened by paleogenetics signaled the effects of warming a decade earlier in the continuously fertilized lake than the oligotrophic systems, pointing to synergisms between eutrophication and climate change. Overall, this dissertation shows that the lake microbiome is responding to rising human impact over the past century at a continental scale, notably to effects across the land-water interface and increasingly to a changing climate.

Acknowledgements

I could write another thesis to express my thanks to David Walsh for his mentorship. Thank you for accepting a dazed undergraduate student into your lab and setting me on this course. Throughout my time as your student, you went above and beyond to provide me with every opportunity to grow as a microbiologist. Thank you for sending me to the Microbial Diversity course in Woods Hole to encounter our community and play with magnetotactic bacteria. It has been an amazing journey: from papering the office windows with printouts of Actinobacteria phylogenies, to riding the “train to Crazy Town” on the sediment paleogenomics study, to feeling inspired and a bit intimidated at ISME18, to getting the cover – thank you for everything.

Rene Gregory-Eaves treated my co-supervision with as much care and attention as many students receive from their primary supervisors. Thank you for making me feel just as at home at McGill as at Concordia, and for nurturing camaraderie within your lab. My thesis exists because you are a visionary of sediment DNA analysis.

Thank you to my committee members, Dylan Fraser and Pedro Peres-Neto, for their valuable input and challenging me to consider my research from different perspectives.

Stephanie Messina-Pacheco, David Colatiriano, Arthi Ramachandran, and Patricia Tran mentored me as an undergraduate student and were unfailing supportive and generous in sharing their knowledge. Susanne Kraemer blazed trails in bioinformatics and spent a good part of her postdoc answering my questions. Zofia Taranu did the impossible: make statistics accessible. Many thanks to Marie-Ève Monchamp for being a wonderful colleague and including me in her freshwater conservation workshop. Joanna Gauthier spent several hours in a small room on the sixth floor of Stewart Biology teaching me how to extract DNA from sediments. It was always great to catch up with Alexandre Baud, Katherine Griffiths, and Cindy Paquette on our paleo work. Michelle Gros got the soccer team together and helped out on projects behind the scenes. Makayla Harrison was an awesome office buddy with an endless Crossfit knowledgebase. Susie McLatchie could always be relied on to guess the ingredients in Free Lunch, set record walking times, bake top-notch sourdough, and knit through every season of The Sopranos. Paul MacKeigan was there (almost) through it all: thank you for working together on a bunch of projects, some of which failed (DNA!), knowing that the best lake is Gun Lake, sitting through an hour of Shostakovich, and always being ready with a Simpsons reference.

I am very grateful to Wentworth Brookes for making the computational research possible by solving countless computer issues. In office GE-300.05, Jay Pimprikar kept the discourse formal and the movements robotic.

Yannick Huot created incredible opportunities to work on impactful, multidisciplinary research through his realization of the NSERC Canadian LakePulse Network. I am so lucky to have had exceptional

experiences traversing Canada to sample our lakes. Marie-Pierre Varin organized the field campaigns which sent me to Nova Scotia, Newfoundland and Labrador, and the British Columbia Interior. François-Xavier Teyssere was a reliable sampling partner and kept clear of the ticks for most of those 92 lakes. Jelena Juric and Maxime Fradette filled every one of my many data requests.

It has been a point of pride to conduct research on the iconic whole-ecosystem experiments in the IISD Experimental Lakes Area. Michael Paterson and Scott Higgins shared expert knowledge of the lake monitoring and paleolimnology. Sonya Havens dug through the freezers to send us part of the precious filter archive.

The *Groupe de recherche interuniversitaire en limnologie* offered extensive training opportunities, but more than that, provided a supportive, extended community.

My graduate studies were tremendously enriched by experiences that took me outside of Montreal. Timothée Poisot and Andrew MacDonald ran the summer school in Data-Driven Ecological Synthesis which pushed my skills in computational biology. Alastair Simpson and Andrew Roger brought me into the fascinating protist microcosmos during my stay at Dalhousie. Thank you to Gordon Lax and Yana Eglit for being wonderful hosts and welcoming me onto their winning pub trivia team. The Microbial Diversity course at the Marine Biological Laboratory was an unforgettable summer because of Scott Dawson, Rachel Whitaker, the incredible faculty and teaching assistants, and my 19 classmates. I will remember it always as a time when everything seemed possible. Special thanks to Team MicroMagneto for their endless enthusiasm and teaching me the ways of the microscope, and to the class for letting me play and replay the magnetotaxis videos. Stefan Wuerz and Hans-Curt Flemming ran the Singapore Centre for Environmental Life Sciences Engineering summer school which gave me a new level of appreciation for biofilms (“It’s a mess, and the mess plays with each other, and the biofilm wins” –H.C.F.).

I am grateful for the scholarship support from the FRQNT, NSERC, GRIL, BIOS², and the Stephan Bronfman Scholarship in Graduate Environmental Studies. Thank you to the Hive Free Lunch, which is hands-down the best program at Concordia.

Outside of university, I must thank Marilyn Sirzyk Gur-Esh for being an extremely forgiving piano teacher and patiently guiding me through Rachmaninoff. Hugs to Megan Benjamin, Chaim Ward, Adam Schwartz, and Leandro Rizk for their calls and staying amazing friends. A big hug also to my Aunt Marion Blumenthal Lazan for her beautiful encouragement. Joshua, Sandra, Ari, and Enzo have cheered me on through this journey. Thank you, Dad, for teaching me everything I know – at least in hockey.

Mom, thank you for always being there for me and for taking care of Floyd those months that I was away on fieldwork.

Thank you, Floyd, for being my best friend; I miss you.

Dedication

To my grandparents

Naomi Elson
née Lazan

Harold Roy
Elson

Cwetla Gasiorowski
née Grinbaum

Jozef Riwen
Gasiorowski

Contribution of Authors

In **Chapter 2**, David A. Walsh conceived the study and designed the analyses with Rebecca E. Garner. Yannick Huot conceived and directed the LakePulse Network. Marie-Pierre Varin coordinated the LakePulse field campaigns and data quality control. R.E.G., Susanne A. Kraemer, and Vera E. Onana performed the bioinformatic analyses. Maxime Fradette performed the geomatic analyses. R.E.G. and D.A.W. wrote the manuscript with improvements by all the authors.

In **Chapter 3**, Rebecca E. Garner, Irene Gregory-Eaves, and David A. Walsh designed the analyses. Yannick Huot directed the LakePulse Network which conducted the national lake assessment. Susanne A. Kraemer and Vera E. Onana extracted the DNA from filters. R.E.G. performed the PCR and computational analyses. R.E.G., D.A.W., and I.G.-E. wrote the manuscript with improvements by all the authors.

In **Chapter 4**, David A. Walsh conceived the study based on a question by Irene Gregory-Eaves. D.A.W. and Rebecca E. Garner designed the analyses. I.G.-E. established the sediment DNA extraction facility. R.E.G. extracted the DNA from sediments and performed the metagenomic analyses. R.E.G., D.A.W., and I.G.-E. wrote the manuscript.

In **Chapter 5**, Rebecca E. Garner, Zofia E. Taranu, Irene Gregory-Eaves, and David A. Walsh designed the analyses. R.E.G. extracted the DNA from sediments and performed the PCR and computational analyses with major inputs on the statistical analyses from Z.E.T. Michael J. Paterson contributed extensive knowledge on lake monitoring and paleolimnology in the International Institute for Sustainable Development Experimental Lakes Area. R.E.G., Z.E.T., I.G.-E., and D.A.W. wrote the manuscript with improvements by all the authors.

For each chapter, all authors read and approved the final manuscript.

Table of Contents

List of Figures	xii
List of Tables.....	xiii
List of Abbreviations	xiv
Chapter 1: Introduction	1
1.1 Freshwater lakes are diverse ecosystems under human impact	1
1.2 The lake microbiome	3
1.3 Thesis objectives and approach	5
Chapter 2: A Genome Catalogue of Lake Bacterial Diversity and Its Drivers at Continental Scale	7
2.1 Abstract.....	7
2.2 Introduction	7
2.3 Results	9
2.3.1 The LakePulse MAG catalogue.....	9
2.3.2 Phylogenetic and metabolic diversity of lake bacteria.....	10
2.3.3 Lake conditions shape assemblage structure and function.....	11
2.3.4 Watershed effects and human impact on the lake microbiome.....	14
2.4 Discussion.....	15
2.5 Methods	16
2.5.1 Lake sampling	16
2.5.2 Environmental data.....	16
2.5.3 DNA extraction and metagenome sequencing.....	17
2.5.4 Metagenome assembly and binning.....	17
2.5.5 Functional analysis	18
2.5.6 MAG read recruitment and coverage normalization	19
2.5.7 Taxonomic classification and phylogenetic analysis	20
2.5.8 Metagenome analysis	20
2.5.9 16S rRNA gene amplicon sequencing and analysis	20
2.5.10 Ecological analysis	21
2.5.11 Reporting summary	21
2.5.12 Data availability	21

2.5.13 Code availability	22
2.6 Acknowledgements	22
2.7 Author contributions	22
2.8 Competing interests	22
2.9 Additional information.....	22
Chapter 3: Protist Diversity and Metabolic Strategy in Freshwater Lakes Are Shaped by Trophic State and Watershed Land Use on a Continental Scale	28
3.1 Abstract.....	28
3.2 Importance	29
3.3 Introduction	29
3.4 Results	30
3.4.1 Sampled lakes and watersheds display high environmental heterogeneity	30
3.4.2 Lakes support taxonomically and functionally diverse protist assemblages	31
3.4.3 Protist assemblages vary regionally and across lake trophic states.....	32
3.4.4 Physicochemical and watershed conditions predict community turnover	33
3.4.5 Trophic strategies exhibit contrasting distributions.....	33
3.5 Discussion.....	34
3.5.1 Protist diversity unveiled at a continental scale	34
3.5.2 Surface water assemblages contain high proportions of phototrophs and heterotrophs	35
3.5.3 Protist communities respond to local environmental conditions	35
3.5.4 Hypereutrophic lakes are taxonomically distinct.....	37
3.5.5 Watersheds influence lacustrine protist diversity	38
3.5.6 Conclusion.....	39
3.6 Materials and methods.....	39
3.6.1 Lake selection and sampling	39
3.6.2 18S rRNA gene amplification and sequencing	40
3.6.3 ASV inference and annotation	40
3.6.4 Sequence similarities with known diversity	40
3.6.5 Environmental data collection.....	41
3.6.6 Diversity analyses	41
3.6.7 Data availability	42
3.7 Acknowledgements	42

3.8 Supplemental material	49
Chapter 4: Sediment Metagenomes as Time Capsules of Lake Microbiomes	51
4.1 Abstract.....	51
4.2 Importance	51
4.3 Introduction	52
4.4 Results	54
4.4.1 Environmental context	54
4.4.2 Preserved metagenome capture.....	54
4.4.3 Taxonomic diversity in free and captured metagenomes	54
4.4.4 Highlights of preservation signals in captured metagenomes.....	55
4.4.5 Exploring unconstrained preservation signals in free metagenomes.....	56
4.5 Discussion.....	57
4.5.1 Preserved taxa in sediment metagenomes	57
4.5.2 Harnessing microbial ecotypes as paleoindicators	59
4.5.3 Preserved metagenome capture.....	59
4.5.4 Conclusion and outlook	60
4.6 Materials and methods.....	61
4.6.1 Lake water and sediment sampling	61
4.6.2 DNA extraction	61
4.6.3 Metagenome shotgun sequencing, assembly, and annotation	62
4.6.4 Read recruitment	62
4.6.5 Statistical analyses.....	63
4.6.6 Phylogenetic analyses.....	63
4.6.7 Data availability	63
4.7 Acknowledgements	63
4.8 Supplemental material	68
Chapter 5: Congruent paleogenetic and monitoring time series document the effects of eutrophication and climate on algal communities in experimental lakes	69
5.1 Abstract.....	69
5.2 Introduction	69
5.3 Methods	71
5.3.1 Lake experimental histories	71

5.3.2 Monitoring data.....	72
5.3.3 Lake sampling	72
5.3.4 Sediment dating and subsectioning	73
5.3.5 IISD-ELA filter archives	73
5.3.6 DNA extraction, amplification, and sequencing.....	73
5.3.7 Bioinformatic analysis of 18S rRNA amplicons	74
5.3.8 Statistical analyses.....	74
5.3.9 Code availability	75
5.3.10 Data availability	75
5.4 Results	75
5.4.1 Phytoplankton monitoring	75
5.4.2 Microeukaryotic diversity in lake sediments	75
5.4.3 Water column diversity not captured in sediment DNA records.....	76
5.4.4 Taxonomic turnover in monitoring and sediment records.....	76
5.4.5 Correlations between monitoring and sediment records	77
5.4.6 Effects of nutrients and climate on algal diversity	77
5.5 Discussion.....	78
5.5.1 Strong congruence between monitoring and paleogenetic records	79
5.5.2 Effects and timing of nutrient and climate drivers in monitoring and paleogenetic time series.....	79
5.5.3 Microeukaryotic diversity in the paleogenetic record	80
5.5.4 Conclusion.....	81
5.6 Acknowledgements	81
Chapter 6: Discussion	90
6.1 Future directions.....	92
References	94

List of Figures

Figure 2.1 The LakePulse MAG catalogue: geographical and environmental scope, quality and patterns of local-scale diversity.....	23
Figure 2.2 Taxonomic, phylogenetic and functional diversity of the LakePulse MAG catalogue.	24
Figure 2.3 Niche breadth and environmental filtering of MAGs based on taxonomic and functional biogeography.	25
Figure 2.4 Diversity of specific bacterial functions among lake assemblages and individual MAGs.	26
Figure 2.5 Influences of watershed soil and land use on lake bacterial assemblages.	27
Figure 3.1 Diversity and distributions of protists across 366 Canadian lakes.	44
Figure 3.2 Accumulation and incidence of genotypes across lakes.	45
Figure 3.3 Local protist diversity of each lake.	46
Figure 3.4 Taxonomic and phylogenetic variation of protist assemblages among lakes.	47
Figure 3.5 Local contribution to β -diversity of protist assemblages across the Canadian landscape.....	47
Figure 3.6 PCA of the taxonomic variation among (A) phototroph, (B) heterotroph, and (C) mixotroph assemblages at the level of individual ASVs.	48
Figure 4.1 Map of Canada with the starred locations of study lakes: Lac Paula, Eightmile Lake, and Grand lac Touradi.	64
Figure 4.2 (a) As illustrated in this experimental design schematic, the recruitment of unassembled metagenome reads to metagenome assemblies tracks the sequence similarities between focal and reference metagenomes. (b) Metagenome capture was performed with a 90% sequence identity threshold derived from the delineation of closely and distantly related sequences.....	65
Figure 4.3 Relative coverage of prevalent phyla in the free and captured metagenomes of three lakes.....	65
Figure 4.4 (a) Relative coverage of prevalent taxonomic orders in the free and captured metagenomes of three lakes. (b) PCA of order-rank taxonomic composition data of free and captured metagenomes.	66
Figure 4.5 Gene counts of selected preserved <i>Cyanobacteria</i> , bacterioplankton, microeukaryotes, and viruses in free surface water and sediment metagenomes in three lakes.	67
Figure 5.1 An overview of the study systems shown through (A) maps of the region and IISD-ELA lakes and (B) lake experimental histories.....	82
Figure 5.2 Microeukaryotic diversity in experimental and unmanipulated lakes recorded in (A) microscopy-based biomass monitoring and (B) ASV assemblages along 0.5-cm sediment intervals.	83
Figure 5.3 PCAs showing the taxonomic variation among algal assemblages in (A) monitoring and (B) sediment layers across all lakes.	84
Figure 5.4 RV coefficients correlating the temporal dynamics of algal assemblages in monitoring and sediment records.....	84
Figure 5.5 Deviance explained by GAMs fitting algal taxonomic variation in monitoring or sediment time series using nutrient (TP and TN) and/or climate (air temperature) predictors.....	85

List of Tables

Table 3.1 Percent deviance explained by generalized dissimilarity models (GDMs) fitting the responses of protist assemblages to environmental gradients.....	49
Table 4.1 Spatial coordinates, morphological characteristics, and surface water physicochemical profiles of study lakes.....	67
Table 4.2 Counts and percentages (normalized by the number of unassembled reads) of reads mapped with a 90% sequence identity threshold in metagenome capture experiments.....	68

List of Abbreviations

ANI	average nucleotide identity
ANOVA	analysis of variance
ASV	amplicon sequence variant
BLAST	Basic Local Alignment Search Tool
BWA	Burrows-Wheeler Aligner
C	carbon
CAZyme	carbohydrate-active enzyme
CBB	Calvin–Benson–Bassham [cycle]
Chl <i>a</i>	chlorophyll <i>a</i>
DIC	dissolved inorganic carbon
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
DOC	dissolved organic carbon
GAM	generalized additive model
GC	guanine-cytosine
GDM	generalized dissimilarity model
GEM	Genomes from Earth's Microbiomes [catalogue]
GH	glycoside hydrolase
GOLD	Genomes OnLine Database
IISD-ELA	International Institute for Sustainable Development Experimental Lakes Area
JGI	Joint Genome Institute
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG orthologue
LCBD	local contribution to beta diversity
MAG	metagenome-assembled genome
MIMAG	Minimum Information about a Metagenome-Assembled Genome
N	nitrogen
NSERC	Natural Sciences and Engineering Research Council [of Canada]
P	phosphorus
PCA	principal component analysis
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
PR ²	Protist Ribosomal Reference database
PUL	polysaccharide utilization locus
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
SSU	small subunit
TAD ₈₀	[central] 80% truncated average depth of coverage
TN	total nitrogen

TP	total phosphorus
tRNA	transfer ribonucleic acid
UV	ultraviolet

Chapter 1

Introduction

1.1 Freshwater lakes are diverse ecosystems under human impact

Fresh water is rare and precious. Although the Earth is unmistakably a blue planet whose surface is 71% covered by ocean, fresh water comprises only ~2.5% of the hydrosphere (Babkin, Klige & Vuglinsky, 2003). While most fresh water is locked up in ice (69.5% in glaciers, ice caps, ground ice, and permafrost) or underground (30.1% in groundwater), a tiny fraction of total fresh water is accessible from lakes (0.25%) and other surface water bodies. Satellite remote sensing and geostatistical modeling place the number of lakes worldwide at 1.43 million for lakes with surface areas $>0.1 \text{ km}^2$ (Messenger *et al.*, 2016) and 117 million including those $>0.002 \text{ km}^2$ (Verpoorter *et al.*, 2014). Together with rivers, lakes represent the most important sources of everyday fresh water use. However, lake water renewal is often slow. With its limited volume, unequal worldwide distribution, and uncertain future supply, fresh water may be considered our most precious resource.

Lakes are formed through the accumulation of water at low positions in the landscape where water drains more slowly than it renews. At these low positions, lakes receive and integrate materials from the catchment, airshed, and groundwater. Linked with the terrestrial environment and hydrologic flow system, lakes are connected by the movement of materials and species within a broader waterscape (Heino *et al.*, 2021). By integrating fluxes of materials and conditions in their surroundings, lakes are sensitive to environmental change and often sentinels at its forefront (Williamson *et al.*, 2008; Adrian *et al.*, 2009).

Across the landscape, lakes and their watersheds reflect a mosaic of climate, landform, geology, and land cover/land use. The physical and chemical profiles of lakes are influenced by catchment characteristics and landscape position and connectivity within the hydrologic flow system (Kratz *et al.*, 1997; Quinlan *et al.*, 2003). Land use and its associated runoff of ions, nutrients, and other solutes further contribute to the variability in watershed and lake characteristics. The influence of landscape gradients explains why neighbouring lakes situated in identical climates and geological settings can exhibit highly contrasting morphometric and physicochemical profiles. Some lake characteristics are temporally dynamic through seasonality (e.g., influenced by ice cover, mixing regimes, and other weather-induced patterns) and the history of human impact. Together, these sources of variability contribute to the immense environmental heterogeneity of lakes at multiple spatial and temporal scales (Kratz, MacIntyre & Webster, 2005).

Lakes provide myriad ecosystem services that support human health and wellbeing (Rinke *et al.*, 2019). Freshwater lakes provision water for domestic drinking supply, agricultural irrigation, and industrial processing, among other uses. Large cultural, recreational, and aesthetic importance are attached to lakes.

In some lakes, fish and invertebrate species are harvested for subsistence or commercial trade. Lakes also render ecosystem functions beyond tangible human benefit or economic value. Compared with terrestrial and marine ecosystems which occupy far greater spatial extents, lakes are disproportionately biodiverse and harbour many endemic species (Strayer & Dudgeon, 2010). In addition to aquatic species, lakes support wetland and terrestrial biodiversity and are important stopover habitats for bird populations along their migration paths. Furthermore, lakes contribute to biogeochemical cycles on a globally significant scale. Fluxes of carbon dioxide and methane in lakes are important factors in global carbon budgets, as they potentially offset terrestrial carbon storage (Cole *et al.*, 2007; Tranvik *et al.*, 2009; Bastviken *et al.*, 2011). Lakes also function as globally important sinks for carbon (Anderson, Heathcote & Engstrom, 2020) and nutrients (Wu *et al.*, 2022).

With increasing land use, water demands, and climate change, lakes are under pressure from multiple anthropogenic stressors and experiencing rapid change (Brönmark & Hansson, 2002). Eutrophication is the foremost freshwater stressor and is associated with nutrient inputs often caused by excessive fertilizer applications in agriculture (Fraterrigo & Downing, 2008). Altered precipitations associated with climate change are leading to increased terrestrial runoff which exacerbates eutrophication (Jeppesen *et al.*, 2009). Other growing pressures on lakes include salinization (Hintz, Fay & Relyea, 2022), calcium decline precipitated by acid rain recovery (Weyhenmeyer *et al.*, 2019), pollution by persistent and emerging contaminants such as plastics, the degradation of littoral habitats by shoreline erosion, and the effects of cumulative stressors (Reid *et al.*, 2019). Lake warming (O'Reilly *et al.*, 2015) is modifying mixing regimes and increasing water evaporation rates as ice coverage periods shorten (Butcher *et al.*, 2015; Woolway, Sharma & Smol, 2022). Warming temperatures and stronger stratification are associated with lake deoxygenation (Jane *et al.*, 2021) and the restructuring of species habitats (Kraemer *et al.*, 2021). Because of humanity's growing freshwater footprint, the sharpest biodiversity declines are being observed in lakes and other freshwater ecosystems (Albert *et al.*, 2021).

Through the deposition of sediments burying materials from the overlying water column, lakes compile natural archives chronicling the histories of the ecosystem and catchment. Lake sediments deposited over the last century record the history of human impact and may be compared with deeper sediments which provide preimpact baselines (Dubois *et al.*, 2018). Lake sediments can be such reliable integrators of human-induced change that the sediment record of Crawford Lake, a small lake in southern Ontario, has been proposed as the global marker for the start of the current Anthropocene epoch (Gushulak *et al.*, 2021). More broadly, techniques in paleolimnology – the collection, analysis, and interpretation of sediment records – are used to reconstruct elements of lake and catchment histories over timeframes beyond the span of monitoring programs, often preceding industrialization or European settlement in the watershed (Smol, 2019).

1.2 The lake microbiome

In lakes and across the world's biomes, microbial diversity is extraordinarily vast. Large populations and short generation times propel the rapid evolution of many microbial lineages. The expansive metabolic diversity of bacteria and archaea is complemented by the morphological diversity enabling complex behaviours in microbial eukaryotes (Keeling & Campo, 2017). The functional diversity of microorganisms is expanded through symbioses, which can recruit metabolisms within and across the domains of life (Gast, Sanders & Caron, 2009; Husnik *et al.*, 2021). Microbial evolution is further directed by viruses as agents of genetic transfer (Bellas *et al.*, 2023). Metabolic versatility in microorganisms is illustrated by mixotrophs, which pair autotrophy and heterotrophy to varying degrees, often depending on environmental conditions (Jones, 1997). Novel microbial lineages continue to be discovered through both classical cultivation (Lax *et al.*, 2018) and culture-independent techniques, which have uncovered enormous microbial dark matter (Hug *et al.*, 2016).

Microorganisms underpin all lake processes and ecosystem services. Lake food webs are controlled through bottom-up forces, including microbial primary production, and top-down forces, including viral controls on microbial populations (Zimmerman *et al.*, 2019) and parasitism. Through their transformations of chemical compounds, microbes regulate large aspects of water quality and nutrient cycling. Microorganisms may also have harmful impacts on lake services. Contamination of lake water by microbial pathogens (e.g., fecal coliforms) are a concern in lakes exploited for drinking water or recreation. A more pervasive issue are algal blooms fueled by excess nutrients which degrade the quality of drinking water, force lake closures, and kill fish and benthic invertebrates (Huisman *et al.*, 2018). Microbial degradation of the dead bloom depletes oxygen from the water column and sediments, restructuring habitats and promoting the use of alternative electron acceptors (Arora-Williams *et al.*, 2018).

Through their spatial and temporal heterogeneity, lakes represent compelling systems to investigate the diversity of microbial form and function and the patterns and processes of community assembly. In a sampling of hundreds of European lakes, physicochemical conditions were the most important drivers of bacterial assemblages, whereas biotic factors ranked first for protists (Bock *et al.*, 2020). Within lakes, gradients and patchworks of environmental conditions create diverse ecological niches, further constrained by species interactions (Pernthaler, 2017). A clear example is the difference in conditions along stratified water columns, wherein, for example, phototrophs are found in the sunlit surface waters, while other metabolic modes dominate in deeper strata (Buck *et al.*, 2021). Lake sediments are microbial habitats defined by zonation, where the top sediment layer represents a zone of intense competition and high habitat complexity, overlying sharp redox gradients formed as resources become scarce with depth (Wurzbacher *et al.*, 2017). Among lakes, dispersal is a key community assembly process in bacteria and algae, more so than for larger organisms like zooplankton (Beisner *et al.*, 2006; Soininen *et al.*, 2011).

Human activities are altering microbial habitats and function in lakes, especially as excess nutrients and new forms of pollution are introduced. Plastics present novel habitats for microbial particle attachment, while their organic leachates, which contain toxic substances, appear to stimulate bacterial growth

(Sheridan *et al.*, 2022). Antibiotic use in humans and livestock is contributing to a reservoir of antimicrobial resistance in freshwater lakes (Spänig *et al.*, 2021; Kraemer *et al.*, 2022). Metagenomic analyses of three New Zealand lakes showed that differences in trophic state and agricultural land use were associated with variation in xenobiotic metabolism potential in surface sediment communities (Biessy *et al.*, 2022). As anthropogenic stressors decrease lake habitat complexity within and among lakes (e.g., through shoreline development and eutrophication), they are predicted to have homogenizing effects on the lake microbiome through space and time. Paleogenetic records in European lakes displayed trends toward the homogenization of Cyanobacteria and microeukaryote assemblages in recent decades (Monchamp *et al.*, 2018; Keck *et al.*, 2020; Huo *et al.*, 2022). As microorganisms are responding to anthropogenic pressures, there is opportunity to leverage microbes as indicators in freshwater monitoring (Sagova-Mareckova *et al.*, 2021).

Often, the lake microbiome is studied in a single lake or at regional scale. While such studies have contributed important insights into the phylogenetic and functional diversity of microorganisms in freshwater lakes (Newton *et al.*, 2011; Bertilsson & Mehrshad, 2022), a lack of standardized sampling and analysis may limit the ability to compare distributions among lakes. The Genomes from Earth's Microbiomes (GEM) catalog was an initiative aimed at reconstructing bacterial and archaeal genomes from metagenomes representing diverse habitats (Nayfach *et al.*, 2021). While impressive in its integration of biomes, the GEM catalog included metagenomes from the water columns of ~30 freshwater lakes, making a strong start but likely underrepresenting the microbial diversity inhabiting the millions of environmentally heterogeneous lakes worldwide. Meanwhile, the continental-scale sampling of lake protist assemblages across >200 European lakes primarily investigated the role of geographic barriers in community assembly (Boenigk *et al.*, 2018).

Without long-term observations, we are living in an invisible present and susceptible to underestimating change (Magnuson, 1990). Insights into the historical diversity and dynamics of the lake microbiome provide context for understanding current patterns and processes. As monitoring programs are rare and often relatively short-lived, paleolimnology provides a retrospective means of accessing lake history via the sediment record (Smol, 2019). Classical techniques in paleolimnology rely on the preservation of subfossils, algal pigments, and other biomolecules to reconstruct past microbial assemblages, but these are limited in their taxonomic breadth or resolution. Emerging paleogenetic techniques leverage microbial DNA archived in lake sediments (Capo *et al.*, 2022). The clear advantage of DNA records is that they capture soft-bodied organisms that do not leave subfossil remains (Domaizon *et al.*, 2017; Capo *et al.*, 2021). One of the key challenges to expanding paleogenetic research on microbial groups is that microbial life thrives in the sediment subsurface and sediment DNA pools represent a complex mixture from live and relic sources (Lennon *et al.*, 2018); signals from extant and historical microbiota are therefore difficult to disentangle. Circumventing this issue, the DNA records of groups like Cyanobacteria and eukaryotic algae whose metabolisms rely on access to sunlight – a resource clearly absent in the sediment subsurface – offer more defined paleogenetic reconstruction. While phytoplankton are

unequivocally of high ecological relevance in lakes, the use of PCR-based methods has narrowly constrained the microbial diversity available to paleogenetic study. Furthermore, while it has been demonstrated that sediment DNA records capture most of the microeukaryotic groups present in the contemporary water column (Capo *et al.*, 2015), it is less clear whether sediments accurately record the temporal dynamics of these communities (Thorpe *et al.*, 2023). Nonetheless, paleogenetic records of algal communities appeared to respond to nutrient and climate variation over the last century (Capo *et al.*, 2017; Monchamp, Spaak & Pomati, 2019; Keck *et al.*, 2020).

1.3 Thesis objectives and approach

The main objectives of this dissertation are to (1) elucidate the structure and function of the lake microbiome; (2) investigate its environmental drivers; and (3) assess the effects and timing of human impact. To address these objectives, this dissertation unites spatial and temporal perspectives on the lake microbiome using a suite of culture-independent molecular genetic and computational techniques.

The study systems used in this dissertation consist of 380 Canadian lakes. Canada is a prime location to study lake ecosystems as it contains the largest resource of lakes worldwide, encompassing nearly 1 million lakes with surface areas $>0.1 \text{ km}^2$ and representing 37% of all lakes globally (Minns *et al.*, 2008). Lakes studied in this dissertation were sampled under the first standardized national lake assessment conducted by the Natural Sciences and Engineering Research Council (NSERC) Canadian Lake Pulse Network (Huot *et al.*, 2019). Lakes were sampled in 12 ecozones and selected across different surface area categories and levels of human impact, as indexed by the types and proportions of land use within watersheds. Through the stratified, random selection scheme, sampled lakes represented the natural and human-mediated variation in lake and watershed characteristics. Lakes within the International Institute for Sustainable Development Experimental Lakes Area (IISD-ELA) in northwestern Ontario were also sampled by the LakePulse Network but complemented with monitoring time series initiated in 1969. Four IISD-ELA lakes were targeted for paleogenetic analysis in this dissertation: the northeast and southwest basins of Lake (L)226, which during the 1970s were respectively fertilized with carbon (C) and nitrogen (N) and C, N, and phosphorus (P); L227, which has been continuously fertilized since 1969 under decreasing N:P regimes; and unmanipulated L224 and L373.

Chapters 2 and 3 investigate the contemporary microbiome and its environmental drivers in the surface waters of >300 Canadian lakes at a continental scale. Chapter 2 uses genome-resolved metagenomics, a technique to construct composite genomes of microbial populations, to analyze the taxonomic and functional diversity of lake bacteria. Chapter 3 examines the biogeography of protists across lakes using 18S rRNA amplicon analysis. In both studies, a broad suite of environmental factors are considered in lakes and their watersheds, including different land use types. After investigating the contemporary patterns, Chapters 4 and 5 use paleogenetic techniques to provide historical contexts on the diversity and dynamics of the lake microbiome. Chapter 4 aims to discover the taxonomic breadth of microorganisms preserved in sediment DNA archives through a novel application of metagenome fragment

recruitment designed to disentangle historical signals from the indigenous sediment microbiota. Chapter 5 leverages the multidecadal monitoring and sediment records from the IISD-ELA to assess whether paleogenetics reliably records algal community temporal dynamics and compare how each time series can be used to detect the multiple stressors of eutrophication and climate change in lakes with highly diverging trophic state histories.

Overall, this dissertation investigates the lake microbiome at continental and multidecadal scales, providing insights into the patterns of taxonomic and functional diversity in ecosystems under increasing human impact.

Chapter 2

A Genome Catalogue of Lake Bacterial Diversity and Its Drivers at Continental Scale

A version of this chapter has been published as:

Garner R.E., Kraemer S.A., Onana V.E., Fradette M., Varin M.-P.V., Huot Y., Walsh D.A. (2023).

A genome catalogue of lake bacterial diversity and its drivers at continental scale.

Nature Microbiology 8, 1920–1934. DOI:[10.1038/s41564-023-01435-6](https://doi.org/10.1038/s41564-023-01435-6)

2.1 Abstract

Lakes are heterogeneous ecosystems inhabited by a rich microbiome whose genomic diversity is poorly defined. We present a continental-scale study of metagenomes representing 6.5 million km² of the most lake-rich landscape on Earth. Analysis of 308 Canadian lakes resulted in a metagenome-assembled genome (MAG) catalogue of 1,008 mostly novel bacterial genomospecies. Lake trophic state was a leading driver of taxonomic and functional diversity among MAG assemblages, reflecting the responses of communities profiled by 16S rRNA amplicons and gene-centric metagenomics. Coupling the MAG catalogue with watershed geomatics revealed terrestrial influences of soils and land use on assemblages. Agriculture and human population density were drivers of turnover, indicating detectable anthropogenic imprints on lake bacteria at the continental scale. The sensitivity of bacterial assemblages to human impact reinforces lakes as sentinels of environmental change. Overall, the LakePulse MAG catalogue greatly expands the freshwater genomic landscape, advancing an integrative view of diversity across Earth's microbiomes.

2.2 Introduction

Freshwater bacteria are a diverse component of lake ecosystems and are central to biogeochemical cycles and water quality (Newton *et al.*, 2011; Pernthaler, 2017; Bertilsson & Mehrshad, 2022). Collectively, Earth's millions of lakes exhibit immense environmental heterogeneity, reflecting a range of influences from regional variations in geology, landform and climate to local compositions of catchment run-off (Kratz, MacIntyre & Webster, 2005). Lakes integrate terrestrial carbon and other materials from the catchment and airshed to an extent that lakes are considered an axis of the global carbon cycle (Tranvik *et al.*, 2009) and sentinels of environmental change (Williamson *et al.*, 2008). The growing number of metagenomic studies is providing important insights into the ecology and evolution of freshwater bacteria in the context of changing environmental conditions (Grossart *et al.*, 2020). However, most studies are focused on a limited number of lakes, restricting our understanding of how environmental variation influences the structure and

function of bacterial assemblages across lake ecosystems. Elucidating freshwater bacterial genomic diversity is critical given the myriad anthropogenic pressures on lakes. Lake warming (O'Reilly *et al.*, 2015), eutrophication (Smith & Schindler, 2009), deoxygenation (Jane *et al.*, 2021), salinization (Dugan *et al.*, 2020), chemical contamination and other emergent stressors (Reid *et al.*, 2019) are expected to modify the freshwater microbiome, but in ways that we are only beginning to understand (Sheridan *et al.*, 2022).

Genome-resolved metagenomics enables the reconstruction of composite genomes from microbial populations (Tyson *et al.*, 2004). Metagenome-assembled genomes (MAGs) supply reference genomes for uncultivated taxa, revealing enormous microbial dark matter (Hug *et al.*, 2016). Currently, MAG analysis is a standard approach to link functions to populations in complex communities. Lake studies applying genome-resolved metagenomics have brought insights into the evolution (Cabello-Yeves *et al.*, 2018; Okazaki *et al.*, 2022), metabolism (Mehrshad *et al.*, 2018; Arora-Williams *et al.*, 2018; Tran *et al.*, 2021) and succession (Rodriguez-R *et al.*, 2020; Kavagutti *et al.*, 2023) of freshwater bacteria. Recently, the approach has been applied at larger scale to generate genomic catalogues from marine (Tully, Graham & Heidelberg, 2018), terrestrial (Anantharaman *et al.*, 2016) and host-associated microbiomes (Pasolli *et al.*, 2019; Nayfach *et al.*, 2019; Almeida *et al.*, 2019), culminating in a genome catalogue of Earth's microbiomes (GEM) (Nayfach *et al.*, 2021). Although impressive in its collection of habitats, GEM encompasses very few freshwater lakes (<30), compared with the immense abundance and heterogeneity of lakes worldwide. Further efforts have begun to elucidate lake genome diversity in stratified water columns (Buck *et al.*, 2021). An essential next step is to expand the representation of freshwater lakes in the global genomic catalogue.

Here we introduce a genome-resolved perspective of freshwater bacterial diversity across 6.5 million km² of Canada, the most lake-rich landscape on Earth (Messenger *et al.*, 2016). This study is a component of the Natural Sciences and Engineering Research Council (NSERC) Canadian LakePulse Network (Huot *et al.*, 2019), a large-scale coordinated effort to assess the health status of lakes using standardized sampling and analysis methods with a major focus on microbial communities (Wight *et al.*, 2020; Kraemer *et al.*, 2020, 2022; Garner, Gregory-Eaves & Walsh, 2020; Garner *et al.*, 2022; Oliva *et al.*, 2022, 2023; MacKeigan *et al.*, 2022; Griffiths *et al.*, 2022). Through metagenome co-assembly and binning, we reconstructed 1,184 high- and medium-quality MAGs from 308 lakes reflecting expansive gradients in limnological and watershed conditions. The LakePulse MAG catalogue was then used to investigate the structure and function of lake bacterial assemblages at continental scale. In addition to clarifying the functional repertoires of freshwater taxa, we pursued a novel analysis of the metabolic variation among MAG assemblages across lake physicochemical and other environmental gradients. The consideration of geomatics variables in the LakePulse survey further allowed us to report on watershed factors influencing bacterial assemblages across the land–water interface, including detectable land use effects signalling evidence of human impact. In resolving genomic diversity from the largest representation of lakes across a temperate to subarctic range, the LakePulse MAG catalogue forms a key resource for investigating the freshwater microbiome on a rapidly changing planet.

2.3 Results

2.3.1 The LakePulse MAG catalogue

To produce the MAG catalogue, we generated metagenomes from the surface waters of 308 lakes in 12 Canadian ecozones (Figure 2.1, [Extended Data Fig. 1](#) and [Supplementary Table 1](#)). Lakes represent a wide range of physicochemical conditions, productivity, morphometry, climate and land use within watersheds (Sánchez Schacht *et al.*, 2022) ([Extended Data Fig. 2](#)). At the continental scale (43–68° N, 62–141° W), a full range of ultraoligotrophic to hypereutrophic lakes were present in the survey (Figure 2.1a). Comparison of lake features by principal component analysis (PCA) revealed large-scale spatial patterns in lake and watershed characteristics (Figure 2.1b). Western Canadian lakes are the deepest and most oligotrophic on average, and often set in watersheds with high proportions of natural landscapes and harvested forests. Northern Canadian lakes experience the coldest climates and lowest land use pressures. Central Canadian lakes comprise the agriculturally intensive Prairies and Boreal Plains ecozones; compared to other Canadian regions, central Canadian lakes are generally shallow, alkaline, nutrient- and ion-rich and highly productive. Eastern Canadian lakes have the warmest surface waters and on average the most extensive built environments in their watersheds. Agriculture is also a common land use in eastern Canada, particularly within the Mixedwood Plains ecozone. Lakes in watersheds with the largest population centres are in the Vancouver and Toronto metropolitan areas within the Pacific Maritime and Mixedwood Plains ecozones, respectively.

We aimed to capture a broad phylogenetic and phenotypic diversity of microorganisms (for example, organism size and lifestyle) in the LakePulse survey. To this end, lake metagenomes were generated from biomass collected across a wide size range (0.22–100 μm). Eleven metagenome co-assemblies representing the 12 ecozones were generated (Boreal and Taiga Cordilleras lakes were combined). Although computationally intensive to produce, co-assemblies facilitated differential coverage binning, ensuring higher-quality MAGs (Albertsen *et al.*, 2013). In total, contig binning of each co-assembly recovered 1,184 MAGs exceeding Minimum Information about a Metagenome-Assembled Genome (MIMAG) medium-quality standards (Bowers *et al.*, 2017) through an additional <10% strain heterogeneity criterion ([Supplementary Table 2](#)). Between 28–233 MAGs were generated for each ecozone ([Extended Data Fig. 3](#)). MAGs were dereplicated across ecozones (average nucleotide identity (ANI) $\geq 95\%$), resulting in a genomospecies-level set of 1,008 MAGs. MAGs were categorized as 136 high- and 872 medium-quality drafts on the basis of MIMAG standards (Bowers *et al.*, 2017), excluding ribosomal (r)RNA gene requirements (Figure 2.1c). MAGs ranged widely in size (0.35–10.39 Mbp), GC content (23.1–75.9%) and coding density (78.4–96.8%) ([Supplementary Table 2](#)). Single metagenome assemblies produced only 563 bins passing our quality control ([Supplementary Table 1](#)); these represented just 202 lakes and 476 genomospecies, over one-third of which were already present in the catalogue. This comparison validated our more intensive co-assembly approach to reconstructing lake bacterial genomic diversity.

To map MAG biogeography across lakes, we calculated the central 80% truncated average depths of coverage (TAD_{80}) from fragment recruitment (Rodríguez-R *et al.*, 2020). Based on TAD_{80} , observed

richness within lakes varied between 11–172 MAGs, with diversity hotspots tending to occur in the shallow, nutrient-rich lakes of the Boreal Plains and Prairies ecozones of central Canada (Figure 2.1d). A species accumulation curve showed that hundreds of lakes were required to adequately sample genomes across the heterogeneous Canadian landscape, highlighting the importance of analysing lake metagenomes across large geographic and environmental gradients to capture the breadth of local lake bacterial diversity and ecosystem function (Figure 2.1e). MAGs captured on average 7.1% (range 0.9–36.0%) of metagenome reads per lake (Figure 2.1f), signifying that the MAG catalogue is a reasonable but incomplete representation of bacterial genome diversity in the lakes. Currently, the LakePulse MAG catalogue represents a substantial expansion in the availability of freshwater genomes and range of lake systems so far.

2.3.2 Phylogenetic and metabolic diversity of lake bacteria

The LakePulse MAG catalogue represents a phylogenetically rich set of genomes (Figure 2.2a and [Supplementary Table 1](#)) exhibiting high taxonomic novelty. Overall, 96% of MAGs are novel candidate species and over a third (349) belong to novel genera (Figure 2.2b), substantially expanding the known genomic diversity of lake bacteria. The largest collection of MAGs was recovered from Bacteroidota, a group that plays a major role in organic particle degradation in marine settings (Fernández-Gómez *et al.*, 2013) but is much less studied in freshwater systems (Bertilsson & Mehrshad, 2022). A third of Bacteroidota MAGs (80 of 275) represent new genera ([Supplementary Table 3](#)), presenting a rich genomic resource for future targeted research on freshwater Bacteroidota ecology and metabolism, including potential zooplankton associations (De Corte *et al.*, 2018). Actinobacteriota and Gammaproteobacteria are the next most MAG-rich groups and include lineages of common and abundant freshwater bacteria (for example, *Rhodoluna*, *Polynucleobacter* and *Limnohabitans*). Other freshwater taxa that are well represented in the MAG catalogue include Verrucomicrobiota, Planctomycetota, Patescibacteria and Myxococcota. Intriguingly, we identified 38 MAGs from Bdellovibrionota, almost all of which represent new genera ([Supplementary Table 3](#)). Bdellovibrionota are described as obligate predators of Gram-negative bacteria (Ezzedine, Desdevises & Jacquet, 2022) and their role in freshwater food webs warrants further attention.

To assess whether the MAGs captured a representative phylogenetic diversity of lake bacteria, we compared the taxonomy between MAGs and 16S rRNA gene amplicon sequence variants (ASVs). The same ten phyla contain the most MAGs and ASVs, although ASVs exhibit an expected long tail of rare diversity ([Extended Data Fig. 4a](#)). Moreover, the three most MAG-rich orders (Chitinophagales, Burkholderiales and Flavobacteriales) are in the top four among ASVs ([Extended Data Fig. 4b](#)). The largest differences in diversity across datasets are within Bacteroidota and Actinobacteriota, which are slightly more represented as a percentage of MAGs. In contrast, Patescibacteria exhibit higher ASV diversity and are potentially underrepresented in the MAGs because of low or divergent marker gene content (Brown *et al.*, 2015). Overall, the LakePulse MAG catalogue encompasses broad phylogenetic diversity reflective of 16S rRNA genes, making it a useful resource for further taxon-focused research.

Bacteria contribute to lake ecosystem function through a diversity of energy-capturing and carbon-cycling metabolisms that regulate biogeochemical cycles. To provide a broad overview of MAG metabolic diversity, we explored the distributions of select metabolic modules across the MAG catalogue (Figure 2.2c and [Supplementary Table 2](#)). Oxygenic photosynthetic Cyanobacteria and anoxygenic photosynthetic *Chlorobia* are present, while bacteriorhodopsins are identified in Thermoleophilia. Photosystem II is the most widespread phototrophic module in the MAGs, including in numerous Alphaproteobacteria, Gammaproteobacteria and Gemmatimonadota MAGs. In combination, photosystem II-encoding MAGs are identified in 95% of lakes, indicating a diverse bacterial contribution to freshwater phototrophy. We identified three of the six known autotrophic carbon fixation pathways (Berg, 2011), including the Calvin–Benson–Bassham (CBB) cycle, the 3-hydroxypropionate/4-hydroxybutyrate (3HP/4HB) cycle and the glyoxylate assimilation pathway. The CBB cycle is linked to Proteobacteria and Cyanobacteria MAGs present in 127 lakes (42%), demonstrating that non-algal carbon fixation is also common in lakes.

Heterotrophic degradation of complex algal and terrestrial organic matter is a central ecological function of lake bacteria. An abundance of Bacteroidota and Gemmatimonadota MAGs are equipped with SusCD-mediated glycan transport systems (Figure 2.2c and [Supplementary Table 2](#)), suggesting broad potential for polysaccharide transport and breakdown. Aromatic compound degradation pathways associated with the metabolism of plant-derived lignin are common within Alphaproteobacteria and Gammaproteobacteria, but also broadly distributed across the MAG phylogeny (Figure 2.2c and [Supplementary Table 2](#)). In addition to heterotrophy, many Burkholderiales MAGs appear to augment their energy metabolism through Sox-mediated sulfur oxidation pathways. We did not find evidence for methane or ammonia oxidation within the MAG catalogue. However, energy conservation via methanol oxidation and redox cycling of nitrogen oxides was evident. Overall, these findings highlight that the MAG catalogue captures a broad diversity of metabolisms for in-depth comparative analysis of functional contributions to lake ecosystems.

2.3.3 Lake conditions shape assemblage structure and function

Biogeographical mapping across Canadian lakes revealed that MAGs occupied on average 20 lakes (7%) and as many as 190 lakes (63%) (Figure 2.3a). Thus, the co-assembly-resolved MAG catalogue appears to comprise relatively widely distributed bacteria. Patescibacteria MAGs are among those with the most restricted ranges, suggesting relatively specialist lifestyles, perhaps dependent on host ranges (He *et al.*, 2021) or input from groundwater environments where they thrive (Tian *et al.*, 2020). Notable biogeographic patterns were observed when MAGs were aggregated at the class to phylum levels (Extended Data Figs. [5](#) and [6](#)). Phyla previously determined to be common in lakes (for example, Bacteroidota, Proteobacteria and Actinobacteriota) are broadly distributed across the continent. Other groups are relatively restricted to distinct regions. For example, *Bacilli* and *Campylobacteria* MAGs are more common in Prairies lakes than elsewhere, perhaps in response to high nutrients or extensive agriculture within watersheds (Extended Data Figs. [5](#) and [6](#)).

We next used the MAG catalogue to investigate biogeographic patterns of taxonomic and functional diversity across continental-scale environmental gradients. Taxonomic variation among MAG assemblages is correlated with lake trophic state, indexed by total phosphorus (TP) concentration (correlations with principal coordinate analysis (PCoA) axes: $r_1 = -0.32$; $r_2 = 0.33$) (Figure 2.3b). ASV community taxonomic profiles are likewise distinguished by trophic state ($r_1 = 0.41$) ([Extended Data Fig. 7a](#)), demonstrating that MAGs are reasonably representative of bacterial community variation. To assess functional variation, we generated lake functional profiles by aggregating the collection of metabolic Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologues (KOs) encoded by the MAG assemblage within each lake. Bacterial taxonomic and functional dissimilarities exhibit a positive but variable relationship (Figure 2.3c). PCoA of functional dissimilarities revealed a distribution of MAG assemblages along axis 1 related to lake trophic state ($r_1 = 0.23$) (Figure 2.3d) that reflects the relationship exhibited by metagenome functional profiles ($r_1 = 0.40$; $r_2 = -0.33$; metagenome functional profiles form the basis of a dedicated study (V.E.O. et al., manuscript in preparation)) ([Extended Data Fig. 7b](#)). These results suggest that nutrient availability and productivity play leading roles in structuring the taxonomic and functional composition of lake bacterial assemblages at continental scale.

To further elucidate the factors influencing bacterial diversity across lakes, we evaluated the importance of lake physicochemistry, morphometry, geography and climate conditions using generalized dissimilarity models (GDMs) (Ferrier *et al.*, 2007). GDMs of membership (Jaccard dissimilarities) explained more deviance compared with structure (Bray–Curtis dissimilarities) ([Supplementary Table 4](#)). As membership is more stable within lake assemblages than structure due to the seasonal succession of bacterial populations, it follows that membership is better predicted across large spatial scales. Lake physicochemistry explains the largest amounts of taxonomic (MAGs 43.9%; ASVs 48.3%) and functional (MAGs 19.4%; metagenomes 32.6%) variation among MAG assemblages and community profiles (Figure 2.3e and [Supplementary Table 4](#)). The consistency in deviance explained by taxonomic or functional models shows that MAG assemblages reliably reflect the environmental filtering of the broader community. The influence of trophic state is evident as taxonomic and functional turnover across MAGs and broader assemblages is consistently responsive to chlorophyll *a* (Chl *a*) concentration, a measure of algal biomass (Figure 2.3f and [Extended Data Fig. 8a](#)). MAG taxonomic, but not functional, turnover is significantly related to TP ([Extended Data Fig. 8a](#)), suggesting that metabolic differences among bacterial assemblages are driven more by the composition of organic matter produced by phytoplankton than by direct nutrient availability. Turnover is most rapid at low Chl *a* and TP, suggesting rapid shifts in bacterial assemblage composition across oligotrophic to mesotrophic lakes (Figure 2.3f).

In addition to the effects of trophic state, GDMs provided insights into other physicochemical factors shaping lake bacterial assemblages at continental scale. MAG and community taxonomic turnover is significantly associated with pH, and MAG functional turnover is additionally predicted by lake surface temperature (Figure 2.3g and [Extended Data Fig. 8a](#)). Temperature and pH are often considered master variables driving large-scale spatial patterns in marine (Sunagawa *et al.*, 2015) and soil (Delgado-Baquerizo *et al.*, 2018) communities, respectively. As aquatic systems that are heavily influenced by their terrestrial

surroundings, it follows that lake bacterioplankton assemblages are influenced by a combination of temperature and pH.

Remarkably, calcium concentration is the strongest predictor of both taxonomic and functional turnover across MAG assemblages and community-level taxonomy (Figure 2.3h and [Extended Data Fig. 8a](#)). A shift in bacterial ecotype diversity along a freshwater calcium gradient was previously reported (Hahn *et al.*, 2021), but how calcium directly influences assemblage composition is unclear. The effect is perhaps indirectly linked to interactions among plankton, as low calcium was previously shown to limit zooplankton growth (Azan & Arnott, 2018), potentially eliciting trophic cascades through the lake food web. Further investigation of the mechanisms and implications of the relationship between calcium and bacterial assemblages is warranted, particularly as lake calcium decline is a legacy effect of emissions-linked acid rain (Weyhenmeyer *et al.*, 2019), while calcium levels may be increasing elsewhere through winter road salt application (Kaushal *et al.*, 2018).

The demonstration that lake physicochemistry shapes the functional composition of MAG assemblages prompted a deeper analysis focused on specific metabolic categories. We generated GDMs for different KEGG metabolic categories implicated in energy and nutrient cycling. As with the full functional profiles, lake physicochemistry consistently explains more deviance than morphometry, geography or climate (Figure 2.4a). Interestingly, explained deviance varies by metabolic category and variation within carbohydrate metabolism is the best predicted by physicochemistry (Figure 2.4a). Significant predictors of carbohydrate metabolism include Chl *a* concentration and lake colour (Figure 2.4b and [Extended Data Fig. 8b](#)). Lake colour reflects terrestrial organic matter input from the watershed. Hence, variation in carbohydrate metabolism across bacterial assemblages appears to be linked to the quantity and composition of autochthonous (algal) and allochthonous (plant) organic matter in lakes. We note that lipid metabolism is the only additional metabolic category where turnover is predicted by Chl *a* and colour (Figure 2.4b and [Extended Data Fig. 8b](#)). Subsequent PCoAs of carbohydrate and lipid metabolisms in lakes show strong relationships with trophic state and large variation across the eutrophic to hypereutrophic spectrum (Figure 2.4c,d). Overall, these results suggest that a diverse and dynamic organic matter pool characterized by punctuated inputs or resuspensions of nutrients in eutrophic to hypereutrophic systems may be a key driver of metabolic diversity within the lake microbiome.

The metabolism of complex carbohydrates of aquatic or terrestrial origin depends on initial degradation through carbohydrate-active enzymes (CAZymes). MAGs contain >13,000 genes assigned to 129 glycoside hydrolase (GH) gene families implicated in cleaving the glycosidic bonds of diverse polysaccharides. Similar to carbohydrate metabolism (KEGG category), GH repertoires within bacterial assemblages are linked to lake physicochemistry (Figure 2.4a). Of GH genes, 41% are identified in Bacteroidota MAGs (Figure 2.4e) and are often encoded in polysaccharide utilization loci (PULs) ([Extended Data Fig. 9](#)), demonstrating that freshwater Bacteroidota are central to complex organic matter degradation comparable to their role in marine and gut ecosystems (McKee *et al.*, 2021). Verrucomicrobiota MAGs are also replete with GH genes (18%), expanding on previous observations that freshwater Verrucomicrobiota are central players in organic matter degradation (He *et al.*, 2017). PCA of GH distributions across MAGs

demonstrated a strong taxonomic structure for carbohydrate-processing potential in lake bacteria (Figure 2.4f). Bacteroidota display the most distinct GH family repertoires, characterized by an enrichment in GH74 implicated in cellulose and xyloglucan degradation, and GH20 implicated in anhydromuropeptide recycling and chitin degradation. Proteobacteria and Bdellovibrionota display overlapping GH repertoires distinct from other taxonomic groups and driven by GH23, GH102 and GH103, which are all lysozymes implicated in peptidoglycan degradation, aligning with potential predatory activity against other bacteria (Harding *et al.*, 2020). The thousands of GH genes encoded in the LakePulse MAG catalogue represent a rich resource to further examine organic matter degradation in freshwaters and to expand on previous global explorations of CAZyme diversity across ecosystems (Berlemont & Martiny, 2016; Garron & Henrissat, 2019).

2.3.4 Watershed effects and human impact on the lake microbiome

While the influence of lake physicochemistry on bacterial ecology has received considerable attention and is elaborated upon in this study, terrestrial influences from the surrounding watershed are less understood. The LakePulse survey was complemented with geomatics descriptions of lake watersheds, including soils and land use (Huot *et al.*, 2019). Here we coupled the geomatics data with the MAG catalogue to provide continental-scale insights into watershed impacts on lake bacterial assemblages across the land–water interface. Soil properties explain significant GDM deviance in both taxonomic (MAGs 12.1%; ASVs 23.2%) and functional (MAGs 16.3%; metagenomes 9.2%) turnover (Figure 2.5a). Notably, across all GDMs developed in this study, the importance of soil is second only to lake physicochemistry, providing evidence for terrestrial influence on lake bacterial assemblage structure and function. Turnover across MAG and community profiles is most strongly associated with soil pH and the volumetric fraction of coarse fragments, a variable describing soil texture (Figure 2.5b and [Extended Data Fig. 8a](#)). Soil texture governs water retention, while pH determines the mobility of ions and nutrients (Delgado & Gómez, 2016). Soil attributes thereby direct the composition and release of catchment run-off, which in turn modifies lake physicochemistry. Hence, terrestrial inputs from soils appear to influence lake bacterial diversity in a moderately predictable manner at continental scale.

Next, we investigated whether bacterial assemblages could be predicted by human impact on the watershed. There is accumulating evidence that the taxonomic diversity of lake bacteria is influenced by land use at large spatial scales (Marmen *et al.*, 2020; Kraemer *et al.*, 2020; Sperlea *et al.*, 2022), yet effects on function are unclear. We were particularly interested in how agriculture (Figure 2.5c) and human population density (Figure 2.5d) within watersheds may shape bacterial diversity, as each contributes to the eutrophication and chemical contamination of lakes. Land use GDMs explain a lesser but still significant amount of taxonomic (MAGs 6.8%; ASVs 9.6%) and functional (MAGs 5.4%) turnover compared with soils. There is no statistically significant response of metagenome function to land use gradients, and given that MAGs are an incomplete representation of the broader community, signals in MAG functional turnover should be interpreted cautiously. Based on evidence from GDMs, MAG assemblages are an

informative subset through which to investigate, and potentially indicate, anthropogenic pressures on watersheds altering lake bacterial assemblages across the land–water interface. The functional effects of land use on MAG assemblages may reflect the responses of taxonomic composition, yet the coupling of taxonomy and function in MAGs presents a dual-faceted resource to study human impact.

Of all human impact variables, agriculture exhibits a significant effect on both MAG taxonomic and functional turnover, while human population density exhibits a significant effect on functional turnover (Figure 2.5e). Intriguingly, when analysing specific metabolic categories, xenobiotic metabolism is the one most strongly explained by land use (Figure 2.5e). Consistent with general functional turnover (Figure 2.5f and [Extended Data Fig. 8a](#)), human population density followed by agriculture is the strongest predictor of turnover within xenobiotic metabolism (Figure 2.5g and [Extended Data Fig. 8b](#)). In both cases, the most rapid turnover occurs where human populations are densest (1,000–3,000 individuals per km²), corresponding to the few lakes sampled within the Vancouver and Toronto metropolitan areas (Figure 2.5d). The underlying effect of population density appears complex: bacterial assemblages within highly populated landscapes are characterized by a lower diversity of KOs linked to xenobiotic metabolism. Moreover, PCoA showed that although xenobiotic metabolism in lakes within high-population watersheds is distinct from that within low-population watersheds, there is also large variation in xenobiotic metabolism among lakes near population centres ([Extended Data Fig. 10](#)). As many of the detected KOs within xenobiotic metabolism are involved in the degradation of aromatic compounds, lake bacteria may be capable of metabolizing aromatic environmental pollutants, such as agricultural pesticides or industrial wastes (Phale, Sharma & Gautam, 2019). The metabolism of land-derived pollution by lake bacteria merits further investigation, considering that watersheds are increasingly converted to agriculture and urban development (Foley *et al.*, 2005), and altered precipitation patterns are increasing run-off (Kaushal *et al.*, 2014). Overall, these results suggest that the size and activities of human populations have a detectable influence on the functional diversity encoded in the freshwater microbiome, as represented in the LakePulse MAG catalogue.

2.4 Discussion

The LakePulse MAG catalogue, compiled from a continental-scale sampling of hundreds of lakes, addresses an urgent need to probe the microbiome underpinning freshwater ecosystems under rapid and accelerating anthropogenic change. This study has substantially expanded the number and diversity of genomes from freshwater ecosystems and demonstrated that lake bacteria are equipped to metabolize diverse substrates originating from both within lakes and their catchments. Trophic state emerged as a leading driver of taxonomic and functional variation among lake bacterial assemblages at continental scale, with implications for understanding bacterial responses to ongoing anthropogenic eutrophication of lakes (Smith & Schindler, 2009). The functional responses of MAG assemblages to land use pressures present a useful system for investigating human impact on the lake microbiome across the land–water interface. The links between lake bacterial diversity and watershed variables, more stable over short timescales than lake

physicochemical conditions, could be leveraged to predict freshwater assemblages using remote sensing data to decrease dependence on in situ surveys and monitoring (Oliva *et al.*, 2022, 2023). Our findings clarify that lake ecosystems cannot be studied in isolation of their watersheds in an era of intensifying land use (Foley *et al.*, 2005).

Through its construction and analysis, the LakePulse MAG catalogue makes strides towards resolving the rich and spatially heterogeneous freshwater fraction of the global microbiome. Among its potential future utilities are as an essential resource for mapping the distributions of freshwater bacteria, tracing evolutionary transitions within the land–ocean continuum, and tracking microbial dispersal and gene transfer from sources of pollution (for example, xenobiotics, antibiotic resistance (Kraemer *et al.*, 2022)) in human-altered watersheds. More lake bacterial diversity and ecosystem function remain to be captured in stratified water columns (Tran *et al.*, 2021; Buck *et al.*, 2021) and sediments (Salcher *et al.*, 2019; Biessy *et al.*, 2022), across seasons (Linz *et al.*, 2018) and potentially through deeper metagenome binning (Rodriguez-R *et al.*, 2020). Finally, microdiversity (Neuenschwander *et al.*, 2018) contained in the MAG catalogue may provide insights into intraspecific metabolic adaptations to the environmental heterogeneity among lakes and watersheds, including under future scenarios of environmental change.

2.5 Methods

2.5.1 Lake sampling

Lake sampling was conducted by the NSERC Canadian LakePulse Network between 2017 and 2019 (Huot *et al.*, 2019). Sampling was conducted in late June to mid-September when some lakes are thermally stratified. To capture the natural and human-mediated heterogeneity in lake and watershed conditions, lakes were selected across three lake area categories (small: 0.1–0.5 km², medium: 0.5–5 km² and large: 5–100 km²) and human impact categories (low, moderate and high as determined by land use types and coverage of the watershed) in 12 terrestrial ecozones of Canada.

Integrated surface water samples were collected over the euphotic zone up to 2 m below the surface at the site of maximum lake depth using an integrated tube sampler (NSERC Canadian Lake Pulse Network, 2021). The euphotic zone was estimated to be twice the Secchi disk depth. Water samples were stored in chilled coolers until they were filtered onshore on the same day. Water was pre-filtered through 100 µm nylon mesh and vacuum-filtered through 0.22 µm Durapore membranes in glass funnels at a maximum pressure of 8 inHg. Approximately 500 ml of water was filtered or until the filter nearly clogged. All sampling equipment was acid-washed and rinsed with lake water before use. Filters were stored in sterile cryovials at –80 °C.

2.5.2 Environmental data

Six categories of environmental explanatory variables were selected for analysis: (1) geography, (2) lake morphometry, (3) surface water physicochemistry, (4) watershed surface soil, (5) land use and (6) climate. Geography variables included latitude, longitude and altitude. Lake morphometry variables 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)30). Shoreline length and average depth were strongly collinear with other morphometry variables ($r \geq 0.7$) and therefore excluded from the analysis. Surface water physicochemical variables included temperature, pH, colour and concentrations of Chl *a*, dissolved inorganic carbon, dissolved organic carbon, total nitrogen, TP, calcium, chloride, magnesium, potassium, sodium and sulfate. Missing physicochemical data were replaced with ecozone medians. Watershed soil variables estimated for the top 0–5 cm soil depth interval were accessed from SoilGrids250m (ref. (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 7 days before lake sampling were accessed from ERA5-Land hourly reanalysis (Muñoz Sabater, 2019) and included mean air temperature, total precipitation, 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 ($<4 \mu\text{g l}^{-1}$), oligotrophic ($4\text{--}10 \mu\text{g l}^{-1}$), mesotrophic ($10\text{--}20 \mu\text{g l}^{-1}$), mesoeutrophic ($20\text{--}35 \mu\text{g l}^{-1}$), eutrophic ($35\text{--}100 \mu\text{g l}^{-1}$) and hypereutrophic ($>100 \mu\text{g l}^{-1}$) (Canadian Council of Ministers of the Environment, 2004). Ecozone shapefiles were accessed from Canada Ecozones (v.5b) (Wiken *et al.*, 1996).

2.5.3 DNA extraction and metagenome sequencing

DNA was extracted from filters using the DNeasy PowerWater kit (QIAGEN) according to manufacturer 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.

2.5.4 Metagenome assembly and binning

Adapter clipping and quality trimming of raw reads were performed in Trimmomatic (v.0.36) using a Phred quality score of 33 and default settings (Bolger, Lohse & Usadel, 2014). We used a co-assembly approach to leverage the differential depth of coverage across contigs and metagenomes and thus maximize the recovery of genomes common to the same ecozone. Metagenomes were co-assembled in MEGAHIT (v.1.2.7) using k-mer lengths of 27, 37, 47, 57, 67, 77 and 87 and a minimum k-mer multiplicity filter of two (Li *et al.*, 2016). Because only three lakes were sampled in the Taiga Cordillera and these bordered the Boreal

Cordillera, an ecozone with similar climate, landform and vegetation, metagenomes from these two ecozones were combined into one co-assembly.

Local alignment of metagenome reads to co-assemblies was performed with the Burrows-Wheeler Aligner (BWA; v.0.7.17) using the BWT-Smith-Waterman (BWT-SW) index algorithm and BWA-Maximal Exact Matches (Li & Durbin, 2009). Binning was performed in MetaBAT (v.2.12.1) using the default minimum scaffold length of 2,500 bp and on the basis of the average differential coverage of each contig determined by the `jgi_summarize_bam_contig_depths` script implementing a 97% sequence identity threshold (Kang *et al.*, 2019). Bin completeness, contamination and strain heterogeneity were estimated in CheckM (v.1.0.7) (Parks *et al.*, 2015). Bins scoring $\geq 50\%$ completeness, $< 10\%$ contamination and $< 10\%$ strain heterogeneity were retained as MAGs. MAGs were dereplicated at ANI $\geq 95\%$ and alignment coverage $\geq 10\%$ in dRep (v.3.2.0), implementing gANI as the second clustering algorithm (Olm *et al.*, 2017). In cases where more than one MAG belonged to the same genomospecies, the highest-quality MAG was selected as the genomospecies representative to be included in the final MAG catalogue and subsequent analysis. MAGs were categorized according to MIMAG standards by which high-quality drafts display completeness $> 90\%$, contamination $< 5\%$ and $\geq 18/20$ transfer (t)RNA genes (16S, 23S and 5S rRNA gene presence criteria were omitted), and medium-quality drafts display completeness $\geq 50\%$ and contamination $< 10\%$ (ref. (Bowers *et al.*, 2017)).

2.5.5 Functional analysis

Protein-coding genes were annotated in Prokka (v.1.12) (Seemann, 2014) which implemented Prodigal (v.2.6.3) for gene prediction (Hyatt *et al.*, 2010). RNA genes were predicted in Infernal (v.1.1.2) (Nawrocki & Eddy, 2013) against Rfam (v.14.2) (Kalvari *et al.*, 2021). Gene functions were annotated in KofamScan using default settings and a bitscore-to-threshold ratio of 0.7 (ref. (Aramaki *et al.*, 2020)). Transporters were predicted in BLASTp against the Transporter Classification Database downloaded on 27 January 2021 (ref. (Saier *et al.*, 2016)) using a maximum e-value of 10⁻²⁰. CAZymes were predicted in hmmsearch from HMMER (v.3.1b2) (Finn, Clements & Eddy, 2011) against dbCAN2 HMMdb (v.9) (Zhang *et al.*, 2018). GH family 109 was omitted from CAZyme analysis due to non-specificity of the hidden Markov model.

Functions were assigned on the basis of the following set of rules. Oxygenic photosynthesis required evidence of at least one marker gene from both photosystems I (*psaA-F*: K02689, K02690, K02691, K02692, K02693 or K02694) and II (*psbA-F*: K02703, K02704, K02705, K02706, K02707 or K02708). For anoxygenic photosynthesis, evidence of photosystem II required the presence of both *pufL* (K08928) and *pufM* (K08929); evidence of photosystem I required the presence of all four proteins (*pscA-D*: K08940, K08941, K08942 and K08943). Bacteriorhodopsin assignment required evidence of *bop* (K04641). Autotrophic carbon fixation via three independent pathways was evidenced by *rbcL* (K01601) and *prkB* (K00855) for the CBB cycle, *mct* (K14470) for the glyoxylate assimilation pathway and *abfD* (K14534) for the 3HP/4HB cycle. Evidence of glycan import mediated by the *susCD* protein complex was determined by the presence of both *susC* (K21573) and *susD* (K21572). Aromatic compound degradation was determined by the

presence of either a funnelling or ring fission pathway: evidence of catabolic funnelling was identified by ferulate to vanillin conversion (evidenced by *fcs* (K12508)), vanillin to protocatechuate conversion (evidenced by *vanA* (K03862)) or protocatechuate ring opening (evidenced by *pcaG* (K00448) or *ligA* (K04100)); ring fission was identified by gallate ring opening (evidenced by *galA* (K04099)), catechol ring opening (evidenced by K00446 or K03381), salicylate to gentisate conversion (evidenced by K00480, K18242 or K18243) or gentisate ring opening (evidenced by K00450). Methanol oxidation was evidenced by *xoxF* (K23995). Nitrogen fixation was evidenced by *nifD* (K02586), *nifH* (K02588) or *nifK* (K02591). Dissimilatory nitrate reduction was evidenced by *narG* (K00370) and *narH* (K00371) or *napA* (K02567) and *napB* (K02568). Evidence of nitrite reduction was determined by the presence of (1) *nirB* (K00362) and *nirD* (K00363), (2) *nrfA* (K03385) and *nrfH* (K15876), (3) *nirK* (K00368) or (4) *nirS* (K15864). Nitric oxide reduction was evidenced by *norB* (K04561) and *norC* (K02305). Nitrous oxide reduction was evidenced by *nosZ* (K00376). Evidence of sulfur oxidation via SOX was determined by the presence of (1) *soxA* (K17222), *soxB* (K17224), *soxX* (K17223), *soxY* (K17226) and *soxZ* (K17227) or (2) *soxC* (K17225) and *soxD* (K22622). Dissimilatory sulfate reduction was evidenced by (1) *aprA* (K00394) and *aprB* (K00395) or (2) *dsrA* (K11180) and *dsrB* (K11181).

PULs were delineated following the algorithm described in ref. (Terrapon *et al.*, 2015). First, adjacent *SusC* and *SusD* genes located on the same DNA strand (termed tandem *susCD* pairs) were identified. Tandem *susCD* pairs within five loci of each other were considered a single *susCD*-containing locus. Next, operons structured around tandem *susCD* pairs were delineated by including genes within intergenic distances of ≤ 200 bp. PUL boundaries were iteratively extended to neighbouring operons when a CAZyme was detected within the first five genes. The selected intergenic distance threshold exceeded the empirically derived 102 bp cut-off proposed in ref. (Terrapon *et al.*, 2015) to include functionally associated genes other than CAZymes that would otherwise be excluded. Only CAZymes implicated in glycan degradation (glycoside hydrolases, polysaccharide lyases, carbohydrate-binding modules and carbohydrate esterases) were included in conditional PUL extension. Finally, only PULs containing specified CAZyme types were retained.

2.5.6 MAG read recruitment and coverage normalization

To avoid recruitment to conserved regions, rRNA and tRNA gene sequences were masked in BEDTools (v.2.26.0) (Quinlan & Hall, 2010) before mapping. To determine the coverage of MAGs across all lake metagenomes, trimmed metagenome reads were initially recruited to MAGs at a sequence identity threshold of 95% in BBDMap (v.37.76) (Bushnell, 2015), then filtered with a hard cut-off of 96% identity using a custom Python script. The selection of a 95% alignment similarity cut-off reflected a prokaryotic genome ANI species delineation (Jain *et al.*, 2018). Mapping files were converted from SAM to sorted BAM format in Samtools (Li *et al.*, 2009). Horizontal coverage was calculated following the method in ref. (Rodriguez-R *et al.*, 2020) as TAD_{80} , the average sequencing depth truncated to the central 80% of mapped positions normalized by the number of genome equivalents. Genome equivalents in each metagenome were

estimated in MicrobeCensus (v.1.1.0) (Nayfach & Pollard, 2015). A positive TAD₈₀ value signified that $\geq 10\%$ of MAG positions were mapped by metagenome reads; the $\geq 10\%$ sequencing breadth threshold has been shown to detect species presence at a confidence of $>95\%$ (ref. (Castro *et al.*, 2018)).

2.5.7 Taxonomic classification and phylogenetic analysis

MAG taxonomic classification was performed in GTDB-Tk (v.1.3.0) by phylogenetic placement into bacterial reference tree (v.r95) (Chaumeil *et al.*, 2019). A de novo phylogenetic tree was constructed in GTDB-Tk using the alignment of 120 marker genes based on the Jones–Taylor–Thornton model of evolution and subsequently plotted with the R package ggtree (Yu *et al.*, 2017).

2.5.8 Metagenome analysis

Single metagenome assemblies were generated in MEGAHIT (Li *et al.*, 2016) using k-mer lengths of 23, 43, 63, 83, 103 and 123 or in metaSPAdes (Nurk *et al.*, 2017) (v.3.13.0) using k-mer lengths of 21, 33, 55, 77, 99 and 127, and mapped in BWA (v.0.7.17) using the BWT-SW index algorithm and BWA-Maximal Exact Matches (Li & Durbin, 2009). Assemblies and coverage files were submitted to the DOE-JGI Metagenome Annotation Pipeline (v.4.16.6, 5.0.20 or 5.0.23) (Huntemann *et al.*, 2016). Metagenome functional diversity was quantified by relative gene counts, calculated as KO counts normalized by total contig average depth of coverage.

2.5.9 16S rRNA gene amplicon sequencing and analysis

A ~300 bp fragment of the V4 region of the 16S rRNA gene was amplified using the primer set 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3') (Caporaso *et al.*, 2011). Each PCR mixture consisted of 5 μ l Phusion high-fidelity buffer, 0.5 μ l deoxynucleoside triphosphates (10 mM), 1.8 μ l of each primer (5 μ M), 0.25 μ l Phusion polymerase, 13.65 μ l H₂O and 10 ng in a 2 μ l volume of template DNA. PCR conditions were 30 s of 98 °C, followed by 22 cycles of 98 °C for 20 s, 54 °C for 35 s, 72 °C for 30 s and a final elongation at 72 °C for 60 s. Sequencing of 250 bp paired-end reads was performed on an Illumina MiSeq platform.

Primer sequences were trimmed in Cutadapt (v.3.1) (Martin, 2011). ASVs were inferred in DADA2 (v.1.16) (Callahan *et al.*, 2016), which performed trimming of low-quality 3'-end positions from forward and reverse reads, quality retention of reads 160–200 bp in length, ASV inference within individual samples informed by machine-learned sequencing error rates, merging of paired reads and removal of consensus chimaeric ASVs. ASVs <250 or >256 bp in length were removed along with ASVs not assigned at the kingdom level and ASVs assigned to archaea, eukaryotes or chloroplasts. ASV data from two lakes were discarded because they did not meet the threshold of 10,000 sequences. ASVs were classified against GTDB r95 16S rRNA gene sequences formatted for DADA2 (ref. (Alishum, 2020)).

2.5.10 Ecological analysis

PCA of lake features was performed in the R package *vegan* (v.2.5.6) (Oksanen *et al.*, 2020) on scaled and centred variables; latitude was passively plotted. Biogeographical analysis was conducted on MAG assemblages in 300 freshwater and oligosaline lakes, identified as having conductivity $<8 \text{ mS cm}^{-1}$ and total major ions $<4,000 \text{ mg l}^{-1}$. To build MAG functional assemblage matrices, we assumed that a function was implied when an encoding MAG was mapped to a lake metagenome. PCoA was performed in the R package *ape* (v.5.4) (Paradis & Schliep, 2019) on pairwise Jaccard dissimilarities between MAG or ASV taxonomic assemblage presence–absence data or Bray–Curtis dissimilarities between metagenome gene count data normalized by total depth of coverage. GDMs fitting assemblage turnover to environmental gradients were constructed in the R package *gdm* (v.1.4.2) using 100 permutations during each step of backward variable elimination (Fitzpatrick *et al.*, 2021). PCA of MAG GH composition was performed on Hellinger-transformed data. Data analysis and visualization were performed in R (v.4.0.1) (R Core Team, 2021) using the tidyverse (v.1.3.0) suite of packages (Wickham *et al.*, 2019).

2.5.11 Reporting summary

Further information on research design is available in the Nature Portfolio [Reporting Summary](#) linked to this article.

2.5.12 Data availability

Raw metagenome reads were archived in the European Nucleotide Archive under study accession PRJEB29238. Metagenome co-assemblies were deposited and annotated at the Joint Genome Institute (JGI) Genomes OnLine Database (GOLD) under study accession Gs0136026 and analysis projects Ga0495746 (Boreal/Taiga Cordilleras), Ga0495744 (Montane Cordillera), Ga0495745 (Pacific Maritime), Ga0495743 (Taiga Plains), Ga0485099 (Semi-Arid Plateaux), Ga0485102 (Boreal Plains), Ga0485100 (Prairies), Ga0364548 (Mixedwood Plains), Ga0373103 (Boreal Shield), Ga0372599 (Atlantic Highlands) and Ga0372598 (Atlantic Maritime). MAGs from co-assemblies and associated annotations were deposited in Dryad (Garner *et al.*, 2023) and at the European Nucleotide Archive under study accession PRJEB62834.

Publicly available datasets used in this study include HydroLakes v.1.0 (<https://www.hydrosheds.org/products/hydrolakes>), SoilGrids250m (<https://www.soilgrids.org/>), ERA5-Land (<https://doi.org/10.24381/cds.e2161bac>), Canada Ecozones v.5b (<https://ccea-ccae.org/ecozones-downloads/>), Rfam v.14.2 (<https://ftp.ebi.ac.uk/pub/databases/Rfam/14.2/>), Transporter Classification Database downloaded on 27 January 2021 (<https://tcdb.org/public/tcdb>), dbCAN2 HMMdb v.9 (<https://bcb.unl.edu/dbCAN2/download/dbCAN-HMMdb-V9.txt>), GTDB r95 (<https://data.gtdb.ecogenomic.org/releases/release95/>) and DADA2-formatted 16S rRNA gene sequences (GTDB_bac120_arc122_ssu_r95.fa.gz; <https://doi.org/10.5281/zenodo.3951383>). Source data are provided with this paper.

2.5.13 Code availability

Scripts associated with this study are available at https://github.com/rebeccagarner/lakepulse_mags.

2.6 Acknowledgements

This study was funded by the NSERC Canadian LakePulse Network (Strategic Network Grant NETGP-479720) and Canada Research Chairs held by D.A.W. and Y.H. R.E.G. and V.E.O. received scholarships from the NSERC CREATE ÉcoLac Training Program in Lake and Fluvial Ecology. R.E.G. also received support from a Fonds de recherche du Québec – Nature et technologies Doctoral Research Scholarship and the Stephen Bronfman Graduate Scholarship in Environmental Studies. The pan-Canadian field campaigns were made possible through the immense commitment and efforts of the LakePulse sampling crews and support teams, and the cooperation and assistance of Indigenous groups, municipal and park employees, lake associations and landowners. We thank the LakePulse researchers involved in the generation, management and quality control of the dataset, and acknowledge contributions from J. Juric, P. MacKeigan, C. Paquette, M. Beaulieu, K. Griffiths, G. Potvin, B. Cremella, G. Diab, V. Fugère, J. Kim, A. Oliva and K. Velghe. We also thank W. Brookes, S. Simpson and Compute Canada for bioinformatics technical support.

2.7 Author contributions

D.A.W. conceived the study and designed the analyses with R.E.G. R.E.G., S.A.K. and V.E.O. performed the bioinformatic analyses. M.F. performed the geomatic analyses. M.-P.V. coordinated the LakePulse field campaigns and data quality control. Y.H. conceived and directed the LakePulse Network. R.E.G. and D.A.W. wrote the manuscript. All authors read and improved the final manuscript.

2.8 Competing interests

The authors declare no competing interests.

2.9 Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41564-023-01435-6>.

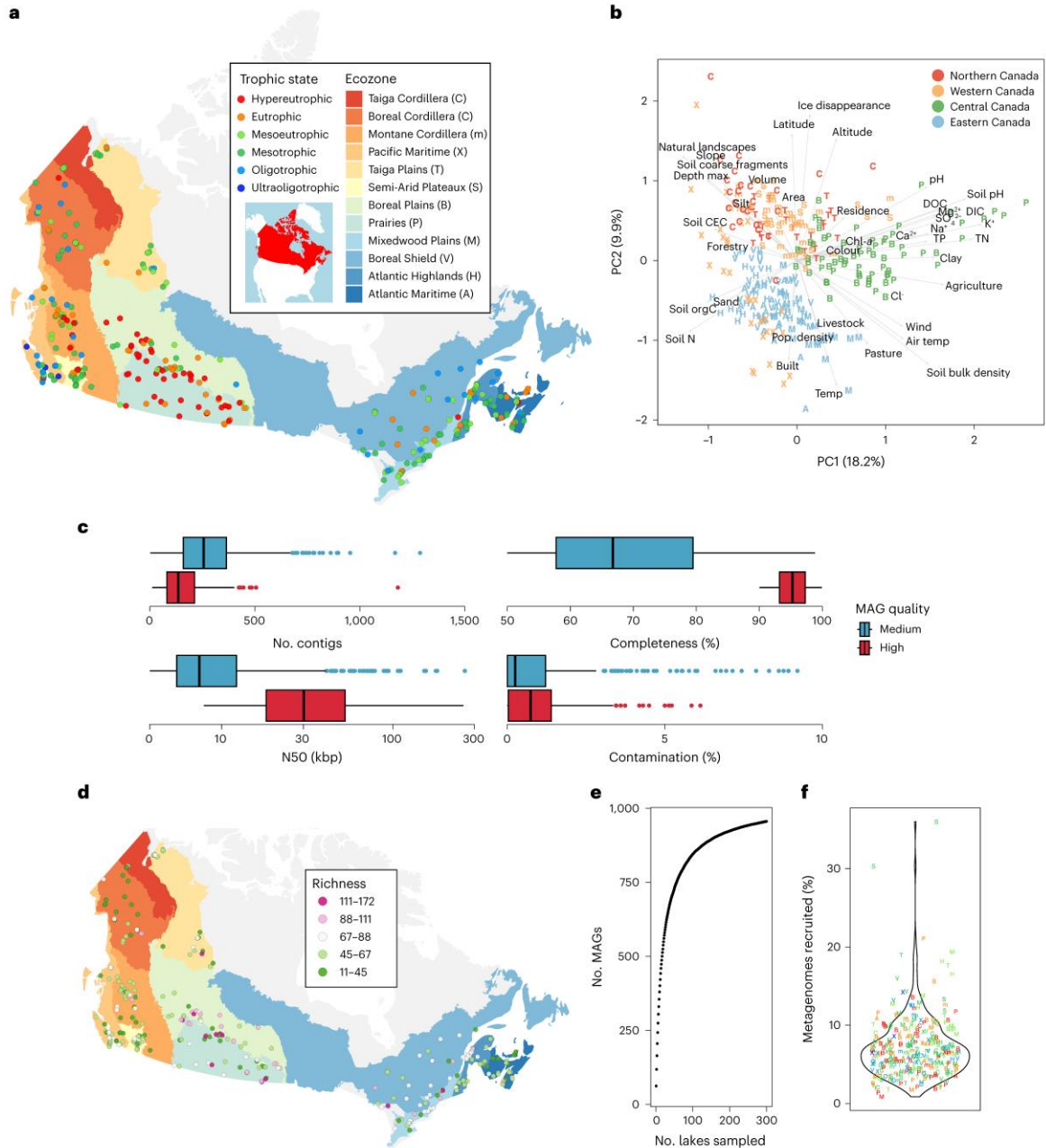


Figure 2.1 The LakePulse MAG catalogue: geographical and environmental scope, quality and patterns of local-scale diversity. The LakePulse MAG catalogue: geographical and environmental scope, quality and patterns of local-scale diversity. a, Map of 308 sampled lakes from across a broad trophic state gradient (shown as point colours) in 12 Canadian ecozones (shown as base map colours). b, PCA illustrating the variation in environmental features among sampled lakes. The letter symbols (see a for legend) and colours of each point represent ecozones and geographic regions of Canada, respectively. Arrows show the contributions of environmental variables to PC loadings, except for latitude which was passively plotted. c, Distributions of MAG characteristics for medium- and high-quality MAGs. Boxplots display median, box bounds first and third quartiles, whisker bounds minimum and maximum (except outliers), and statistical outliers evaluated as $1.5 \times$ interquartile range ($n = 1,008$ MAGs). d, MAG richness within lakes. Point colours represent richness, which is the number of MAGs detected at each site based on positive TAD80 values (see a for ecozone colour legend). e, MAG accumulation curve (that is, the number of new genomes detected) in a random ordering of lakes. Points represent the mean richness of 100 random species accumulation permutations for 1–300 lakes and error bars represent s.d. f, Fraction of LakePulse metagenomes captured within MAGs (measured as the percentage of each metagenome recruited to the MAGs). Point colours represent lake trophic states and letters represent ecozones (see legend in a).

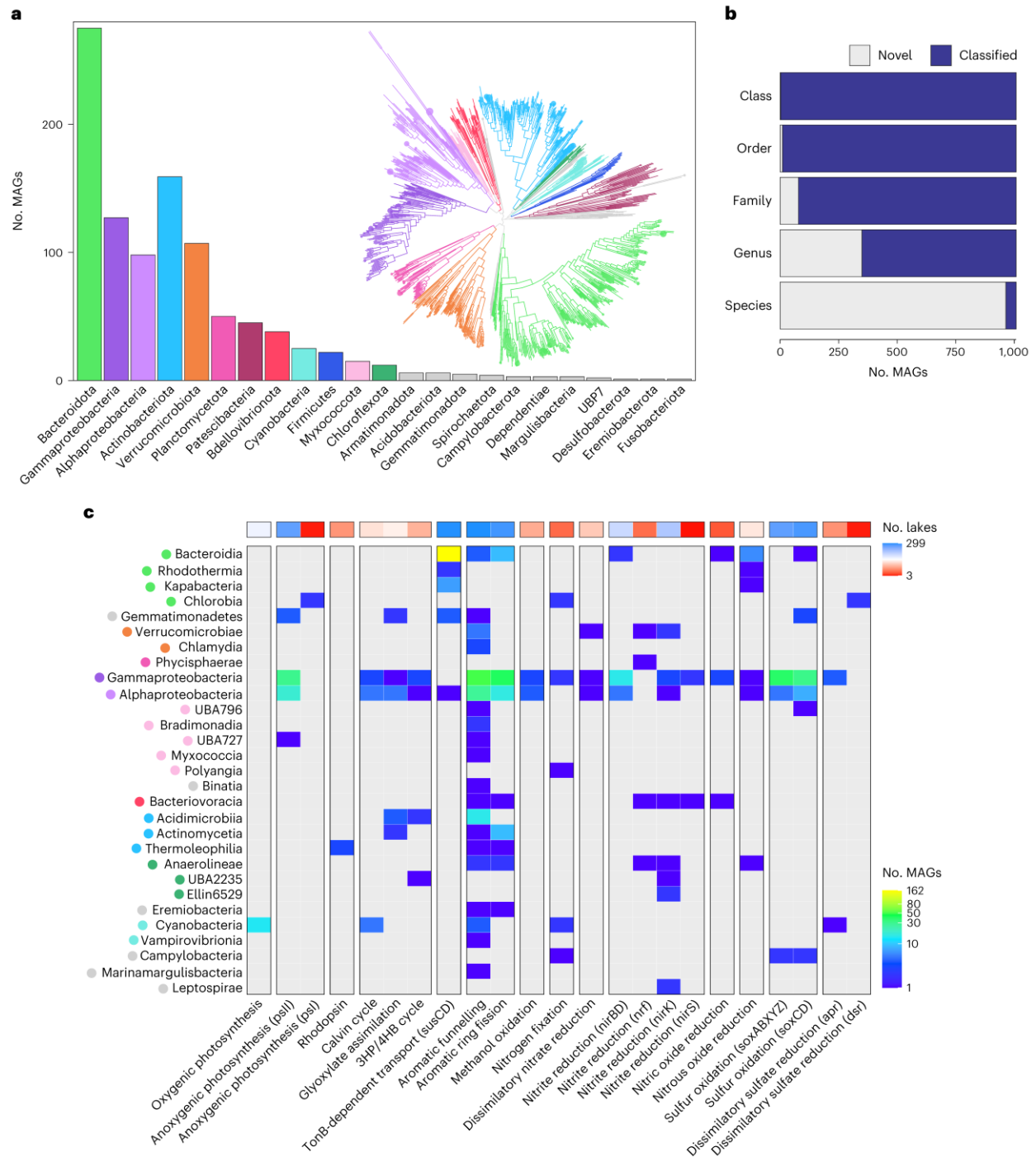


Figure 2.2 Taxonomic, phylogenetic and functional diversity of the LakePulse MAG catalogue. a, Number of MAGs assigned to each phylum and phylogenetic tree of MAGs constructed from 120 bacterial marker genes. b, Taxonomic novelty of the LakePulse MAG catalogue. Barplots are divided by colour to show the number of novel (that is, unclassified) and classified MAGs from class to species level. c, Functional diversity across taxonomic orders. The lower heat map shows the number of MAGs within each class containing genetic markers for a selection of metabolic functions (grey, no evidence; point colours next to class names indicate the phylum-level affiliation of MAGs shown in a). The upper heat map shows the number of lakes containing MAGs with evidence of metabolic functions.

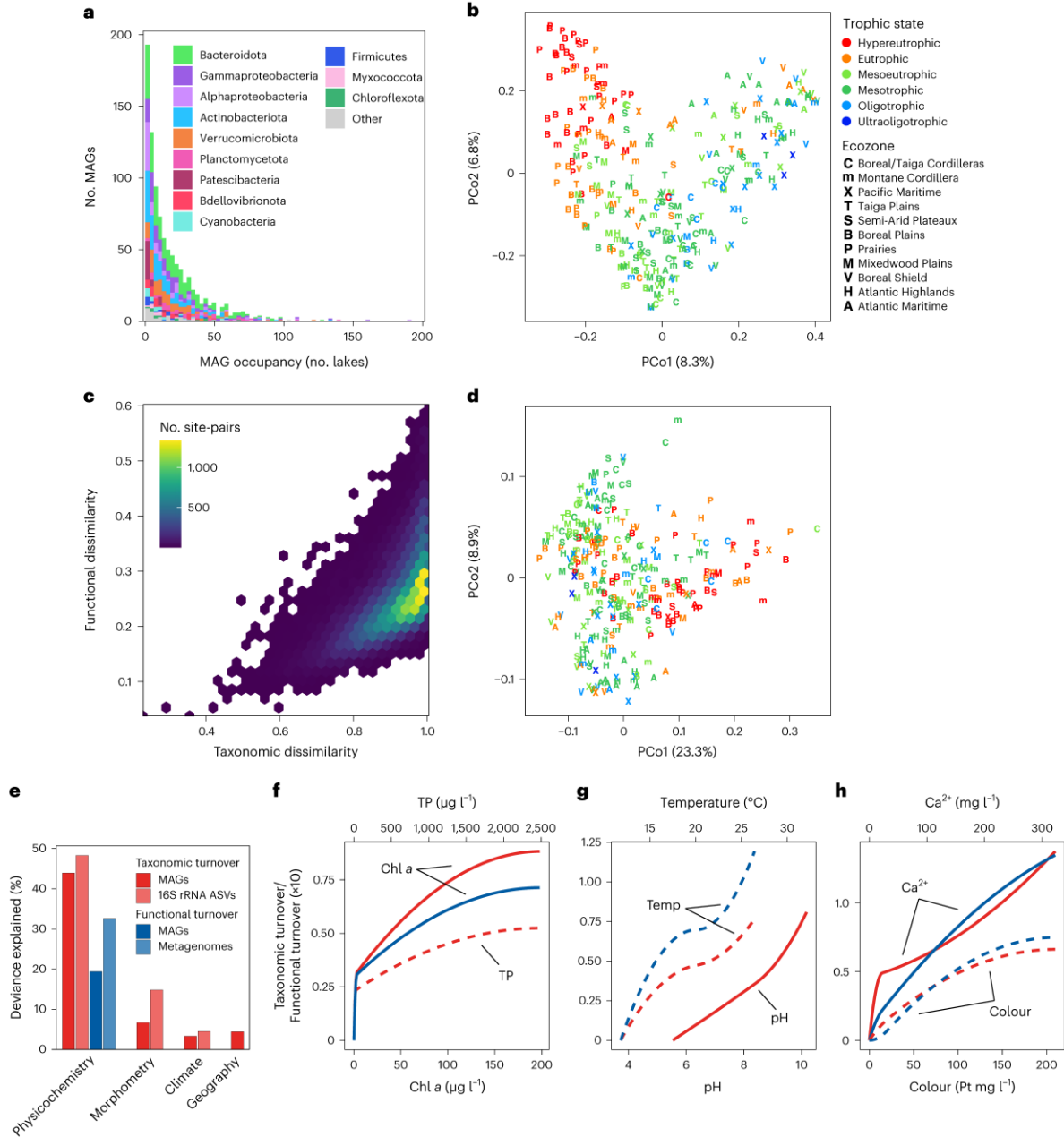
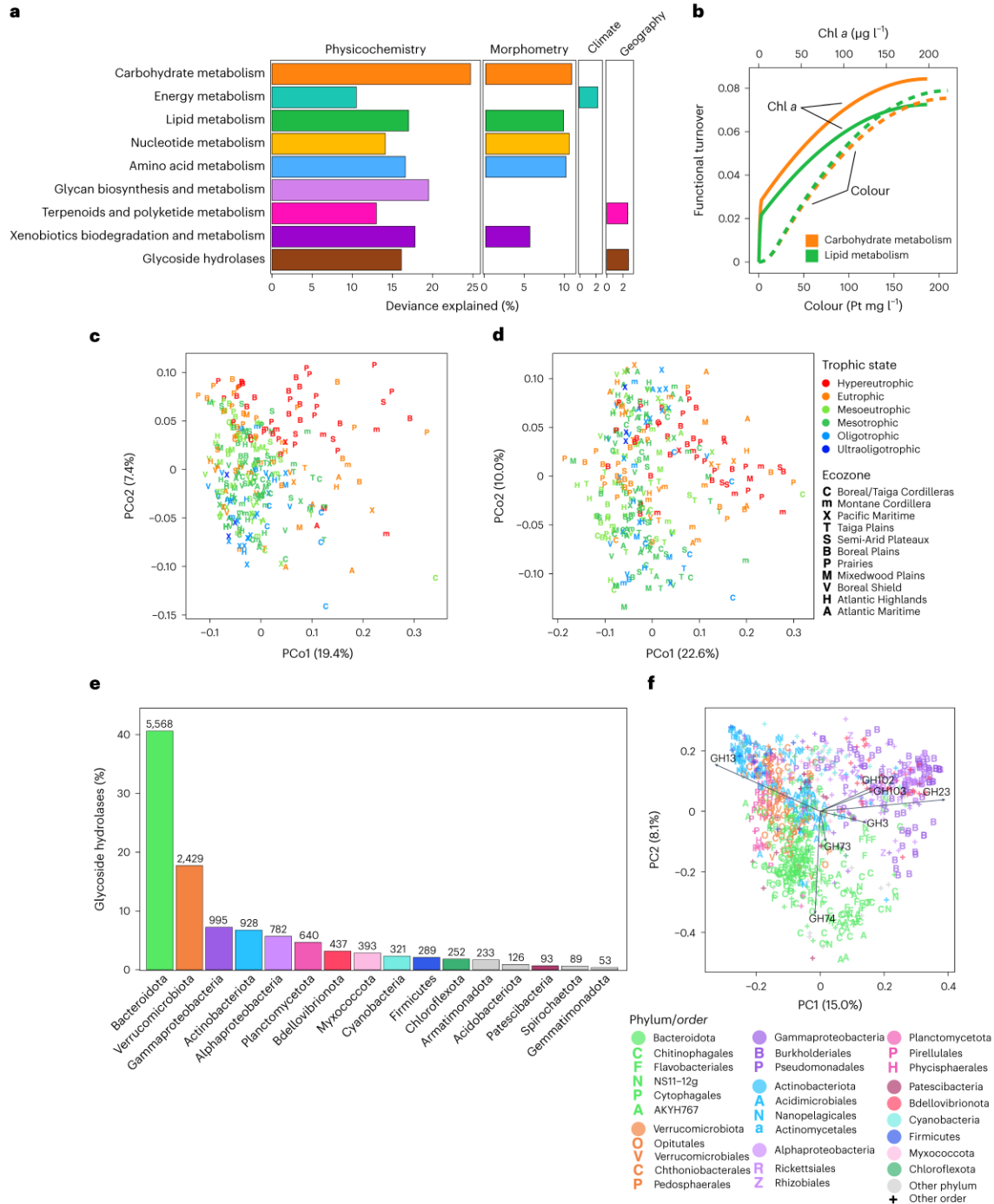


Figure 2.3 Niche breadth and environmental filtering of MAGs based on taxonomic and functional biogeography. **a**, Frequencies of MAG incidence (that is, the number of lakes occupied by each MAG). Colours show MAG phylum-level assignments. **b**, PCoA showing the variation in MAG taxonomic composition among lakes based on Jaccard dissimilarities. Point colours represent lake trophic state and letter symbols represent ecozones. **c**, Relationship between functional and taxonomic diversity among MAG assemblages in different lakes. The colour gradient shows the density of pairwise comparisons. **d**, PCoA showing the variation in functional gene content among MAG assemblages based on Jaccard dissimilarities. Point colours and letter symbols are as described in the legend of **b**. **e**, Percent deviance explained by GDMs fitting the responses of taxonomic or functional MAG assemblages or community profiles to physicochemical, morphometric, geographic or climatic gradients. **f–h**, Partial effects of total phosphorus (TP) and Chl *a* (**f**), temperature and pH (**g**), and calcium and lake colour (**h**) on the taxonomic and functional turnover of MAG assemblages across lakes. Curve shapes represent the rate of turnover and the maximum height of curves represents the magnitude of turnover across environmental gradients. Line colours represent different MAG assemblage response data (red, taxonomic

turnover; blue, functional turnover) and line types (solid or dashed) represent different environmental explanatory variables.



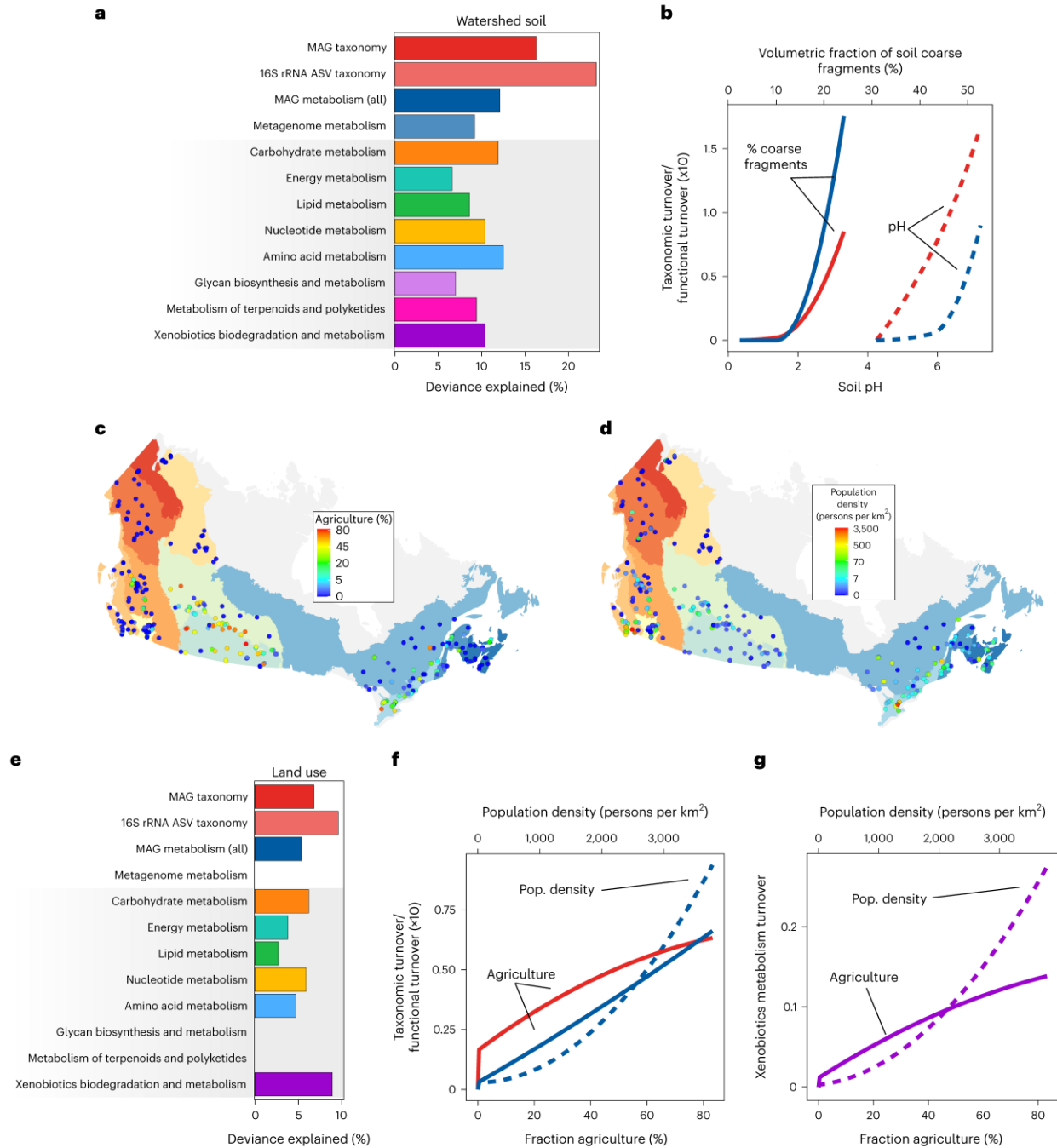


Figure 2.5 Influences of watershed soil and land use on lake bacterial assemblages. a, Percent deviance explained by GDMs fitting the responses of taxonomic or functional MAG assemblages or community profiles to watershed soil characteristics. b, Partial effects of soil pH and volumetric fraction of coarse fragments on the taxonomic and functional turnover of MAG assemblages across lakes. c, Map of proportion of land used for crop agriculture in lake watersheds. d, Map of human population densities in lake watersheds. e, Percent deviance explained by GDMs fitting the responses of taxonomic or functional MAG assemblages or community profiles to watershed land use characteristics. f,g, Partial effects of crop agriculture and human population density on the taxonomic and functional turnover of MAG assemblages across lakes (f) and the turnover of xenobiotics biodegradation and metabolism within MAG assemblages across lakes (g).

Chapter 3

Protist Diversity and Metabolic Strategy in Freshwater Lakes Are Shaped by Trophic State and Watershed Land Use on a Continental Scale

A version of this chapter has been published as:

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 7(4), e00316-22. DOI:[10.1128/msystems.00316-22](https://doi.org/10.1128/msystems.00316-22)

3.1 Abstract

Protists play key roles in aquatic food webs as primary producers, predators, nutrient recyclers, and symbionts. However, a comprehensive view of protist diversity in freshwaters has been challenged by the immense environmental heterogeneity among lakes worldwide. We assessed protist diversity in the surface waters of 366 freshwater lakes across a north temperate to subarctic range covering nearly 8.4 million km² of Canada. Sampled lakes represented broad gradients in size, trophic state, and watershed land use. Hypereutrophic lakes contained the least diverse and most distinct protist communities relative to nutrient-poor lakes. Greater taxonomic variation among eutrophic lakes was mainly a product of heterotroph and mixotroph diversity, whereas phototroph assemblages were more similar under high-nutrient conditions. Overall, local physicochemical factors, particularly ion and nutrient concentrations, elicited the strongest responses in community structure, far outweighing the effects of geographic gradients. Despite their contrasting distribution patterns, obligate phototroph and heterotroph turnover was predicted by an overlapping set of environmental factors, while the metabolic plasticity of mixotrophs may have made them less predictable. Notably, protist diversity was associated with variation in watershed soil pH and agricultural crop coverage, pointing to human impact on the land-water interface that has not been previously identified in studies on smaller scales. Our study exposes the importance of both within-lake and external watershed characteristics in explaining protist diversity and biogeography, critical information for further developing an understanding of how freshwater lakes and their watersheds are impacted by anthropogenic stressors.

3.2 Importance

Freshwater lakes are experiencing rapid changes under accelerated anthropogenic stress and a warming climate. Microorganisms underpin aquatic food webs, yet little is known about how freshwater microbial communities are responding to human impact. Here, we assessed the diversity of protists and their myriad ecological roles in lakes varying in size across watersheds experiencing a range of land use pressures by leveraging data from a continental-scale survey of Canadian lakes. We found evidence of human impact on protist assemblages through an association with lake trophic state and extending to agricultural activity and soil characteristics in the surrounding watershed. Furthermore, trophic state appeared to explain the distributions of phototrophic and heterotrophic protists in contrasting ways. Our findings highlight the vulnerability of lake ecosystems to increased land use and the importance of assessing terrestrial interfaces to elucidate freshwater ecosystem dynamics.

3.3 Introduction

Protists have evolved a vast morphological and ecological diversity (Simpson & Eglit, 2016). In aquatic ecosystems, protists play key roles in the transfer of energy and nutrients by converting sunlight into chemical energy, remineralizing organic matter, controlling microbial biomass, feeding higher trophic levels, and maintaining symbioses, some of which recruit prokaryotic metabolisms (Gadd & Raven, 2010; Worden *et al.*, 2015; Caron *et al.*, 2017).

Elucidating protist diversity in lakes is relevant for clarifying microbial distributions across a wide array of environmental conditions and for tracking the health of critical freshwaters. Covering <1% of Earth's surface (Messager *et al.*, 2016), lakes contribute disproportionately to the global carbon cycle (Tranvik *et al.*, 2009; Beaulieu, DelSontro & Downing, 2019; Anderson, Heathcote & Engstrom, 2020) and hold essential water resources (Heino *et al.*, 2021). Lakes display a rich environmental heterogeneity, generated by integrating fluxes of materials and energy from their catchments and airsheds (Kratz *et al.*, 2005; Williamson *et al.*, 2008). Freshwater lakes are hotspots of biodiversity, collectively containing higher levels of eukaryote richness and endemism (Strayer & Dudgeon, 2010) and protist community turnover (Singer *et al.*, 2021) than the marine and terrestrial realms. In the Anthropocene, lakes are increasingly altered by eutrophication (Smith & Schindler, 2009), warming temperatures (O'Reilly *et al.*, 2015), deoxygenation (Jane *et al.*, 2021), salinization (Dugan *et al.*, 2020), and myriad other persistent and emergent stressors (Reid *et al.*, 2019). There is accumulating evidence that anthropogenic modifications to lake habitats affect protist assemblages (Capo *et al.*, 2017; Keck *et al.*, 2020; Salmaso & Tolotti, 2021), in turn influencing ecosystem dynamics.

Efforts to study protists from a variety of biomes have unearthed a vast diversity and begun to map their distributions on a global scale (Massana *et al.*, 2004; de Vargas *et al.*, 2015; Mahé *et al.*, 2017). Such investigations have provided insights into protist environmental preferences and community assembly processes (Giner *et al.*, 2019; Logares *et al.*, 2020; Oliverio *et al.*, 2020; Vass *et al.*, 2020), food web dynamics (Azam *et al.*, 1983; Sherr & Sherr, 2002; Lima-Mendez *et al.*, 2015; Bjorbækmo *et al.*, 2020; Piwosz *et al.*,

2021), symbioses (Gast, Sanders & Caron, 2009; Monteil *et al.*, 2019; Husnik *et al.*, 2021), viruses (Coy *et al.*, 2018), functional traits (Ramond *et al.*, 2019), and bioindicator values (Payne, 2013; Sagova-Mareckova *et al.*, 2021). Large-scale lake surveys have shown that protist assemblages are shaped by broad biogeographic patterns (Filker *et al.*, 2016; Khomich *et al.*, 2017; Boenigk *et al.*, 2018) but are also influenced by local environmental factors and interactions with bacteria (Bock *et al.*, 2020). Phytoplankton surveys recapitulate these observations while exclusively investigating photosynthetic taxa including Cyanobacteria (Ptacnik *et al.*, 2008; Stomp *et al.*, 2011; Maileht *et al.*, 2013). However, there is no clear understanding yet of the distributions of different trophic life history strategies and the environmental drivers underlying their diversity. This knowledge gap is particularly glaring, since heterotrophy is likely the most abundant trophic mode (Keeling & Campo, 2017). Meanwhile, mixotrophs have dramatically altered our view of plankton food webs by combining primary production and prey consumption (Stoecker *et al.*, 2017), sometimes surpassing obligate heterotrophs as the leading grazers of bacteria in lakes (Berninger, Caron & Sanders, 1992; Hitchman & Jones, 2000).

In this study, 18S rRNA gene amplicon sequencing was used to investigate the distributions of protists in the surface waters of 366 freshwater lakes across a north temperate to subarctic continental range. We hypothesized that (i) protist diversity at the local scale and community turnover decrease under high-nutrient conditions; (ii) trophic groups respond to different environmental factors, and specifically, phototrophs are more sensitive than heterotrophs to bottom-up resource availability; and (iii) because lakes integrate their catchments, protist diversity in lakes should reflect watershed conditions. This project was conducted within the LakePulse survey, which sampled hundreds of lakes of different sizes in watersheds across a gradient of human impact with the primary aim to assess lake health through a multidisciplinary lens (Huot *et al.*, 2019). The current study fills a gap in the mapping of microbial biogeography through this first standardized assessment of protist diversity across Canada, which stewards the greatest abundance of lakes worldwide (Minns *et al.*, 2008). Our study draws attention to the diversity of protists and the ecological patterns that emerge in a broad collection of newly explored habitats being reshaped by increasing human impact.

3.4 Results

3.4.1 Sampled lakes and watersheds display high environmental heterogeneity

Protist assemblages were surveyed in the euphotic zones of 366 freshwater to oligosaline lakes in 12 ecozones across Canada (43 to 68°N, 53 to 141°W) (Figure 3.1). Watersheds ranged widely in area (0.3 to 9,332.3 km²) and were characterized by a variety of human population densities (0 to 3,785 people/km²) and land use, including different proportions of crop agriculture (0 to 81%) and built development (0 to 93%). Lakes had a wide range of surface areas (0.05 to 99.66 km²) and maximum depths (1 to >150 m) and were either vertically mixed (~40% of lakes) or thermally stratified (~60%) at the time of sampling. Physicochemical conditions differed substantially across lakes, as represented by a broad pH gradient (5.6

to 10.2) and ultraoligotrophic to hypereutrophic states evaluated by total phosphorus (TP) concentrations (2 to 2,484 µg/L) (Figure 3.1).

Sampled lakes and watersheds reflected regional variation in environmental conditions, including anthropogenic gradients (see Fig. S1 in the supplemental material). Lakes in northern Canada (Taiga Cordillera, Boreal Cordillera, and Taiga Plains ecozones) were subject to the coldest climates and lowest intensities and proportions of land use within their watersheds. Lakes in western Canada (Montane Cordillera, Pacific Maritime, and Semi-Arid Plateaux) were the deepest on average and located in watersheds with the largest proportions of harvested forests. In central Canada (Boreal Plains and Prairies), a region dominated by plains and prairies with extensive agriculture, lakes were generally shallow, exposed to winds, productive, and high in pH, carbon, ions, and nutrients. Lakes in eastern Canada (Mixedwood Plains, Boreal Shield, Atlantic Highlands, and Atlantic Maritime) had the warmest surface waters and generally the most built-up landscapes, a feature shared with watersheds in the Pacific Maritime. Overall, nutrient-, ion-, and carbon-rich waters, and high-pH lake conditions were most often observed in agricultural watersheds with alkaline soils (Fig. S2).

3.4.2 Lakes support taxonomically and functionally diverse protist assemblages

Eukaryotic diversity was assessed through the sequencing of 18S rRNA gene fragments amplified from DNA collected in 0.22- to 100-µm surface water particles. A total of 15,848 amplicon sequence variants (ASVs) were inferred in 17,749,930 sequences across 366 lake samples. A final data set of 13,046 putative protist ASVs encompassing 14,622,273 sequences was retained after ASVs assigned to animals, fungi, and plants were removed (Table S1). The rarefaction of pooled samples showed that the sampling of new ASVs plateaued toward 2,000,000 sequences, signaling that the global sequencing effort had exhaustively captured the protist diversity targeted by the primer pair (Fig. S3). The most abundant taxonomic groups were Ochrophyta (24% of all sequences), Cryptophyta (18%), Ciliophora (15%), and Dinoflagellata (11%) (Figure 3.1). Lineages with the highest ASV richness were Ochrophyta (22% of all ASVs), Dinoflagellata (12%), and Chlorophyta (11%).

We analyzed ASV incidence to assess the contribution of individual assemblages to total landscape diversity. New genotypes accumulated at a high rate in the first ~100 randomly ordered assemblages, followed by a gradual deceleration (Figure 3.2A). The majority of ASVs were restricted to one or a few lakes (Figure 3.2B). A smaller number of ASVs were distributed widely, including one ASV assigned to *Cryptomonas curvata* (Cryptophyta) that was ubiquitous yet highly variable in relative abundance (0.0031 to 64%) across all the lakes sampled.

Trophic functions were assigned to ASVs representing 85% of the total sequence space by leveraging natural history descriptions summarized from the literature (Figure 3.1; Table S1). Of the functionally annotated ASVs, most were classified as obligate phototrophs (32% of all sequences). The most abundant phototrophs were classified as Chrysophyceae (Ochrophyta; 16% of all sequences), Bacillariophyta (Ochrophyta; 3%), and Chlorophyceae (Chlorophyta; 5%). Bacterivory (20% of all

sequences) and mixotrophy (20%) were the next most abundant trophic modes. The most abundant mixotrophs were classified as Cryptophyceae (Cryptophyta; 17% of all sequences) and Dinophyceae (Dinoflagellata; 2%). Cytotrophy (i.e., feeding on other protists), parasitism, commensalism, saprotrophy, and osmotrophy were detected to a lesser extent. Parasites (3% of all sequences) were most abundant in the Coccidiomorphea (Apicomplexa; 2%) and Oomycota (Pseudofungi; 1%). Heterotrophs – which broadly encompassed bacterivores, cytotrophs, saprotrophs, and osmotrophs – accounted for 27% of all sequences. Heterotrophy was most abundant in the ciliates Spirotrichea (9% of all sequences), Oligohymenophorea (2%), and Litostomatea (2%) and in other lineages, including Bicoecia (Opalozoa; 3%) and Katablepharidaceae (Katablepharidophyta; 2%).

Genotypic similarity to known protist 18S rRNA gene diversity was assessed through the global alignment of ASVs to V7 region fragments in the Protist Ribosomal Reference database (PR²) (Fig. S4). Of the 12,511 ASVs that met the threshold global alignment length, the majority showed high sequence similarity with references in the database. Most ASVs (58%) were $\geq 96\%$ identical to references. ASVs with 100% similarity to reference sequences occupied 56% of the total sequence space, while ASVs with $\geq 96\%$ similarity occupied 91% of sequences. The most novel genotypes included the 738 ASVs with $< 90\%$ sequence similarity to PR² references and were either not assigned to a supergroup (415 ASVs) or primarily classified as Opisthokonta (141 ASVs), Alveolata (88 ASVs), or Stramenopiles (30 ASVs).

Local diversity at each lake was estimated by richness (67 to 1,275 ASVs) (Fig. S5), the Shannon index (0.17 to 5.67) (Figure 3.3A), Pielou's evenness index (0.04 to 0.83), and Faith's phylogenetic diversity index (8.15 to 76.04). The Shannon index was negatively correlated with magnesium ($r = -0.2033$), total nitrogen (TN; $r = -0.2032$), dissolved inorganic carbon (DIC; $r = -0.2031$), and potassium ($r = -0.2030$) concentrations (all correlations had P values of < 0.001). Analyses of variance (ANOVAs) followed by *post hoc* Tukey's tests comparing the association of trophic state with local diversity showed that mean richness, Shannon diversity, and evenness were significantly lower in hypereutrophic lakes than in eutrophic, mesoeutrophic, mesotrophic, or oligotrophic lakes (all $P \leq 0.002$) (Figure 3.3B).

3.4.3 Protist assemblages vary regionally and across lake trophic states

Next, we looked at how communities varied in taxonomic and phylogenetic composition among lakes. A principal component analysis (PCA) of ASV assemblages showed a clear pattern of taxonomic variation by lake trophic state and ecozone along the first dimension (Figure 3.4A). Assemblages in the typically nutrient- and ion-rich lakes of the Prairies and Boreal Plains were distinguished from assemblages in the lower-nutrient lakes of the Boreal Shield and other eastern regions. Cryptophyte diversity (*Cryptomonas*, *Geminigera*, *Plagioselmis*, and *Komma* species) contributed the most strongly to the variation among assemblages (Figure 3.4A). A principal coordinate analysis (PCoA) of generalized UniFrac distances between assemblages showed patterns of phylogenetic variation that were highly congruent with the taxonomic variation observed in the PCA, as evaluated by an RV coefficient correlating the first two dimensions of each ordination (RV = 67%, $P = 1.3 \times 10^{-98}$) (Figure 3.4B).

Taxonomically distinct assemblages as quantified by local contributions to β -diversity (LCBD) were mostly localized in the Prairies and Boreal Plains, with a few high-LCBD assemblages scattered across other regions (Figure 3.5). Beta-diversity partitioning showed that the taxonomic dissimilarities between assemblages were primarily generated through ASV turnover (75% of total variance) and to a lesser extent through differences in richness (25%). Highly significant ($P < 0.001$) positive correlations were detected between LCBD and a complement of physicochemical variables, including potassium ($r = 0.53$), TN ($r = 0.50$), DIC ($r = 0.45$), TP ($r = 0.44$), dissolved organic carbon (DOC; $r = 0.42$), magnesium ($r = 0.40$), sodium ($r = 0.38$), sulfate ($r = 0.37$), chlorophyll *a* ($r = 0.33$), lake colour ($r = 0.22$), and pH ($r = 0.22$). LCBD further correlated ($P < 0.001$) with watershed crop cover ($r = 0.35$) and soil pH ($r = 0.27$) and correlated negatively with soil nitrogen ($r = 0.25$). LCBD was negatively correlated with local diversity, estimated as ASV richness ($r = 0.44$) and the Shannon ($r = 0.54$) and Faith's phylogenetic diversity ($r = 0.34$) indices (all $P < 0.001$).

3.4.4 Physicochemical and watershed conditions predict community turnover

Following the previous observations of high community variability across an environmentally heterogeneous set of lakes, we evaluated the drivers of taxonomic and phylogenetic turnover based on five categories of environmental variables: (i) geography (i.e., latitude, longitude, and altitude), (ii) weather, (iii) lake morphometry, (iv) physicochemistry, and (v) watershed characteristics, including land use and edaphic properties.

We employed generalized dissimilarity models (GDMs) to detect nonlinear trends between community turnover and environmental gradients. Physicochemical factors explained the highest deviance in GDMs modeling taxonomic or phylogenetic turnover. In order of decreasing strength, DIC, TP, chlorophyll *a*, magnesium, pH, potassium, lake colour, surface temperature, and DOC were statistically significant predictors of taxonomic turnover, whereas chlorophyll *a*, magnesium, TN, pH, calcium, and colour were significant predictors of phylogenetic turnover. After physicochemistry, watershed characteristics were the most important predictors of community turnover, with both taxonomic and phylogenetic diversity responding most strongly to soil pH and then proportion of cropland cover. The volumetric fraction of soil coarse fragments and natural landscape coverage were additional predictors of taxonomic and phylogenetic turnover, respectively. Lake morphometry, comprising maximum depth, watershed slope, and shoreline circularity, had relatively weak effects on turnover. Weather and geography did not generate statistically significant GDMs. Model deviances are summarized in Table 3.1, and the partial effects of individual variables are summarized in Fig. S6 and Table S2.

3.4.5 Trophic strategies exhibit contrasting distributions

To examine the distributions of different trophic strategies, we investigated the taxonomic variation of phototroph, heterotroph, and mixotroph communities in separate PCAs (Figure 3.6). Reflecting the taxonomic variation emerging at the whole-community level, assemblages of each trophic mode were

distinguished by lake trophic state and ecozone along the first dimension. However, phototrophs displayed taxonomic variation patterns contrasting those of heterotrophs and mixotrophs. To compare the taxonomic turnover of phototrophs, heterotrophs, or mixotrophs among lakes of the same trophic state, we measured the distances of assemblages to the trophic state median within the two-dimensional principal coordinate space. The mean distance of phototroph assemblages to the trophic state median (i.e., turnover) was significantly lower in hypereutrophic lakes than in eutrophic, mesoeutrophic, mesotrophic, or oligotrophic lakes (all $P < 0.001$) (Fig. S7). In contrast, the mean distance of heterotroph assemblages was significantly higher in eutrophic lakes than in mesotrophic or oligotrophic lakes (all $P \leq 0.007$) (Fig. S7). Compared with either phototroph or heterotroph assemblages, mixotroph assemblages were highly dispersed within trophic state groups (Fig. S7).

Partitioning protist assemblages by trophic function further allowed us to examine the responses of different groups to environmental conditions. GDMs showed that phototrophs, heterotrophs, and mixotrophs each responded most strongly to physicochemical gradients. Chlorophyll *a*, DIC, potassium, and pH were predictors common to the three trophic modes, while calcium, colour, and chloride were additional predictors of phototroph turnover, and TP, surface temperature, sulfate, TN, colour, and chloride were additional predictors of heterotroph turnover. Chlorophyll *a* followed by potassium were the top predictors of phototroph turnover, while DIC followed by chlorophyll *a* were the top predictors of heterotroph and mixotroph turnover. Watershed characteristics explained an important amount of deviance in phototroph and heterotroph turnover: phototroph turnover was explained by soil pH, crop coverage, soil organic carbon density, and built development, whereas heterotroph turnover was additionally explained by soil coarse fragments. Only phototroph turnover was predicted by lake morphometry variables (watershed slope, lake maximum depth, and circularity). Heterotroph turnover was weakly explained by weather, specifically air temperature and wind speed, and geography, specifically, altitude and distances between lakes. Mixotroph turnover was the least predictable from environmental conditions, as physicochemical factors explained less deviance than for obligate phototroph or heterotroph turnover and GDMs modeled on geography, weather, lake morphometry, and watershed characteristics were not statistically significant. Model deviances are summarized in Table 3.1, and the partial effects of individual variables are summarized in Fig. S8 and Table S2.

3.5 Discussion

3.5.1 Protist diversity unveiled at a continental scale

Establishing a comprehensive perspective of freshwater protist diversity is challenging given the substantial environmental heterogeneity of the millions of lakes distributed globally. Our study fills a sizable gap by mapping protist distributions across hundreds of lakes spanning an area covering nearly 8.4 million km² in the largest study of its kind to employ a standardized sampling scheme (Huot *et al.*, 2019). Protist diversity was determined in the sunlit surface waters of 366 freshwater lakes varying in size and degree of human impact on the watershed. The sampled biomass bridged cell diameters across 4 orders of magnitude

(0.22 to 100 mm), allowing us to recover pico- to microscale organisms, many of which are not resolved under standard light microscopy and which constitute an expansive microbial diversity encompassing the major eukaryotic lineages and trophic strategies. Our global rarefaction analysis showed that the sequencing effort provided a reasonable estimate of the genotypic variation captured by the primer pair. The rapid accumulation of genotypes across sites indicated that sampling hundreds of lakes was required to assess landscape diversity. The integration of our work with the collection of recent large-scale surveys is leading to a synoptic view of protist ecology (Parfrey *et al.*, 2014; de Vargas *et al.*, 2015; Massana *et al.*, 2015; Mahé *et al.*, 2017; Khomich *et al.*, 2017; Ortiz-Álvarez *et al.*, 2018; Boenigk *et al.*, 2018; Oliverio *et al.*, 2020; Giner *et al.*, 2020).

3.5.2 Surface water assemblages contain high proportions of phototrophs and heterotrophs

Lacustrine protist diversity was dominated by ochrophytes, which accounted for both the highest sequence abundance and ASV richness. Cryptophytes, ciliates, dinoflagellates, and chlorophytes were also highly represented, reflecting taxonomic profiles typical of freshwater biomes on other continents and identified with primer pairs targeting other gene regions (Debroas *et al.*, 2017). The main taxa contributing to the dissimilarity between assemblages were mixotrophic cryptophytes, whose distributions as major bacterivores may be dependent on the occurrence of specific prey (Grujic *et al.*, 2018).

Phototrophic taxa accounted for the greatest richness and relative abundance of ASVs, perhaps not unexpectedly given that assemblages were sampled from the euphotic zone. Notably, the dominance of phototrophs within the protist fraction appears to be a distinct feature of freshwater photic zones in contrast with the prevalence of heterotrophs in the sunlit ocean and surface soils (Singer *et al.*, 2021). We showed that heterotrophs and mixotrophs were numerically important groups after phototrophs, which, along with a smaller collection of parasites, illustrates that lake surface waters harbour a broad array of microbial functions linking multiple trophic levels.

3.5.3 Protist communities respond to local environmental conditions

The key environmental drivers of protist diversity in Canadian lakes differed from those observed in previous large-scale surveys. Assemblages showed the strongest responses to physicochemical factors, including nutrient, major ion, and chlorophyll *a* concentrations, pH, and lake colour. The environmental drivers in lakes differed from those in marine and soil ecosystems, where protist diversity on a large scale is generally predicted by temperature (Ibarbalz *et al.*, 2019) and annual precipitation (Oliverio *et al.*, 2020), respectively. The differences in environmental filtering between biomes likely represent fundamental differences in the types of habitats and degrees of ecosystem connectivity. Compared with the more spatially continuous and expansive ocean and soil macroenvironments, lakes are fragmented across the landscape, and neighbouring lakes can exhibit widely contrasting physicochemical attributes (Kratz *et al.*, 2005; Martin & Soranno, 2006). Lake heterogeneity is amplified by temporal variability and punctuated perturbations. Higher community turnover has previously been measured among lakes than within marine or soil

ecosystems (Strayer & Dudgeon, 2010; Singer *et al.*, 2021), which we reassert is linked to physicochemical heterogeneity.

We found that the influence of local environmental conditions far outweighed the effects of geographic variation across the continental extent. Among LakePulse sites, systems with the highest trophic states (typically, Prairies and Boreal Plains lakes) were located at intermediate longitudes, latitudes, and altitudes. A study of Scandinavian boreal lakes spanning a longitudinally aligned and narrower trophic state gradient (oligotrophic to mesoeutrophic) reported that geography explained more protist community variation than water chemistry (Khomich *et al.*, 2017), complementing our assessment that regional physicochemical heterogeneity is a major determinant of protist diversity. Other surveys identified biogeographic patterns of protist diversity structured by the isolation and dispersal limitation of mountain lake communities (Filker *et al.*, 2016; Boenigk *et al.*, 2018). Geographic barriers (e.g., the Rocky Mountains) did not appear to generate strong compositional divisions in our set of protist assemblages but were identified as having an important influence on the distributions across LakePulse sites of crustacean zooplankton (Paquette, Gregory-Eaves & Beisner, 2021), a group with greater dispersal limitation due to their larger body sizes. Instead, we found the greatest taxonomic and phylogenetic divisions between regions distinguished by differences in lake trophic state and other local environmental conditions.

Partitioning protist trophic diversity allowed us to examine how different components of freshwater food webs respond to the environment. We observed contrasting distribution patterns for each trophic mode. Phototroph assemblages among hypereutrophic lakes exhibited significantly lower taxonomic turnover (as evaluated by mean distance to the trophic state median) than lakes at lower nutrient states, whereas heterotroph assemblages did not follow this trend but turned over significantly more rapidly among eutrophic lakes than mesotrophic or oligotrophic lakes. However, the turnover within each trophic mode was predicted by a mostly overlapping suite of physicochemical factors, including chlorophyll *a*, DIC, ions, pH, and colour, although rank order of importance varied among groups. Nutrient (TP and TN) concentrations and surface temperature were exclusive predictors of heterotroph turnover. All of the environmental predictors of mixotroph turnover were common to both phototrophs and heterotrophs, with no predictors unique to mixotrophs. Changes in low levels of chlorophyll *a* were associated with the most rapid turnover in phototroph composition, which is to be expected given that chlorophyll *a* is linked to phytoplankton biomass and phototroph diversity shifts along a lake productivity gradient. The next strongest predictor of phototroph turnover was potassium, which is not a limiting resource (Jaworski, Talling & Heaney, 2003) but displayed extreme regional variation, peaking in the Prairies, followed by the Boreal Plains. Heterotroph and mixotroph turnover was primarily predicted by DIC concentrations. Overall, the environmental drivers of heterotroph diversity mostly reflect the bottom-up controls on primary producers traversing multiple trophic levels but are rendered more complex by the added effects of nutrient and temperature factors. Bacterial prey and top-down controls, encompassing predation and parasitism (not measured in this study), likely also determine trophic functional diversity.

Mixotroph distributions were the least aligned with trophic state and the least predictable from environmental conditions, which should be expected for organisms with the metabolic versatility to occupy

variable niche spaces. The balance between primary production and prey consumption is dependent on a mixotroph's phenotypic plasticity (Wilken, Choi & Worden, 2020). While mixotrophy is competitively advantageous over obligate phototrophy or heterotrophy under low-nutrient conditions (Katechakis & Stibor, 2006), primarily phototrophic mixotrophs prevail in oligotrophic lakes and are replaced by primarily heterotrophic mixotrophs as trophic state increases (Saad *et al.*, 2016). Another variable driving mixotroph diversity in aquatic ecosystems is light availability (Fischer *et al.*, 2017; Wilken *et al.*, 2018), which is modulated by lake colour, a predictor identified in this study. Here, too, bacterial prey and zooplankton predators with a preference for nutritious mixotrophs likely also exert controls (Hansson *et al.*, 2019).

3.5.4 Hypereutrophic lakes are taxonomically distinct

Hypereutrophic lakes, located mostly in agricultural watersheds, contained protist assemblages with the lowest diversity and highest taxonomic distinctness (i.e., LCBD) relative to other lakes in the landscape. Specifically, Shannon diversity was inversely related to ion and nutrient concentrations, while taxonomic distinctness tracked with ion- and nutrient-rich conditions. Hypereutrophic conditions potentially filter protist communities by creating relatively extreme conditions (e.g., light attenuation), tolerated by a small number of taxa assembling into uneven communities distinct from those in lakes at lower-nutrient states. Phosphorus is often the limiting nutrient of phytoplankton in numerous freshwater systems (Schindler *et al.*, 1971; Liang, Soranno & Wagner, 2020), yet TP was a predictor of turnover exclusive to heterotrophs. Phosphorus was found to be an important predictor of littoral protist diversity in European lakes (Bock *et al.*, 2020) and of long-term microeukaryote community turnover evidenced from paleolimnological trends (Capo *et al.*, 2017). Interactions between protists and Cyanobacteria likely also play a role in determining protist assemblages in hypereutrophic lakes, especially as Cyanobacteria of the genus *Microcystis* were found to be associated with high-nutrient conditions across LakePulse sites (MacKeigan *et al.*, 2022).

Ordinations of separate trophic modes showed that the high community variation among eutrophic lakes was generated by heterotroph diversity, whereas obligate phototroph assemblages were the least varied under hypereutrophic conditions. A positive relationship between compositional heterogeneity and trophic state runs counter to the expectation of reduced community variability (e.g., for phytoplankton (Maloufi *et al.*, 2016)) that is predicted to follow the leveling of abiotic conditions among lakes induced by land use and eutrophication. Biotic homogenization appears to have trended with long-term climate warming and eutrophication in *Cyanobacteria* (Monchamp *et al.*, 2018) and protist (Keck *et al.*, 2020) assemblages reconstructed from sediment core chronologies. Following our observation of increased compositional heterogeneity among lakes as a function of trophic state, we posit that protist communities in productive lakes are less stable over time, including over the same season, as observed in bacterial time series (Aguilar & Sommaruga, 2020). The temporal fluxes in abiotic conditions prompting succession may be induced by allochthonous inputs or nutrient resuspension from sediments accrued at higher rates in regions of extensive lake use and high populations densities (Baud *et al.*, 2021). Furthermore, given that many Prairies and Boreal Plains lakes are shallow and exposed to winds, temporary stratification followed

by destratification is not uncommon (Mitchell & Prepas, 1990). We speculate that taxon replacement linked to land use and eutrophication may force a re-evaluation of human impact on biodiversity. In particular, anthropogenic pressures may not inherently decrease diversity but instead increase turnover, possibly at the expense of rare or specialist taxa disappearing from the landscape pool (Dornelas *et al.*, 2014).

3.5.5 Watersheds influence lacustrine protist diversity

While the importance of physicochemical factors for lacustrine protist diversity has been described (Khomich *et al.*, 2017; Bock *et al.*, 2020) and elaborated upon in this study, the influence of the watershed, in particular soil properties and land use, which are often but not entirely correlated with lake physicochemical attributes, until now has not been documented on a continental scale. We found that the taxonomic distinctness of local assemblages (i.e., LCBD) corresponded strongly with the proportion of crop agriculture in the watershed, while surface soil pH was an important predictor of community turnover. Because of their concave topographies and position in the landscape, lakes are recipients of major allochthonous subsidies, with global effects (e.g., carbon storage) that are disproportionate to the spatial extent of lakes (Tranvik *et al.*, 2009). The influence of terrestrial catchments on within-lake community dynamics is compelling evidence for why lakes cannot be studied apart from their watersheds.

The filtering of lacustrine protist diversity by watershed soil chemistry points to the influence of external factors on lake conditions. Soil buffering capacity, determined by soil texture, organic matter content, and mineral composition, is a main abiotic control on lake water pH (Dillon *et al.*, 1978). Furthermore, soil properties control the mobility of nutrients and their eventual input into lakes. Specifically, soil pH determines the availability and chemical forms of nutrients, and particle size governs the movement of groundwater carrying released nutrients (Delgado & Gómez, 2016). Soil properties also determine the composition and activity of soil microbiomes, which perform biogeochemical transformations modulating the availability of nutrients (Fierer, 2017) and seed potential colonists from soils to lakes. Altered precipitation regimes and warming temperatures associated with climate change are expected to increase soil erosion (Nearing, Pruski & O’Neal, 2004) as well as terrestrial nutrient exports (Jeppesen *et al.*, 2009).

Given the continual increase in land surface transformed by agricultural production and urbanization (Foley *et al.*, 2005), accelerated watershed land use conversions are widespread. Changes in soils associated with human activities range from increased nutrient loading and acidification by nitrogenous fertilizers in agriculture (Barak *et al.*, 1997) to shifts in carbon storage precipitated by changes in land management practices or climate (Ontl & Schulte, 2012). Moreover, land use and climate change interact to increase the frequency and magnitude of nutrient and carbon pulses to water bodies (Kaushal *et al.*, 2014). In the interest of securing a healthy future for critical freshwaters, we suggest that current soil chemistry heterogeneity across the landscape can inform predictions about the potential consequences of anthropogenic watershed alterations for lakes. In particular, work can be done to understand the microbial diversity and food web dynamics driven by various soil states in future land use scenarios. Overall, the

ability to predict lacustrine protist diversity from watershed conditions, as demonstrated in this study, highlights an expanded potential for monitoring lake ecosystems using remote sensing products (Oliva *et al.*, 2022).

3.5.6 Conclusion

This is the first study to examine the taxonomic and trophic functional variation in protist diversity across the expansive and lake-rich Canadian landscape. We showed that on this continental scale, lakes displayed broad environmental heterogeneity, including substantial variation in local physicochemical conditions driving taxonomic and phylogenetic community turnover. Watershed soil pH and crop agriculture additionally predicted community turnover and exceptional local-scale diversity. Hypereutrophic lakes were found to contain less diverse and more distinct assemblages than lower-nutrient lakes, primarily as a product of their variable heterotroph and mixotroph compositions. In contrast, phototroph assemblages were more similar among hypereutrophic lakes. While phototrophy was the prevailing nutritional strategy in lake euphotic zones, heterotrophy was nearly as numerically important; each of these trophic modes was highly predictable from physicochemical and other environmental factors. Our survey and findings serve as a valuable resource for mapping species distributions and provide a basis for future research into the increasing anthropogenic impact on lake microbiomes.

3.6 Materials and methods

3.6.1 Lake selection and sampling

Hundreds of lakes were sampled between July and early September in 2017 to 2019 by the Natural Sciences and Engineering Research Council of Canada (NSERC) Canadian Lake Pulse Network (Huot *et al.*, 2019). Sampling was timed to coincide with the summertime period of water column thermal stratification, where relevant. Lakes were sampled across 12 terrestrial ecozones, regions defined by landform, geology, and vegetation (Wiken *et al.*, 1996). Lake selection was stratified across lake surface area and watershed land use impact categories to capture natural and human-mediated lake heterogeneity. Only natural lakes with a maximum depth of at least 1 m and within 1 km from a road were considered. Freshwater to oligosaline lakes (identified as having conductivity of <8 mS/cm and total major ion concentrations of <4,000 mg/L) were selected for this analysis.

Water was collected using an integrated tube sampler from the euphotic zone over a depth of up to 2 m below the surface at the deepest point in the lake (NSERC Canadian Lake Pulse Network, 2021). The site of maximum lake depth was located by depth sounding with the aid of bathymetric maps where available. The depth of the euphotic zone was estimated as twice the Secchi disk depth. All water sampling equipment was acid washed and rinsed three times with lake water before use. Carboys were stored in icepack-chilled coolers until water filtration later in the day. Water was prefiltered through 100- μ m nylon mesh and vacuum filtered on 47-mm-diameter 0.22- μ m Durapore membranes through a glass funnel at a

maximum pressure of 8 inHg. Up to 500 mL of water was filtered until the filter was nearly clogged. Filters were stored in sterile cryovials at -80°C.

3.6.2 18S rRNA gene amplification and sequencing

DNA was extracted using the DNeasy PowerWater kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions with the addition of two optional steps: after bead beating and centrifugation, 1 µL RNase A was added to samples, followed by 30 min incubation at 37°C. DNA was quantified using the Qubit dsDNA BR assay (Invitrogen, Carlsbad, CA, USA). A ~265-bp fragment of the 18S rRNA gene V7 region was amplified with the primers 960F (5'-GGCTTAATTTGACTCAACRCG-3') (Gast, Dennett & Caron, 2004) and NSR1438 (5'-GGGCATCACAGACCTGTTAT-3') (Van de Peer, 2000) to broadly target microeukaryotes (Capo *et al.*, 2016). Each PCR contained a total 25-µL mixture of 14.25 µL Milli-Q water, 5 µL 5X High-Fidelity buffer, 1.25 µL of each 10 mM primer, 0.5 µL 10 mM deoxynucleoside triphosphates (dNTPs), 0.25 µL dimethyl sulfoxide (DMSO), 0.5 µL Phusion DNA polymerase (Thermo Fisher Scientific, Waltham, MA, USA), and 2 ng DNA template. PCR conditions consisted of an initial denaturation at 98°C for 1 min, 30 cycles of 98°C for 10 s, 60°C melting for 30 s, and 72°C for 20 s, and a final extension at 72°C for 5 min. PCR products were loaded with Orange G dye on an ethidium bromide-stained 2% agarose gel and electrophoresed at 40 V for 100 min. DNA bands aligned at the target fragment length against a 100-bp DNA ladder were excised with razor blades and gel extracted with the QIAquick gel extraction kit (Qiagen, Hilden, Germany), modified by final elution into Milli-Q water. PCR products were submitted to Genome Quebec for library barcoding and sequencing of 250-bp paired-end reads in three sequencing runs on an Illumina MiSeq platform.

3.6.3 ASV inference and annotation

Primer sequences were removed in Cutadapt v. 3.1 (Martin, 2011). Trimmed reads were processed into ASVs through DADA2 v. 1.16 (Callahan *et al.*, 2016). Samples were pooled for ASV inference using otherwise default parameters. Taxonomy was assigned with naive Bayesian classification trained on PR² v. 4.12.0 (Guillou *et al.*, 2012). Potentially spurious ASVs were removed by visually inspecting a de novo alignment performed in MAFFT (Katoh, Rozewicki & Yamada, 2019). To retain only putative protist ASVs, ASVs assigned to Metazoa, Fungi, and Embryophyceae were removed. Taxa were assigned to trophic functional groups as either photoautotrophs, heterotrophs (bacterivores, cytotrophs, saprotrophs, or osmotrophs), mixotrophs, or parasites according to lineage-specific feeding habits summarized by Adl *et al.* (2019).

3.6.4 Sequence similarities with known diversity

A database of 18S rRNA gene references restricted to the V7 region was constructed by applying the 960F/NSR1438 primer pair to PR² v. 4.13.0 in Cutadapt. ASV top hits were queried against the PR² database

in BLAST v. 2.6.01 (Camacho *et al.*, 2009). Sequence identities were reported for ASVs that were globally aligned to references over a length ≥ 220 nucleotides.

3.6.5 Environmental data collection

Lake trophic states were assigned based on TP concentration thresholds estimated for Canadian freshwater systems: ultraoligotrophic (TP concentration, $<4 \mu\text{g/L}$), oligotrophic (4 to $10 \mu\text{g/L}$), mesotrophic (10 to $20 \mu\text{g/L}$), mesoeutrophic (20 to $35 \mu\text{g/L}$), eutrophic (35 to $100 \mu\text{g/L}$), and hypereutrophic ($>100 \mu\text{g/L}$) (Canadian Council of Ministers of the Environment, 2004). Meteorological conditions recorded over 7 days leading up to sampling and ice disappearance day data were accessed from 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 (Messenger *et al.*, 2016). Watershed surface soil properties were accessed from SoilGrids250m (Hengl *et al.*, 2017). Land cover information was compiled as described by Huot *et al.* (2019). Maps were constructed in R with the NAD 83 coordinate reference system and using the coordinates of Canada from the package maps (Becker *et al.*, 2018) and ecozone shapefiles sourced from the Canada Council of Ecological Areas (Wiken *et al.*, 1996).

Environmental data were categorized into thematic groups of variables. Latitude, longitude, and altitude were categorized as geography variables. Ice disappearance day and meteorological variables (air temperature, precipitations, and net solar radiation) were categorized as weather variables. Lake surface area, circularity, volume, maximum depth, discharge, residence time, watershed slope within 100 m of the shoreline, watershed area, and lake-to-watershed area ratio were categorized as lake morphometry variables. Watershed land use (crop agriculture, pasture, built development, and clear-cut forestry) and natural land cover fractions, human population density, and mean surface soil properties (bulk density of the fine earth fraction, cation exchange capacity, nitrogen, pH, organic carbon density, organic carbon content in the fine earth fraction, volumetric fraction of coarse fragments, clay, sand, and silt) were categorized as watershed variables. Surface water temperature, calcium, magnesium, potassium, sodium, chloride, sulfate, TP, TN, DIC, DOC, and chlorophyll *a* concentrations, pH, and lake colour were categorized as lake physicochemical variables. Missing physicochemical data were replaced with ecozone median values. Highly colinear variables, evaluated by a Pearson's correlation *r* value of ≥ 0.7 , within all categories except physicochemistry were removed.

3.6.6 Diversity analyses

ASVs were aligned in the SILVA Incremental Aligner v. 1.7.2 (Pruesse, Peplies & Glöckner, 2012) against the SILVA 138.1 SSU Ref NR 99 database (August 27, 2020 release) (Quast *et al.*, 2012). A maximum-likelihood phylogeny was constructed in FastTree v. 2.1.11 using the Generalized Time-Reversible model of nucleotide evolution (Price, Dehal & Arkin, 2010). Phylogenetic dissimilarities between ASV assemblages were calculated as generalized UniFrac distances, which are sensitive to compositional changes in lineages

of intermediate abundance (Chen *et al.*, 2012). Generalized UniFrac distances ($\alpha = 0.5$) were computed using the GUniFrac package.

To deal with uneven total sequence abundance across samples, sequence count composition was scaled to relative abundance. Rarefaction analysis was conducted on the total data set 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 package *vegan* (Oksanen *et al.*, 2020). Local diversity indices (richness, Pielou's evenness, Shannon diversity, and Faith's phylogenetic diversity) were calculated from rarefied community data (i.e., randomly subsampled to the lowest sample abundance equaling 10,069 sequences) in the R packages *vegan* (Oksanen *et al.*, 2020) and *picante* (Kembel *et al.*, 2010). PCAs were computed on Hellinger-transformed community data in *vegan*.

LCBD and β -diversity partitioning analyses were performed on β -diversity estimated using 100 permutations from Hellinger-transformed community data in the package *adespatial* (Dray *et al.*, 2021). RV coefficients were computed from the first two principal components or coordinates in the package *FactoMineR* (Lê, Josse & Husson, 2008). Nonlinear relationships between β -diversity and untransformed environmental gradients were modeled in GDMs (Ferrier *et al.*, 2007; Rosauer *et al.*, 2014) in the package *gdm* (Fitzpatrick *et al.*, 2021). To create GDM site-pair tables, 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. To assess the dispersion in taxonomic composition among lakes of the same trophic state, the mean Bray-Curtis distances of assemblages to trophic-state medians (i.e., centroids) calculated across the first two principal coordinates were compared using ANOVAs, followed by *post hoc* Tukey's tests.

Data wrangling and statistical analysis were performed in R v. 4.0.2 (R Core Team, 2021).

3.6.7 Data availability

Sequence data have been deposited in the European Nucleotide Archive under study accession number PRJEB42538 (www.ebi.ac.uk). Scripts are accessible from https://github.com/rebeccagarner/lakepulse_protists.

3.7 Acknowledgements

This research was funded by the NSERC Canadian Lake Pulse Network (Strategic Network Grant NETGP-479720) and Canada Research Chairs held by D.A.W., I.G.-E., and Y.H. R.E.G. and V.E.O. were supported by the NSERC CREATE ÉcoLac training program in lake and fluvial ecology. R.E.G. also acknowledges support from a *Fonds de recherche du Québec – Nature et technologies* doctoral research scholarship and the Stephen Bronfman Scholarship in Environmental Studies.

We thank the LakePulse sampling crews and the many landowners, municipal and park employees, lake associations, and First Nations who welcomed and facilitated the sampling effort. We also thank the researchers in LakePulse who contributed to the data set, Zofia Taranu for helpful discussions on statistical

methods, and Paul MacKeigan, Cindy Paquette, Marieke Beaulieu, and Katherine Griffiths for their collaboration on data quality assessment and curation.

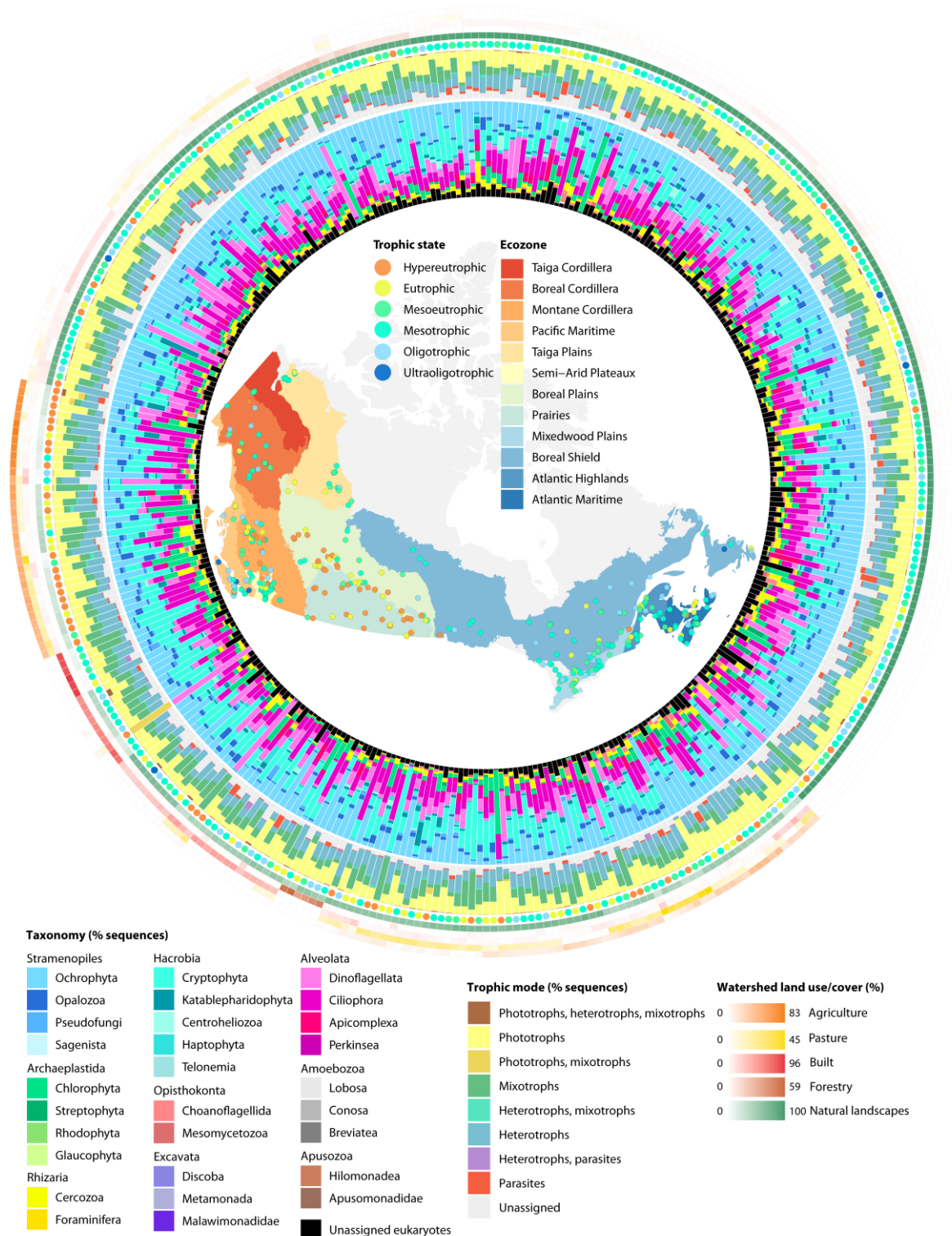


Figure 3.1 Diversity and distributions of protists across 366 Canadian lakes. Lake trophic states and ecozones are shown on the map of sampling sites as coloured circles and polygons, respectively. The relative sequence abundance of protist taxonomic divisions in each lake is represented in the inner tract of bar plots. The middle tract of bar plots shows the relative sequence abundance of trophic modes. The heat map on the outer edge illustrates the proportions of land use and land cover associated with the watershed of each lake, whose trophic state is represented in an adjacent coloured circle. Watershed land use proportions are hierarchically clustered to highlight the relationship between agriculture and trophic state.

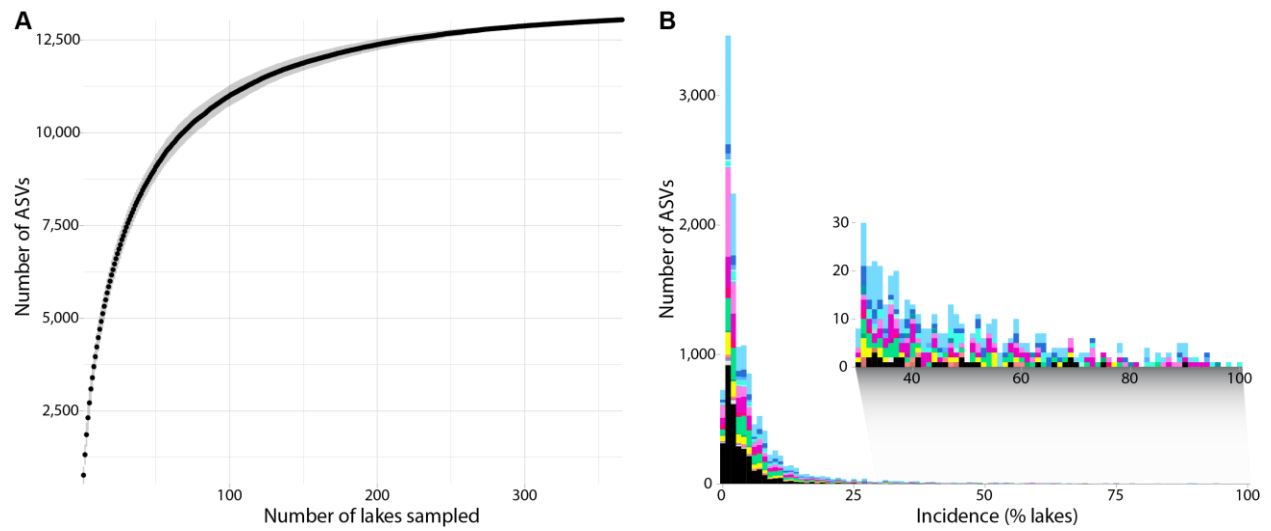


Figure 3.2 Accumulation and incidence of genotypes across lakes. (A) Accumulation curve of genotypes in a random ordering of lakes. Vertical bars are standard deviations. The accumulation of new genotypes was rapid in the first ~100 lakes, followed by a gradual deceleration. (B) Incidence of genotypes across lakes. Most taxa are distributed across one or a few lakes, whereas a few taxa (magnified in the inset) are widely distributed. ASV taxonomic classifications are coloured according to the taxonomic divisions in Figure 3.1.

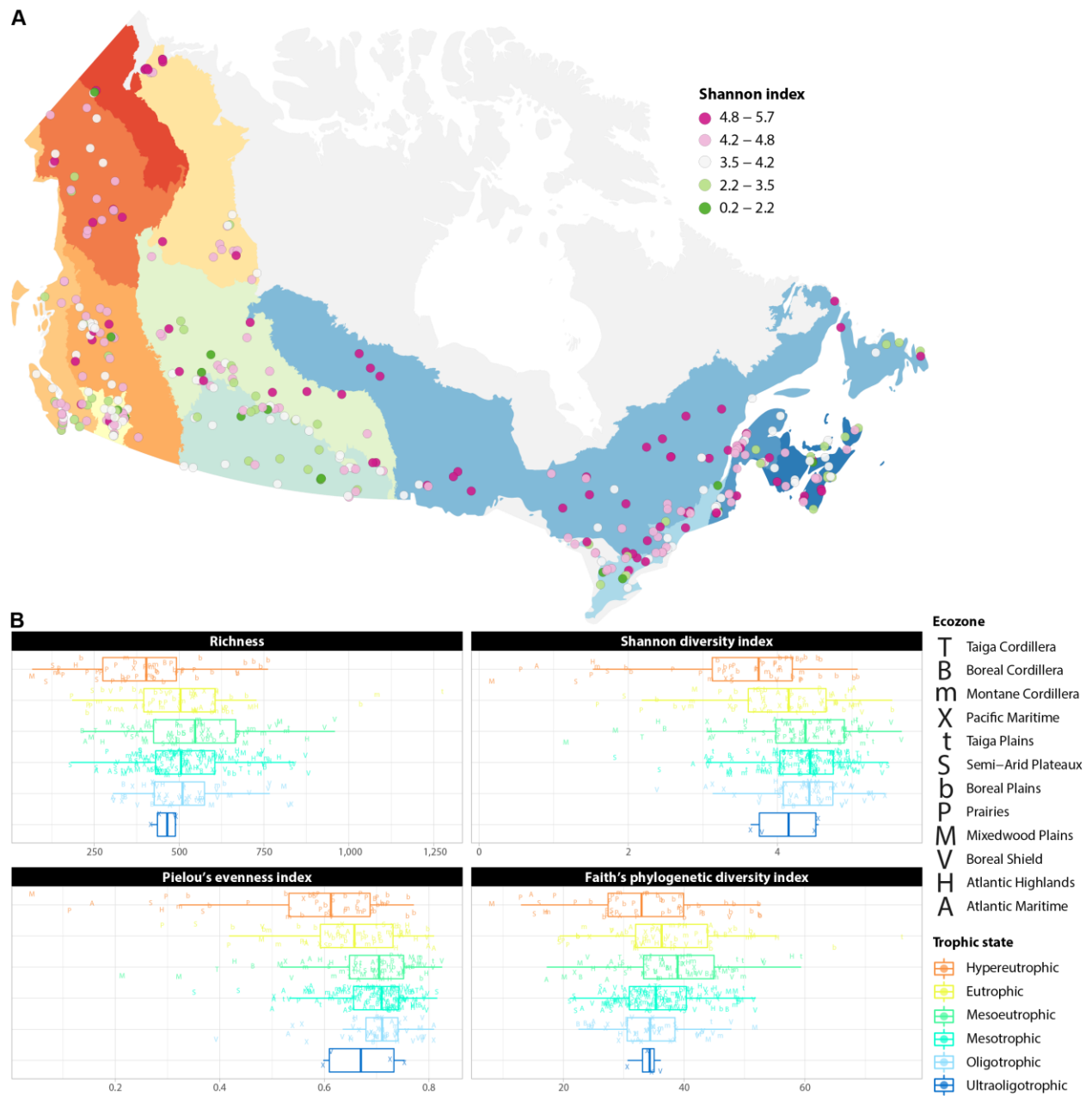


Figure 3.3 Local protist diversity of each lake. (A) Rarefied Shannon diversity index calculated for each protist assemblage across the Canadian landscape. Ecozones are identified in the key of Figure 3.1. (B) Local diversity metrics categorized by lake trophic state. TP concentration thresholds and sample sizes of lakes categorized under each trophic state are as follows: ultraoligotrophic (TP concentration, $<4 \mu\text{g/L}$; $n = 4$), oligotrophic (4 to $10 \mu\text{g/L}$; $n = 48$), mesotrophic (10 to $20 \mu\text{g/L}$; $n = 132$), mesoeutrophic (20 to $35 \mu\text{g/L}$; $n = 71$), eutrophic (35 to $100 \mu\text{g/L}$; $n = 61$), and hypereutrophic ($>100 \mu\text{g/L}$; $n = 50$).

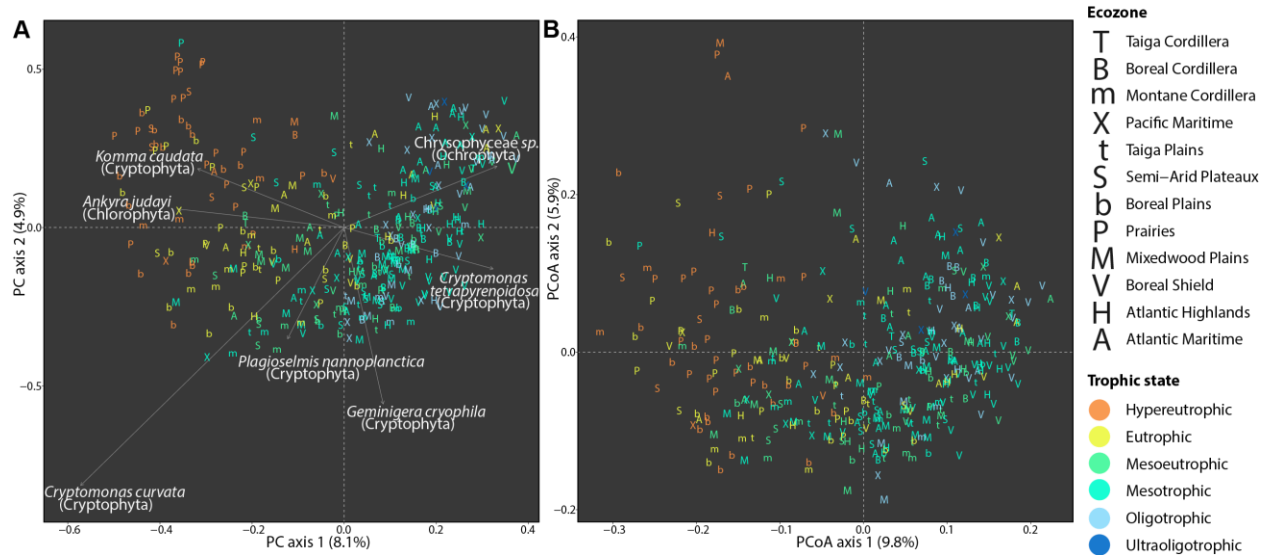


Figure 3.4 Taxonomic and phylogenetic variation of protist assemblages among lakes. (A) PCA of the taxonomic variation among protist assemblages at the level of individual ASVs. The relative contributions and taxonomic assignments are shown for the top 7 ASVs (an arbitrary cutoff selected for illustrative clarity) contributing to the variation explained by the first two principal component dimensions. (B) PCoA of the phylogenetic variation among protist assemblages.

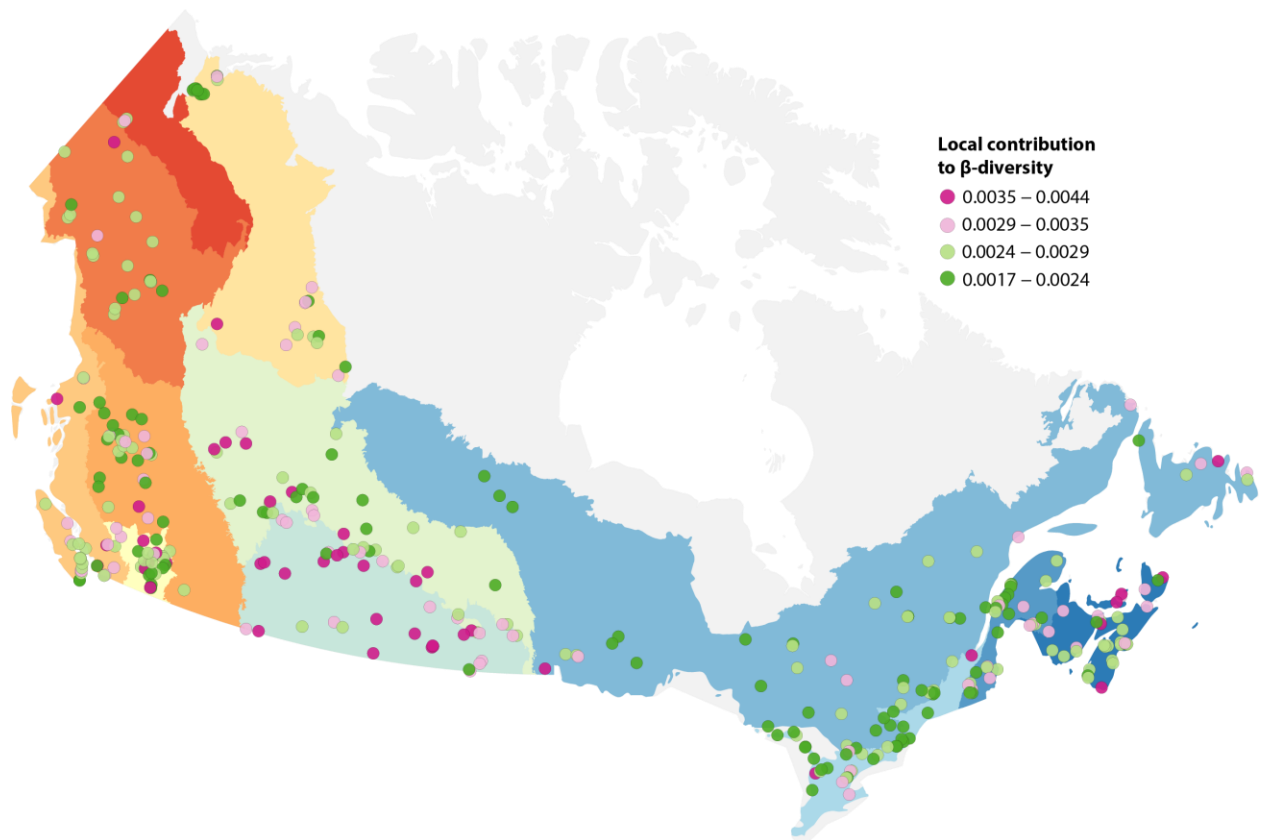


Figure 3.5 Local contribution to β -diversity (LCBD) of protist assemblages across the Canadian landscape. LCBD describes the taxonomic uniqueness of a given assemblage, i.e., how much the taxonomic composition differs from the rest of the communities in the landscape. Ecozones are identified in the map legend of Figure 3.1.

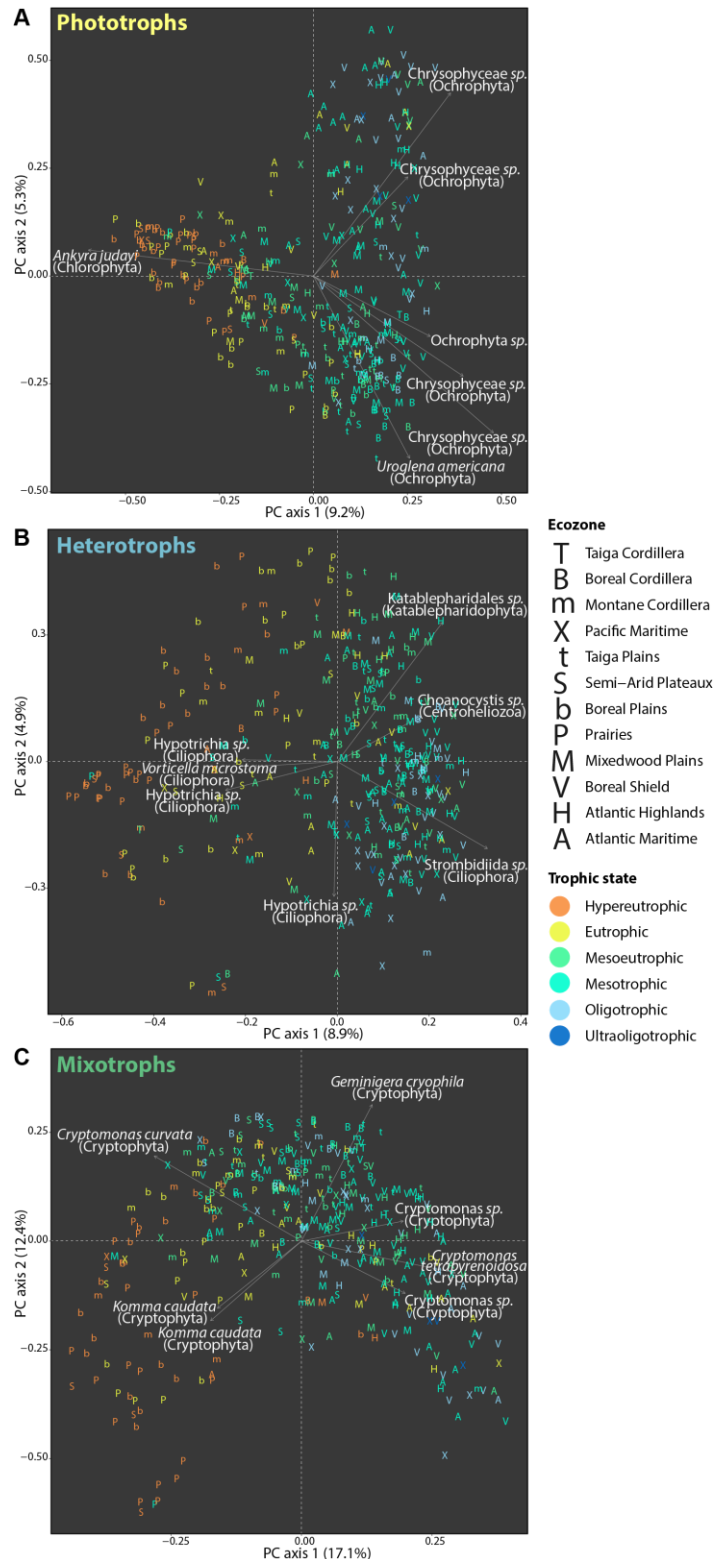


Figure 3.6 PCA of the taxonomic variation among (A) phototroph, (B) heterotroph, and (C) mixotroph assemblages at the level of individual ASVs. The relative contributions and taxonomic assignments are shown for the top 7 ASVs contributing to the variation explained by the first two PC dimensions. Ecozone affiliations and trophic state classifications of lakes are represented by letter symbols and colour, respectively.

Table 3.1 Percent deviance explained by generalized dissimilarity models (GDMs) fitting the responses of protist assemblages to environmental gradients^a.

Explanatory variables	Deviance explained (%)				
	All protists		Trophic mode ^b		
	Taxonomy	Phylogeny	Phototrophs	Heterotrophs	Mixotrophs
Physicochemistry	38	33	35	42	18
Watershed	15	16	20	19	NS
Morphometry	6	8	5	NS	NS
Weather	NS	NS	NS	3	NS
Geography	NS	NS	NS	2	NS

^aGDMs were constructed using various community response data and categories of environmental explanatory variables. NS, model was not statistically significant ($P \geq 0.05$). ^bAnalysis was performed only on taxonomic composition response data.

3.8 Supplemental material

Figure S1. Distributions of (A) geography, (B) lake morphometry, (C) watershed, (D) weather, and (E) physicochemical variables by ecozone. Dashed lines denote ecozone medians.

Figure S2. Principal component analysis of sites based on lake and watershed environmental characteristics.

Figure S3. Rarefaction curve showing the number of ASVs detected as a function of sample size across all sites.

Figure S4. Sequence similarities of ASVs to known 18S rRNA gene V7 region diversity. (A) Frequency of ASVs as a function of sequence identity with the top hit reference in PR². (B) Number of sequences associated with the sequence identities of top hits. Bar colour represent the division-rank taxonomic assignment top hit references in PR².

Figure S5. Map of rarefied ASV richness.

Figure S6. Partial effects of (A) physicochemical, (B) watershed, and (C) lake morphometry variables on the taxonomic and phylogenetic turnover of protist assemblages. Points along the x axis denote the distribution of explanatory variables and are coloured according to lake trophic state (see the map key in Figure 3.1).

Figure S7. Phototroph, heterotroph, and mixotroph assemblage distances from trophic state centroids.

Figure S8. Partial effects of (A) physicochemical, (B) watershed, (C) lake morphometry, (D) weather, and (E) geography variables on the taxonomic turnover of phototroph, heterotroph, and mixotroph assemblages. Points along the x axis denote the distribution of explanatory variables and are coloured according to lake trophic state (see the key in Figure 3.1).

Table S1. ASV taxonomic and trophic functional annotations.

Table S2. Explanatory variables and relative importance of individual predictors in GDMs. Partial effects (i.e., relative importance) of individual predictors within models (parenthetical and presented in descending order of variable importance) were determined as the sum of I-spline coefficients equaling the

magnitude of turnover. NS, nonsignificant model. *, For models of turnover in trophic mode assemblages, analysis was performed only on taxonomic composition response data.

Chapter 4

Sediment Metagenomes as Time Capsules of Lake Microbiomes

A version of this chapter has been published as:

Garner R.E., Gregory-Eaves I., Walsh D.A. (2020).

Sediment Metagenomes as Time Capsules of Lake Microbiomes.

mSphere 5(6), e00512-20. DOI:[10.1128/msphere.00512-20](https://doi.org/10.1128/msphere.00512-20)

4.1 Abstract

The reconstruction of ecological time series from lake sediment archives can retrace the environmental impact of human activities. Molecular genetic approaches in paleolimnology have provided unprecedented access to DNA time series, which record evidence of the microbial ecologies that underlaid historical lake ecosystems. Such studies often rely on single-gene surveys, and consequently, the full diversity of preserved microorganisms remains unexplored. In this study, we probed the diversity archived in contemporary and preindustrial sediments by comparative shotgun metagenomic analysis of surface water and sediment samples from three eastern Canadian lakes. In a strategy that was aimed at disentangling historical DNA from the indigenous sediment background, microbial preservation signals were captured by mapping sequence similarities between sediment metagenome reads and reference surface water metagenome assemblies. We detected preserved *Cyanobacteria*, diverse bacterioplankton, microeukaryotes, and viruses in sediment metagenomes. Among the preserved microorganisms were important groups never before reported in paleolimnological reconstructions, including bacteriophages (*Caudovirales*) and ubiquitous freshwater *Betaproteobacteria* (*Polynucleobacter* and *Limnohabitans*). In contrast, ultramicroscopic *Actinobacteria* (“*Candidatus* Nanopelagicales”) and *Alphaproteobacteria* (*Pelagibacterales*) were apparently not well preserved in sediment metagenomes even though they were numerically dominant in surface water metagenomes. Overall, our study explored a novel application of whole-metagenome shotgun sequencing for discovering the DNA remains of a broad diversity of microorganisms preserved in lake sediments. The recovery of diverse microbial time series supports the taxonomic expansion of microbiome reconstructions and the development of novel microbial paleoindicators.

4.2 Importance

Lakes are critical freshwater resources under mounting pressure from climate change and other anthropogenic stressors. The reconstruction of ecological time series from sediment archives with

paleolimnological techniques has been shown to be an effective means of understanding how humans are modifying lake ecosystems over extended timescales. In this study, we combined shotgun DNA sequencing with a novel comparative analysis of surface water and sediment metagenomes to expose the diversity of microorganisms preserved in lake sediments. The detection of DNA from a broad diversity of preserved microbes serves to more fully reconstruct historical microbiomes and describe preimpact lake conditions.

4.3 Introduction

Lakes support human wellbeing and a rich suite of ecological functions. Relative to the surface area they occupy worldwide, lakes contribute disproportionately to global elemental cycles (Tranvik *et al.*, 2009; Anderson *et al.*, 2020). Lakes are also natural repositories of long-term environmental data. Internal lake dynamics and changes driven by terrestrial and atmospheric inputs are integrated and archived through sediment deposition (Williamson *et al.*, 2008; Adrian *et al.*, 2009). As climate change and other anthropogenic stressors exert increasing pressures on lakes (Reid *et al.*, 2019), it is of great interest to examine how lakes are being modified from their preindustrial states. In the absence of long-term monitoring, chronologies reconstructed from lake sediment records can describe departure from preimpact baselines (Dubois *et al.*, 2018) and provide insights into the myriad effects of the anthropogenic impact observed since circa 1880 or earlier (Perga *et al.*, 2015; Giosan *et al.*, 2018; Smol, 2019). Classical microbial paleoindicators (e.g., diatom and chrysophyte assemblages) identified from subfossils, pigments, and other proxies are widely applied in paleolimnology for inferring key aspects of historical water columns (Smol, 1988; Dixit *et al.*, 1992).

Molecular genetic techniques have revolutionized environmental microbiology by exposing a vast, uncultured diversity of microorganisms (Handelsman, 2004). Molecular genetic analysis, advanced by PCR gene amplification and shotgun metagenomics, is continually expanding the body of knowledge on the diversity and importance of microbial communities in all habitable environments, including aquatic ecosystems (Grossart *et al.*, 2020). The development of molecular genetic methods in paleolimnology has enhanced the detection of soft-bodied microorganisms that lack preserved morphological features but that can be tracked in sediment DNA archives (Domaizon *et al.*, 2017; Bálint *et al.*, 2018). The inclusion of a broader diversity of microorganisms in paleolimnological reconstructions allows access to the microbiomes which underlaid lake ecosystems and biogeochemical processes in the past. The detection of preserved microbial diversity further offers opportunities to explore the still largely untapped potential of microbes as paleoindicators, grounded in the ubiquity of microorganisms in aquatic systems, their unparalleled phylogenetic and ecological breadth (Falkowski, Fenchel & Delong, 2008), and their rapid responses to change (Glasl, Webster & Bourne, 2017). Microbial time series reconstructed from lake sediment DNA have so far been used to investigate the emergence of toxic and bloom-forming *Cyanobacteria* (Martínez de la Escalera *et al.*, 2014; Monchamp *et al.*, 2016; Monchamp, Spaak & Pomati, 2019; Pilon *et al.*, 2019), microeukaryotic community succession in response to eutrophication and climate change (Capo *et al.*, 2016, 2017a), microbial evolution in response to anthropogenic mercury loading (Poulain *et al.*, 2015), and

past methane oxidation dynamics (Belle & Parent, 2019). The PCR-directed methods dominating molecular genetic analysis in paleolimnology are useful for reporting partial taxonomic and functional components of historical microbiomes. However, reconstructing historical diversity through PCR amplicon sequencing is limited by taxonomic blind-spots (Eloe-Fadrosh *et al.*, 2016), restricted gene quantification value (Piwosz *et al.*, 2020), and dependency on intact DNA template (Taberlet *et al.*, 2012). Furthermore, many major microbial groups have never been targeted in standard single-gene surveys, leaving unexplored any new insights they might yield on past lake ecosystems.

Metagenomics has the potential to provide a powerful new perspective to probe the unconstrained diversity preserved in sediment archives. Shotgun sequencing has gained traction in other ancient DNA disciplines (e.g., studies of ancient hominins in evolutionary anthropology) to recover ultrashort DNA fragments not amplifiable by PCR and to avoid other PCR biases (Ziesemer *et al.*, 2015). To date, only a handful of paleolimnological studies have applied shotgun sequencing. Pedersen *et al.* reconstructed the postglacial succession of plants, algae, and animals from lake sediment metagenomes (Pedersen *et al.*, 2016). Subsequent research demonstrated that shotgun sequencing expands the richness of catchment flora detected in lake sediments (Parducci *et al.*, 2019) and that ancient chloroplast and mitochondrial genomes can be resolved from the DNA remains of vegetation (Schulte *et al.*, 2021) and algae (Lammers, Heintzman & Alsos, 2021). Aside from the studies on eukaryotic algae, the only other paleolimnological reconstruction to describe autochthonous microbiomes with shotgun metagenomics profiled the succession of archaea in an ancient lake (Ahmed *et al.*, 2018). In these studies, historical DNA represented tiny fractions of sediment metagenomes but still furnished new information (i.e., not overlapping with classical proxies) about past lacustrine or catchment biota. As these works explore the potential of metagenomics, they underscore the primary challenges associated with reconstructing historical records from sediment metagenomes. Sediment DNA pools derive foremost from indigenous sediment microbes, and historical DNA is subject to high turnover (Torti, Jørgensen & Lever, 2018). Historical signals are difficult to detect because they are weaker than, and are distorted by, indigenous sediment populations.

In this study, whole-metagenome shotgun sequencing was used to probe the diversity of microorganisms preserved in contemporary and preindustrial sediment archives to expand the taxonomic scope of lake microbiome reconstructions. To identify historical DNA in sediment metagenomes, microbial preservation signals were captured by mapping sediment metagenome reads to surface water metagenome assemblies. This fragment recruitment strategy was used to disentangle historical DNA from the indigenous sediment background, using the contemporary diversity in the overlying water column as a reference. The detection of preserved DNA archives was subsequently expanded in the unconstrained explorations of whole-sediment metagenomes, guided by leads developed through metagenome capture. By teasing out the DNA preservation signals of historical lake microbiomes in the sediment record, we are able to harness a more complete perspective on the diversity of lakes and their ecological functioning over decades to millennia.

4.4 Results

4.4.1 Environmental context

Whole-metagenome shotgun sequencing was performed on surface water and sediment DNA sampled from three regionally representative eastern Canadian lakes: Lac Paula, Eightmile Lake, and Grand lac Touradi (Figure 4.1; Table 4.1). Lac Paula is a medium-sized, deep lake set in a small, equal parts clear-cut and naturally forested watershed in central Quebec in the Boreal Shield ecozone. Eightmile Lake is a small lake of intermediate depth set in a small, ~38% clear-cut watershed in northern New Brunswick in the Atlantic Highlands ecozone. Grand lac Touradi is a large, deep lake in an expansive, naturally forested watershed in eastern Quebec in the Atlantic Highlands ecozone. Top sediments were sampled from the recently deposited, upper 1 cm of each core, and bottom sediments were sampled from preindustrial deposits (on average, 40 cm below the water-sediment interface). A conservative lead ^{210}Pb (lead-210) dating model placed all bottom sediment ages in the preindustrial background (i.e., pre-1880 Common Era [C.E.]) (see Fig. S1 in the supplemental material).

4.4.2 Preserved metagenome capture

We define captured metagenomes as products of *in silico* read recruitment and free metagenomes as whole metagenomes unconstrained by capture. Microbial DNA preserved in sediment metagenomes was captured by mapping top and bottom sediment metagenome reads to surface water metagenome assemblies (Figure 4.2a). Some low-similarity alignment was expected to arise from distant evolutionary relationships between freshwater microorganisms and indigenous sediment populations. To disentangle historical and indigenous sediment microbiome DNA, capture of closely related sequences was performed with a stringent 90% identity threshold, empirically derived from the consistent delineation of closely ($\geq 90\%$) and distantly ($< 90\%$) related populations (Figure 4.2b). Most scaffolds in captured metagenomes did not contain highly conserved ribosomal or transfer RNA genes ($97.9\% \pm 1.0\%$) (see Table S1 in the supplemental material). Still, to avoid overinflating the estimates of DNA preservation from false-positive detections of distantly related organisms aligned at conserved sequences, scaffolds containing ribosomal or transfer RNA genes were removed (Table S1). Mapping between sediment and surface water metagenomes was detected across all lakes (0.2 to 0.8%), but it was lower than mapping between bottom and top sediment metagenomes (5.5 to 21.0%) (Table 4.2). On average, $98.9\% \pm 0.8\%$ of scaffolds in captured metagenomes contained at least one annotated gene, and the majority of scaffolds across captured metagenomes contained a single gene ($56.9\% \pm 11.8\%$) (Fig. S2a).

4.4.3 Taxonomic diversity in free and captured metagenomes

We performed taxonomic assessments of free and captured metagenomes to compare the compositions of natural surface water and sediment microbiomes with the microbial diversity preserved in sediment DNA archives (Figure 4.3; Table S2). Free surface water metagenomes in all lakes contained high relative coverage of *Actinobacteria*, *Betaproteobacteria*, *Alphaproteobacteria*, and *Bacteroidetes*. Free top sediment

metagenomes also contained high relative coverage of *Actinobacteria* and *Betaproteobacteria* but were also rich in *Deltaproteobacteria*, *Euryarchaeota*, and *Gammaproteobacteria*. Free bottom sediment metagenomes contained high relative coverage of *Euryarchaeota*, *Deltaproteobacteria*, *Firmicutes*, and *Chloroflexi*.

Analyses of the captured metagenomes considered at the phylum level demonstrated a relatively consistent assemblage of preserved microbiota, with a few differences between contemporary and preindustrial sediments (Figure 4.3; Table S2). Comparisons of the captured surface water signals in either top or bottom sediment metagenomes contained high relative coverage of viruses, *Actinobacteria*, *Betaproteobacteria*, *Cyanobacteria*, and *Alphaproteobacteria*. Captured surface water signals in bottom sediment metagenomes also contained high relative coverage of *Planctomycetes*. Captured top sediment signals in bottom sediment metagenomes contained high relative coverage of *Euryarchaeota* and *Deltaproteobacteria*.

To elucidate the taxonomic diversity of free and captured metagenomes at higher phylogenetic resolution, taxonomic analyses were performed at the order level (Figure 4.4a; Table S3). A principal-component analysis (PCA) of these data demonstrated a clear distinction between free and captured metagenomes along the first principal component (PC) axis, reflecting the taxonomic distinctness of surface water and sediment microbiomes (Figure 4.4b). Free and captured surface water metagenomes were separated along the second PC axis, suggesting that only a subset of microbial diversity is preserved in sediments (Figure 4.4b). The first two PC dimensions together explained 64.8% of the variation between metagenome taxonomic compositions. “*Candidatus* Nanopelagicales” and *Pelagibacterales* were highly covered in the free surface water metagenomes of all lakes, but they were not highly covered in captured surface water metagenomes (Figure 4.4a; Table S3). Free and captured sediment metagenomes contained high relative coverage of *Methanomicrobiales*, *Syntrophobacterales*, *Desulfobacterales*, *Clostridiales*, *Desulfuromonadales*, *Myxococcales*, *Nitrosomonadales*, *Rhizobiales*, and *Nitrospirales* (Figure 4.4a; Table S3).

4.4.4 Highlights of preservation signals in captured metagenomes

Since most freshwater *Cyanobacteria* are photosynthetic, *Cyanobacteria* DNA detected in the metagenomes of subsurface sediments can be distinguished from the indigenous sediment background and identified as the preserved DNA of *Cyanobacteria* that inhabited historical sunlit lake biomes. Captured surface water metagenomes exhibited higher relative coverage of *Cyanobacteria* than free surface water metagenomes, highlighting recruitment capture as a viable strategy for detecting well-preserved historical microorganisms (Figure 4.3; Table S2). Of the three lakes, Grand lac Touradi displayed the most sizable relative capture of preserved *Cyanobacteria* in contemporary and preindustrial sediments (Figure 4.3; Table S2). *Synechococcales* were the most highly represented *Cyanobacteria* in captured surface water metagenomes (Figure 4.4a; Table S3).

Captured surface water metagenomes also showed that sediments accumulate DNA from some – but not all – bacterioplankton, as well as viruses. The most highly covered bacterioplankton that were represented in captured surface water metagenomes were *Corynebacteriales*, *Burkholderiales*, *Planctomycetales*, *Sphingomonadales*, and *Rhizobiales* (Figure 4.4a; Table S3). *Burkholderiales* were highly covered in both free and captured surface water metagenomes and in free top sediment metagenomes, suggesting that some planktonic *Burkholderiales* are preserved in sediments, while other *Burkholderiales* are indigenous to sediments and may potentially confound historical signals without higher-resolution taxonomic assessments (Figure 4.4a; Table S3). Some highly abundant bacterioplankton (*Ca. Nanopelagicales*, *Pelagibacteriales*) were only marginally covered in captured surface water metagenomes (Figure 4.4a; Table S3). Viruses were highly covered in the captured surface water metagenomes of all lakes, but they were not highly covered in the free surface water metagenomes (Figure 4.3; Table S2). *Caudovirales* were the most highly captured viruses, especially in Eightmile Lake (Figure 4.4a; Table S3). A phylogeny inferred from the terminase large subunit protein (TerL) identified a wide diversity of *Caudovirales*, but host identities could not be resolved (Fig. S3).

4.4.5 Exploring unconstrained preservation signals in free metagenomes

While metagenome capture was practical for developing leads on preserved taxa, experimental limitations rendered this approach unsuitable for recruiting unconstrained preservation signals for historical lake microbiomes. Specifically, the seasonal, depth-discrete, and contemporary community structures of surface water snapshot samples constrained the capture of a broad microbial diversity. To expand the detection of preserved DNA to include the diversity that escaped capture, the search for preserved taxa was extended to the unconstrained DNA libraries of free sediment metagenomes. Across all lakes, free top and bottom sediment metagenomes contained traceable accumulations of genes from *Cyanobacteria*, bacterioplankton, and microeukaryotes (Figure 4.5). Genes assigned to *Cyanobacteria* (*Synechococcus*, *Aphanothece*, and *Cyanobium*) numbered in the thousands in free surface water metagenomes and were detected to a lesser extent in top and bottom sediment metagenomes (Figure 4.5). Sediment metagenomes contained genes from other bacterioplankton, including *Betaproteobacteria* (*Rhodoferrax*, *Polynucleobacter*, and *Limnohabitans*), *Alphaproteobacteria* (*Sphingomonas*), *Actinobacteria* (*Mycobacterium*, *Mycolicibacterium*, *Ilumatobacter*, and *Rhodoluna*) (Figure 4.5), and others (Fig. S4). A small number of genes from eukaryotic plankton, namely, diatoms (*Phaeodactylum* and *Thalassiosira*), were also detected in sediment metagenomes (Figure 4.5). Some bacterioplankton (“*Candidatus* Methylopusillus,” “*Candidatus* Fonsibacter,” “*Candidatus* Planktophila,” and “*Candidatus* Nanopelagicus”) that were highly covered in surface water metagenomes were represented by relatively low gene counts in free sediment metagenomes (Figure 4.5).

To demonstrate how preserved microbial diversity might be identified at high phylogenomic resolution, surface water and sediment metagenome reads were mapped to publicly available reference genomes from *Polynucleobacter* and *Limnohabitans*, two genera of *Betaproteobacteria* with well-described

freshwater microdiversity and that were common in captured metagenomes. Surface water metagenomes from Eightmile Lake and Grand lac Touradi contained *Polynucleobacter* and *Limnohabitans* populations with high sequence similarity ($\geq 90\%$) to reference genomes (Fig. S5). In contrast, reference genomes were not captured in any of the top or bottom sediment metagenomes (Fig. S5).

4.5 Discussion

This study illustrates that shotgun sequencing can access a broad diversity of microorganisms preserved in contemporary and preindustrial lake sediments. Preserved taxa were identified by mapping sediment metagenome reads to reference surface water metagenome assemblies with a stringent 90% sequence identity threshold derived to disentangle freshwater signals from the indigenous sediment background. Once well-preserved lineages were identified through the *in silico* capture of surface water signals in sediment metagenomes, unconstrained explorations for preserved taxa were conducted in free sediment metagenomes to uncover preserved diversity that escaped capture. This strategy recovered a wide diversity of freshwater microorganisms previously overlooked by more classical paleolimnological approaches. Given that microorganisms have critical functions in lake ecosystems, including carbon and other elemental cycling, studying the dynamics of lake microbiomes in the past could be quite insightful.

4.5.1 Preserved taxa in sediment metagenomes

Taxonomic analysis of contemporary and preindustrial sediment metagenomes showed that these metagenomes contained hundreds to thousands of preserved genes from freshwater microbiota within the *Cyanobacteria* (*Synechococcus*, *Aphanothece*, and *Cyanobium*), *Betaproteobacteria* (*Rhodoferrax*, *Polynucleobacter*, and *Limnohabitans*), *Alphaproteobacteria* (*Sphingomonas*), *Actinobacteria* (*Mycobacterium*, *Mycolicibacterium*, *Ilumatobacter*, and *Rhodoluna*), microeukaryotes (*Phaeodactylum* and *Thalassiosira*), and viruses (*Caudovirales*).

Betaproteobacteria were abundant in surface waters and well preserved in top and bottom sediment metagenomes. The ubiquity of *Burkholderiales* in water column and sediment microbiomes illustrates the importance of applying a stringent sequence identity cutoff to capture the elements of sediment metagenomes closely related to freshwater taxa as an approach to identifying historical DNA. *Burkholderiales* exhibit wide physiological and ecological diversity and have been isolated from a range of different sources, including soils and freshwater lakes (Coenye, 2014; Willems, 2014). The sheer diversity of *Burkholderiales* makes it difficult to pin down specific traits which can explain their preservation in sediment metagenomes.

In general, taxa were not equally preserved in sediment metagenomes. *Actinobacteria* and *Alphaproteobacteria* were numerically important in surface water metagenomes, reflecting distribution patterns typical of freshwater lakes (Newton *et al.*, 2011; Salcher, Pernthaler & Posch, 2011). Despite dominating surface water microbiomes, *Ca. Nanopelagicales* (*Ca. Nanopelagicus* and *Ca. Planktophila*), *Pelagibacterales* (*Ca. Fonsibacter*), and other ubiquitous bacterioplankton (*Ca. Methylopusillus*) were not

well preserved in top or bottom sediments. *Ca. Nanopelagicales*, *Pelagibacterales*, and *Ca. Methylopusillus* are ultramicrobacteria characterized by free-living lifestyles and streamlined genomes with low guanine-cytosine (GC) contents (Henson *et al.*, 2018; Neuenschwander *et al.*, 2018; Salcher *et al.*, 2019), traits held in common with the well-preserved ultramicroscopic *Burkholderiales*. A variety of mechanisms has been put forth to explain the differential DNA preservation across microbial groups previously observed in single-gene time series, including sporulation (Boere *et al.*, 2011), cellular architecture and membrane biochemistry (Capo *et al.*, 2015), genome GC content (Vuillemin *et al.*, 2017), and autoecology. However, further work is needed to elucidate how DNA is differentially preserved across taxa and under different lake conditions.

The robust preservation of viruses in the captured metagenomes of all three lakes suggests that historical viromes are at least partially accessible to reconstruction from sediment archives. The best preserved viruses were *Caudovirales*, tailed bacteriophages infecting a wide range of hosts through both lytic and lysogenic replication (Maniloff & Ackermann, 1998). A phylogeny based on TerL appeared to resolve the *Caudovirales* as phages associated with several bacterial taxa; however, the tree lacked both the strong bootstrap confidence and clear delineation of clades necessary for definitive host identifications. The strong preservation of bacteriophage DNA is likely related to two mechanisms. First, prophages are widely integrated in the genomes of infected microorganisms (Canchaya *et al.*, 2003), whose own DNA is variably preserved in sediments. Paleoeological time series can exploit the concurrent preservation of viral and host DNA to reconstruct virus-host evolution and illuminate the influence of environmental change on virus-host dynamics (Coolen, 2011). The second mechanism of preservation is the hardiness of encapsidated virions, which are suited to withstand adverse conditions and remain intact under long-term cold storage (Ng *et al.*, 2014). As abundant biological entities that exert enormous control over microbial populations in aquatic ecosystems (Fuhrman, 1999; Kavagutti *et al.*, 2019), viruses could lend valuable insights into historical lake microbiomes.

A major issue in ancient DNA disciplines is that historical signals can be distorted or diluted by cross-contamination from contemporary DNA (Domaizon *et al.*, 2017). The exposure of sediment DNA to modern contaminants can potentially occur at any stage of sediment core sampling or laboratory work. The first precaution we took to minimize cross-contamination was to retrieve the innermost sediments furthest from the core tube walls at top and bottom intervals using sterile spatulas and syringes. Another precaution was to perform DNA extractions from water filtrates and sediments in separate dedicated laboratories under positive-pressure UV hoods. Precautions aside, the historical microbial signals detected in sediment metagenomes appeared valid and not contaminated by contemporary DNA, as evidenced by the observed trends in DNA preservation across groups. If sediment metagenomes were contaminated by surface water DNA, we would expect the preferential recruitment of sediment metagenome reads to DNA from common and abundant lacustrine microorganisms (e.g., ultramicroscopic *Actinobacteria* and *Alphaproteobacteria*). Instead, preferential recruitment of sediment metagenome reads to rarer surface water taxa was observed.

4.5.2 Harnessing microbial ecotypes as paleoindicators

The detection of preserved DNA from a broad diversity of microorganisms expands the taxonomic scope of paleolimnology. To access the paleoindicator potential of preserved microbiota, we suggest that preimpact lake conditions can be reconstructed from records of past microbial ecologies. In the future, paleolimnological reconstructions focused on well-preserved heterotrophic bacteria could provide key insights into past ecosystem functions. The ecologies of ubiquitous lacustrine bacterioplankton are extensively detailed in published literature (Newton *et al.*, 2011; Pernthaler, 2017). For some freshwater bacteria, ecotypic diversity has been described, i.e., genetically differentiated lineages with specific ecological requirements (Rocap *et al.*, 2003; Cohan, 2006). We propose that the ecologies of well-studied bacterioplankton, particularly those with resolved ecotypes, can clarify the niche dimensions of historical microbiomes.

Our explorations of captured and free metagenomes detected key bacterioplankton groups for which ecotypic diversity is well described, e.g., *Polynucleobacter* and *Limnohabitans*. The microdiversity of ubiquitous *Polynucleobacter* and *Limnohabitans* has been investigated in a wide variety of freshwater lakes and has been physiologically characterized through culture, transplant experiments, and genome sequencing (Hahn *et al.*, 2016a b; Nuy *et al.*, 2020; Props & Denef, 2020). The niche specificities of *Polynucleobacter* or *Limnohabitans* ecotypes in DNA time series could potentially describe historical water column conditions (e.g., pH, carbon and other nutrient substrates, light penetration, temperature, and oxygen levels). Historical ecotypes might be identified by mapping sediment metagenomes to ecotype reference genomes, or a reduced approach might focus on recovering marker genes with sufficient phylogenetic information to resolve microdiversity. Toward the identification of ecotypes, we attempted to map sediment metagenomes to publicly available *Polynucleobacter* and *Limnohabitans* reference genomes. While reference genomes were captured in the surface water metagenomes of Eightmile Lake and Grand lac Touradi, they were not captured in the sediment metagenomes. The low-similarity alignment of sediment metagenomes and reference genomes representing genera detected in sediment metagenomes suggests that these contemporary reference genomes do not reflect historical microdiversity. Ecotypes could potentially be resolved if sets of genomes or metagenome-assembled genomes representative of local to regional pre- and postimpact diversity were made available, especially through largescale metagenomic lake surveys (Huot *et al.*, 2019). Comparing the ecological dependencies reconstructed in ecotype time series could then retrace environmental change.

4.5.3 Preserved metagenome capture

Metagenome capture was designed to frame a novel methodology and proof of concept for the detection of historical DNA with shotgun metagenomics. However, with the approach we adopted in this pilot study, we are limited in our abilities to represent the totality of preserved microbial diversity. First, our surface water metagenomes were snapshot samples of DNA collected from the photic zone during summer stratification. These reference metagenomes comprised microbial assemblages partially representative of

the vertical structure of stratified lakes, which varies seasonally. Second, surface water metagenomes represented contemporary diversity. Some degree of community turnover is expected across the more than 150 years separating preindustrial and contemporary microbiomes, and preindustrial diversity no longer occurring in contemporary microbiomes cannot be recruited. Finally, metagenome assembly only partly incorporated unassembled metagenomes, with the effect that reference assemblies further constricted the scope of sampled diversity. These limitations parametrizing the capture of snapshot surface water metagenomes effectively set up the capture of preservation signals to fail. Yet even so, we recovered partial preserved metagenomes from sediments with novel taxonomic information.

Metagenome capture was most prolific between sediment metagenomes, which displayed 20- to 30-fold-higher sequence mapping than captured surface water metagenomes. The continuity between top and bottom sediment metagenomes was observed in each lake, despite a depositional age difference of over 150 years and commonly observed steep geochemical gradients along sediment horizons (Melton *et al.*, 2014). In contrast, the high taxonomic variation between surface water and sediment metagenomes can be explained by the stark environmental dissimilarities between photic zone and sediment subsurface habitats (Walsh *et al.*, 2016) and by the low composition of sediment DNA that is of historical origin (35). Experimental capture showed that very small fractions of sediment metagenomes originated from historical freshwater microorganisms (<1%, under the limitations of recruitment to snapshot reference surface water metagenomes). The disparity between the proportions of indigenous and historical DNA in sediments underscores the need to develop strategies to optimize historical DNA recovery while cancelling out signals from background sediment populations. One obvious path forward would be to apply the capture approach on a combined assembly of reference water column metagenomes, collected over a larger set of lakes or across multiple seasons. Nonetheless, the historical DNA detected by metagenome capture was sufficient to partially reconstruct historical lake microbiomes. PCR gene amplification is likely the most practical approach for targeting the population dynamics of preserved groups identified in this study. Future works might also attempt to extract historical signals embedded in indigenous sediment microbiomes (Orsi *et al.*, 2017; More *et al.*, 2019).

4.5.4 Conclusion and outlook

In conclusion, we put forth two ideas to meet some of the unrealized potential of molecular genetic analysis in paleolimnology. First, whole-metagenome shotgun sequencing can access a broad diversity of microorganisms preserved in lake sediment DNA archives. Second, microorganisms are underexploited paleoindicators whose ecotypic diversity can potentially retrace environmental flux. This study tested a proof of concept that DNA from historical lake microbiomes can be recovered from sediment metagenomes. Our strategy of baiting unassembled sediment metagenome reads with reference surface water metagenome assemblies was designed to disentangle historical microbial signals from the indigenous sediment background. Future works are encouraged to improve on this approach to optimize the recovery of historical signals, e.g., by querying population-level gene variation in microbial time series with PCR

amplification. All told, our application of shotgun DNA sequencing has expanded the detectable microbial diversity available to paleolimnological reconstructions to include heterotrophic bacteria and viruses, which we posit are as important in historical contexts as they are in contemporary lake microbiomes.

4.6 Materials and methods

4.6.1 Lake water and sediment sampling

Surface water and sediment cores were sampled from three eastern Canadian lakes between 31 July and 25 August 2017 during the Canadian Lake Pulse Network field campaign (Huot *et al.*, 2019) (Figure 4.1; Table 4.1). Lakes were selected by meeting the criteria of a maximum depth of at least 1 m and accessibility within 1 km from a road (Huot *et al.*, 2019). At each lake, surface water and sediment cores were collected at the site of maximum depth, located by depth sounding and using bathymetric maps as guides when available. All water sampling devices were acid washed and triple rinsed before sample processing. Water was collected from the euphotic zone over an integrated depth of up to 2 m below the surface with an integrated tube sampler. Carboys were stored in chilled coolers until same-day water filtration on shore. Up to 500 ml of water was prefiltered through 100- μ m synthetic nylon (Nitex) mesh and then vacuum filtered on 47-mm-diameter 0.22- μ m Durapore membranes (MilliporeSigma, Darmstadt, Germany) through a glass funnel apparatus at a maximum pressure of 8 in Hg. Sediment core target length was estimated to reach preindustrial depositional age based on regional sedimentation rates synthesized from the literature. Sediment cores were retrieved using a gravity corer fitted with a 68-mm-diameter plastic core tube. Sediment cores were extruded on shore and subsampled at top and bottom 1-cm intervals. Bottom sediment subsections collected at core depth below the water-sediment interface were relatively consistent: the 40- to 41-cm interval in Lac Paula and the 39- to 40-cm interval in Grand lac Touradi and Eightmile Lake. Samples for molecular genetic analysis were collected from innermost sediments, furthest from the walls of the core tube, using a sterile syringe or spoon. Water filters and sediment samples were stored at -80 and -20°C, respectively. To determine the depositional ages of bottom sediments, ^{210}Pb and bismuth-214 (^{214}Bi) radionuclide activities were measured using gamma spectrometry and a constant rate of supply model (Schelske *et al.*, 1994). When overlap was observed between ^{210}Pb and ^{214}Bi activities within 1 standard deviation, sediments were determined to have been deposited within the last 100 to 150 years (i.e., reaching preindustrial background).

4.6.2 DNA extraction

DNA was extracted from filters using the DNeasy PowerWater kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions with the addition of the following optional steps: after vortex bead beating and centrifugation (step 7 of manufacturer's detailed protocol), 1 μ l of RNase A was added to samples before a 30-min incubation at 37°C. DNA extractions from sediments were performed in a UV cabinet in a separate dedicated laboratory. DNA was extracted from 0.500 ± 0.015 g of wet sediment using the NucleoSpin soil kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's instructions

with lysis buffer SL1 and enhancer SX options, without protocols for separating intra- and extracellular DNA. DNA was quantified using the Qubit dsDNA BR assay kit (Invitrogen, Carlsbad, CA, USA) in a Qubit 2.0 fluorometer.

4.6.3 Metagenome shotgun sequencing, assembly, and annotation

DNA samples were submitted to Genome Quebec for library barcoding and shotgun sequencing of 150-bp paired-end reads on an Illumina NovaSeq 6000 platform with flow cell type S2. The sequencing depth ranged from 25,717,497 to 73,554,802 reads per sample. Sequencing adapters, low-quality bases, and ultrashort reads were trimmed with Trimmomatic v. 0.38 (Bolger *et al.*, 2014). Metagenome assembly from paired reads was performed with MEGAHIT v. 1.0.6 using k-mer lengths 23, 43, 63, 83, 103, and 123 (Li *et al.*, 2016). Assemblies were submitted to the Integrated Microbial Genomes & Microbiomes (IMG) system for annotation in IMG Annotation Pipeline v. 4.16.6 (Huntemann *et al.*, 2016). Taxonomy was predicted by phylodist (i.e., the taxonomic affiliation of the closest gene homolog) in the IMG pipeline.

4.6.4 Read recruitment

For each lake, read recruitment was conducted between surface water, top sediment, and bottom sediment metagenomes, always in the direction tracking the downward preservation of DNA (i.e., reads from the underlying metagenome were mapped to the overlying metagenome assembly; Figure 4.2a). Global read recruitment was performed with BMap v. 35.68 (Bushnell, 2015) using an initial, highly permissive nucleotide sequence identity threshold of ~70% to assess alignment trends between metagenomes. A subsequent, empirically derived sequence identity threshold of 90% was applied to align closely related sequences. This cutoff was selected after observing consistent shifts in recruitment trend across capture experiments toward 90% sequence identity, which appeared to delineate closely ($\geq 90\%$) and distantly ($< 90\%$) related populations (Figure 4.2b). To reject false-positive results overestimating DNA preservation, scaffolds containing ribosomal or transfer RNA genes were removed from captured metagenomes because these conserved sequences are shared with relatively high similarity among distantly related organisms and can produce biased alignment. The number of mapped reads was normalized by the number of unassembled reads in the recruitment effort. Scaffold lengths and gene counts on scaffolds in captured metagenomes were evaluated to determine whether the taxonomic annotations of protein-coding sequences could be extended to taxonomic assignments of whole scaffolds (see Fig. S2 in the supplemental material).

To increase the taxonomic resolution for groups with well-described freshwater microdiversity, surface water and sediment metagenome reads were recruited to *Polynucleobacter* and *Limnohabitans* reference genomes available from IMG (listed in Table S4 in the supplemental material). Ribosomal and tRNA genes were masked in reference genomes with BEDTools maskfasta v. 2.26.0 (Quinlan & Hall, 2010) to avoid read recruitment from distantly related organisms to conserved sequences. Read recruitment from surface water and sediment metagenomes to concatenated reference genomes was performed in BMap

with the default ~70% sequence identity threshold in order to visualize recruitment success as a function of genetic similarity.

4.6.5 Statistical analyses

Principal-component analysis (PCA) was performed on metagenome taxonomic composition data to evaluate variation among captured and free metagenomes from all study lakes. Relative coverage-weighted metagenome composition data of order-rank taxonomy were chord-transformed with the function `decostand(method = "norm")` in the R package `vegan` (Oksanen *et al.*, 2020). The PCA was computed with the `vegan` function `rda()`. Dendrograms of prevalent taxa were based on Ward's hierarchical clustering of the Bray-Curtis dissimilarities between metagenome taxonomic compositions across lakes. Data wrangling and visualization were programmed in R (R Core Team, 2021) using the tidyverse package suite (Wickham *et al.*, 2019).

4.6.6 Phylogenetic analyses

To resolve the host identities of captured bacteriophages, a *Caudovirales* phylogeny was constructed based on TerL, a common taxonomic marker for this group (Roux *et al.*, 2012). *Caudovirales* TerL amino acid sequences detected in metagenomes were aligned with publicly available reference sequences in MAFFT online service using default settings (Katoh *et al.*, 2019). A maximum likelihood tree was inferred using the Jones-Taylor-Thornton substitution model (Jones, Taylor & Thornton, 1992) with 5 discrete gamma categories and 100 bootstrap replicates in MEGA-X (Kumar *et al.*, 2018).

4.6.7 Data availability

The annotated metagenome assemblies in this study are available from the Integrated Microbial Genomes & Microbiomes system (U.S. Department of Energy Joint Genome Institute, Berkeley, CA, USA) at <https://img.jgi.doe.gov> under GOLD Study Gs0136026. Links to the GOLD Analysis Projects are listed in Table S5. All scripts are available at <https://github.com/rebeccagarner/paleocapture>.

4.7 Acknowledgements

We are grateful to the NSERC Canadian Lake Pulse Network's staff and field crews for their enormous efforts in collecting the lake water and sediment samples and for compiling the lake metadata. Thank you to Susanne Kraemer for performing the filter DNA extractions and for preparing the DNA samples for sequencing. Thank you to Adam Jeziorski for performing the sediment dating.

This work was supported by the Natural Sciences and Engineering Research Council of Canada and the Canada Research Chairs Program.

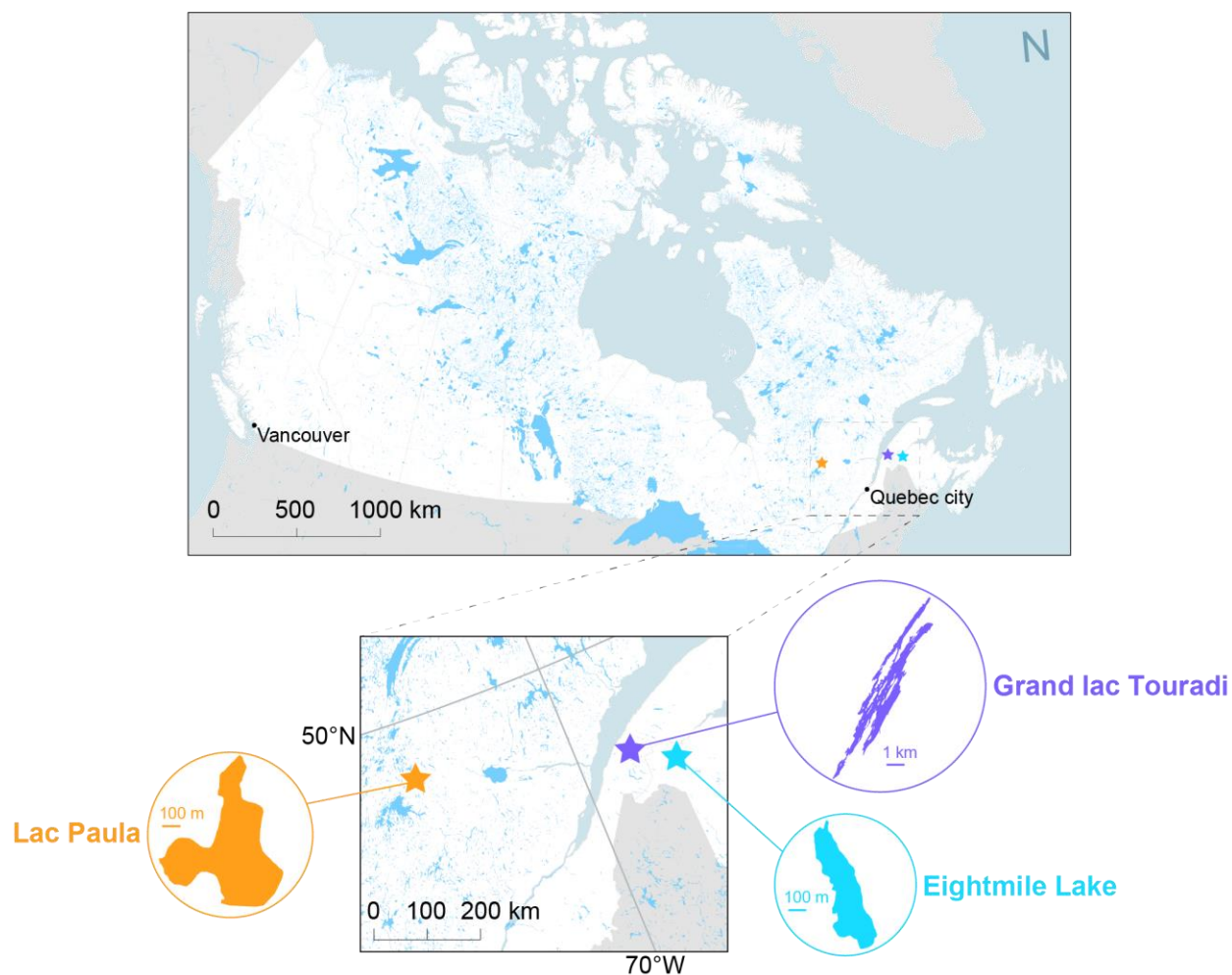


Figure 4.1 Map of Canada with the starred locations of study lakes: Lac Paula, Eightmile Lake, and Grand lac Touradi. The map was created in QGIS v. 3.10 (Team & Project, 2020) with the Canada Atlas Lambert projection and shapefile data sourced for the hydrographic features (Lakes, Rivers and Glaciers in Canada - CanVec Series - Hydrographic Features, 2019) and political boundaries of Canada (Boundary files, 2011 census). Spatial coordinates, morphological characteristics, and physicochemical profiles of study lakes are summarized in Table 4.1.

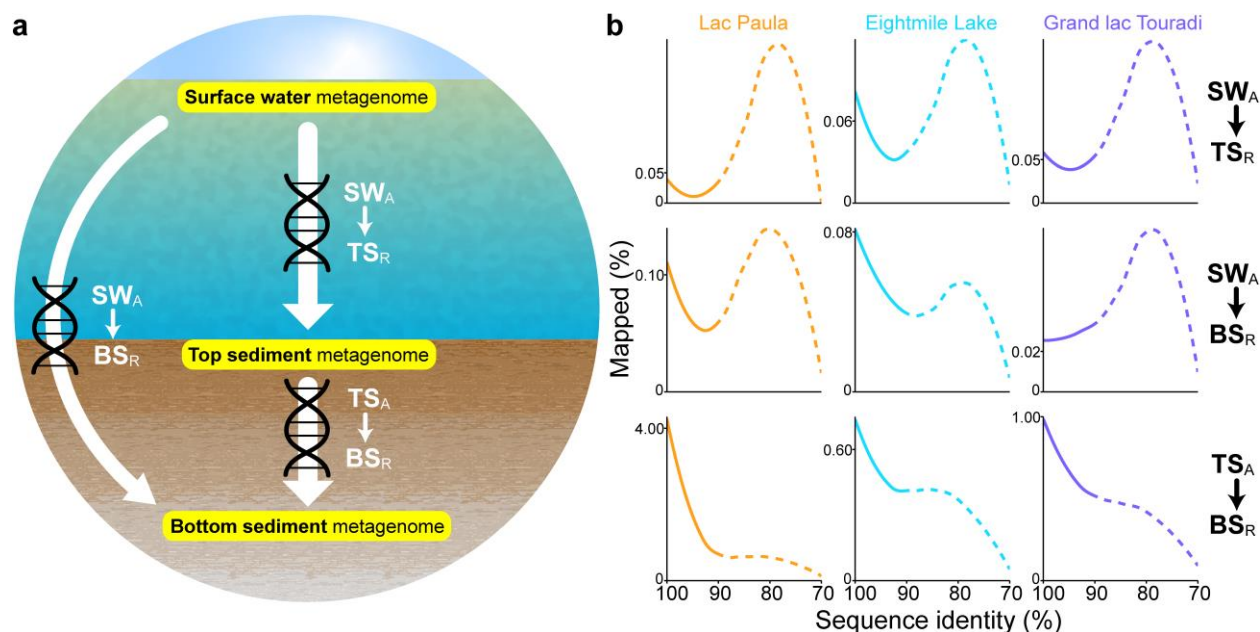


Figure 4.2 (a) As illustrated in this experimental design schematic, the recruitment of unassembled metagenome reads (abbreviated subscript R) to metagenome assemblies (subscript A) tracks the sequence similarities between focal and reference metagenomes. Captured metagenomes are abbreviated $SW_A \rightarrow TS_R$ (top sediment reads mapped to the surface water assembly), $SW_A \rightarrow BS_R$ (bottom sediment reads mapped to the surface water assembly), and $TS_A \rightarrow BS_R$ (bottom sediment reads mapped to the top sediment assembly). Arrows point in the direction of DNA preservation, from the overlying source to the underlying sediment archive. (b) Metagenome capture was performed with a 90% sequence identity threshold derived from the delineation of closely (solid curve, $\geq 90\%$) and distantly (dashed curve, $< 90\%$) related sequences. Reads mapped to scaffolds containing ribosomal or transfer RNA genes are included (exceptionally) in this figure, with negligible impact on the results presented (see Table S1 in the supplemental material).

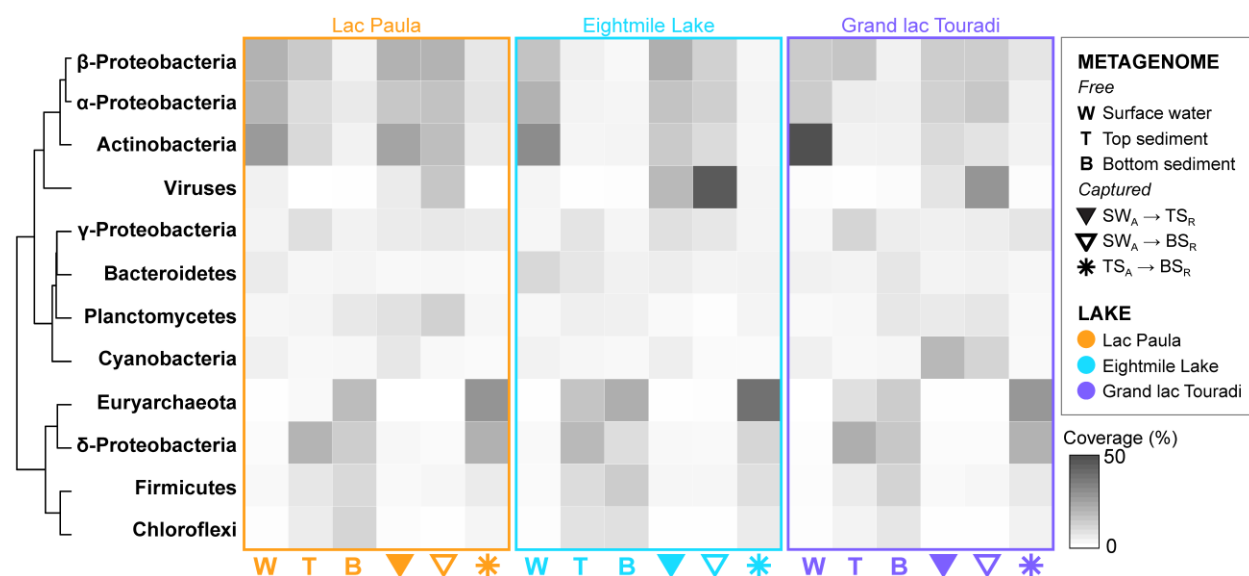


Figure 4.3 Relative coverage of prevalent phyla in the free and captured metagenomes of three lakes. Represented in this heat map are taxa with $\geq 10\%$ relative coverage in at least one metagenome.

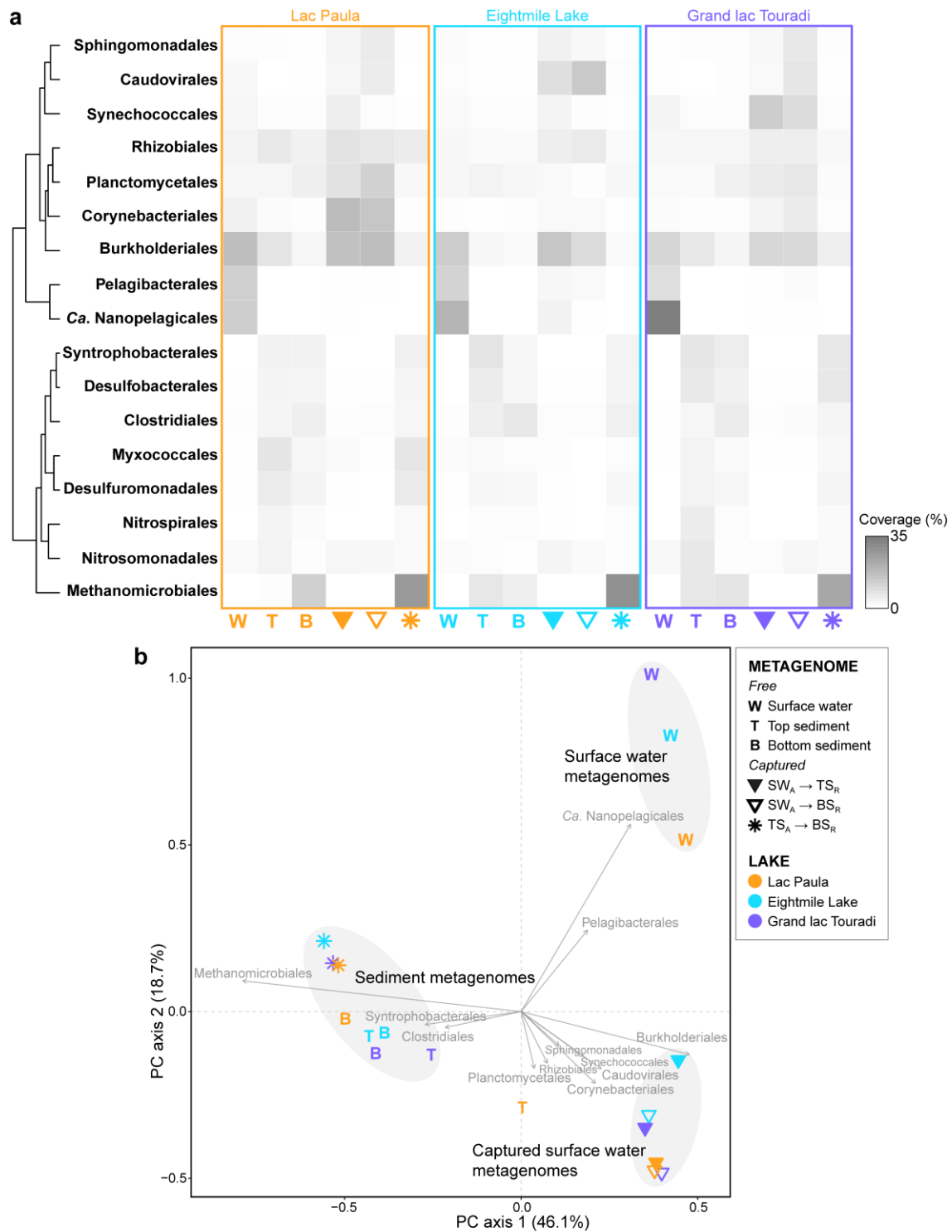


Figure 4.4 (a) Relative coverage of prevalent taxonomic orders in the free and captured metagenomes of three lakes. Represented in this heat map are taxa with $\geq 5\%$ relative coverage in at least one metagenome. (b) PCA of order-rank taxonomic composition data of free and captured metagenomes. The predominant taxon contributions to the first two PC axes are shown. Groups of taxonomically similar metagenomes are highlighted by grey-shaded ellipses.

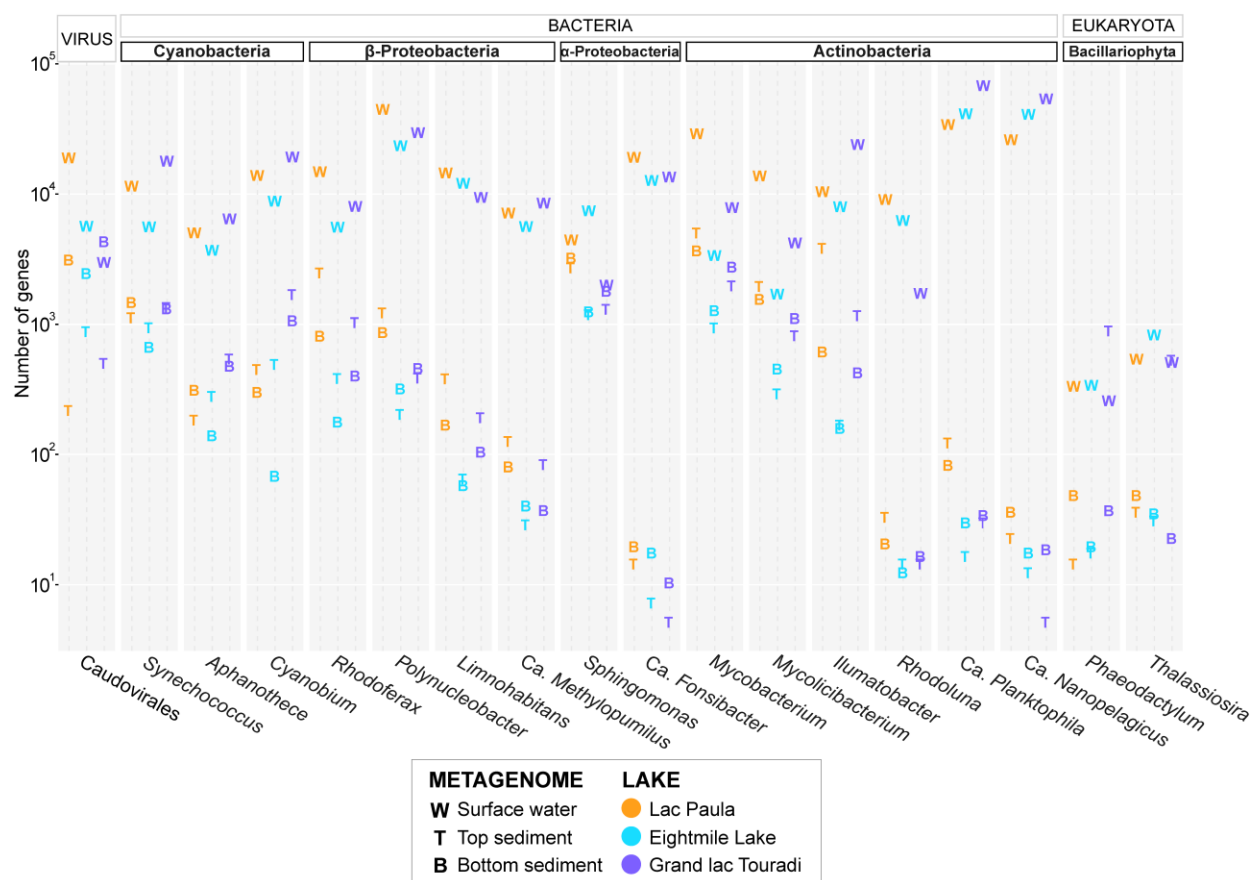


Figure 4.5 Gene counts of selected preserved *Cyanobacteria*, bacterioplankton, microeukaryotes, and viruses in free surface water and sediment metagenomes in three lakes.

Table 4.1 Spatial coordinates, morphological characteristics, and surface water physicochemical profiles of study lakes.

Lake	Latitude (°N)	Longitude (°W)	Maximum depth (m)	Surface area (km ²)	Watershed area (km ²)	Chlorophyll <i>a</i> concn (mg/L)	Dissolved inorganic carbon concn (mg/L)	Dissolved organic carbon concn (mg/L)	Soluble reactive phosphorus concn (µg/L)
Lac Paula	48.991254	74.028099	17.0	0.6	3.2	1.7	1.1	5.9	3
Eightmile Lake	47.694086	67.642841	4.5	0.2	1.7	3.5	14.6	6.5	2
Grand lac Touradi	48.131057	68.666779	17.0	7.7	117.7	2.6	15.6	6.5	3

Table 4.2 Counts and percentages (normalized by the number of unassembled reads) of reads mapped with a 90% sequence identity threshold in metagenome capture experiments.

Lake	Captured metagenome ^a	Unassembled read count	Mapped read count	Mapped (%)
Lac Paula	SW _A → TS _R	32,622,245	69,318	0.2
	SW _A → BS _R	67,840,581	511,831	0.8
	TS _A → BS _R	67,840,581	14,234,026	21.0
Eightmile Lake	SW _A → TS _R	29,873,977	145,692	0.5
	SW _A → BS _R	29,740,441	188,125	0.6
	TS _A → BS _R	29,740,441	1,622,267	5.5
Grand lac Touradi	SW _A → TS _R	29,290,940	143,854	0.5
	SW _A → BS _R	30,470,979	97,815	0.3
	TS _A → BS _R	30,470,979	2,324,829	7.6

^aCaptured metagenome abbreviations are defined in the legend to Figure 4.2a.

4.8 Supplemental material

Figure S1. Preindustrial (i.e., pre-1880) bottom sediment depositional ages were inferred when overlap was observed between bottom sediment lead-210 and bismuth-214 activities within 1 standard deviation.

Figure S2. (a) Frequencies of scaffolds containing n genes. (b) Average fold and lengths of scaffolds in captured metagenomes.

Figure S3. Terminase large subunit protein (TerL) phylogeny of *Caudovirales* detected in free and captured metagenomes.

Figure S4. Extension of Figure 4.5 to include all genus-rank taxa with $\geq 1\%$ coverage in free or captured surface water metagenomes.

Figure S5. Recruitment of surface water and sediment metagenome reads to concatenated *Polynucleobacter* and *Limnohabitans* reference genomes.

Table S1. Comparisons of the number of scaffolds and mapped reads in captured metagenomes before and after scaffolds containing ribosomal or transfer RNA genes were removed.

Table S2. Relative coverage of phyla in the free and captured metagenomes of three lakes.

Table S3. Relative coverage of prevalent taxonomic orders ($\geq 5\%$ relative coverage in at least one metagenome) in the free and captured metagenomes of three lakes.

Table S4. Accession information for *Betaproteobacteria* reference genomes downloaded from IMG.

Table S5. GOLD Analysis Project identifiers (IDs) and links for annotated metagenome assemblies in this study.

Chapter 5

Congruent paleogenetic and monitoring time series document the effects of eutrophication and climate on algal communities in experimental lakes

Garner R.E., Taranu Z.E., Paterson M.J., Gregory-Eaves I., Walsh D.A.

5.1 Abstract

Eutrophication and climate change are synergistically altering lake ecosystems. Yet, the often slow and insidious effects of each of these stressors are challenging to detect and disentangle within the typical span of lake monitoring, where such programs exist. Here, we couple the 50-year monitoring and 120-year sediment DNA records of experimentally fertilized and unmanipulated lakes in the IISD Experimental Lakes Area to investigate the effects of nutrients and warming on algal assemblages. Through RV coefficient analysis, we identified strong congruence between the temporal dynamics of monitoring and sediment DNA records, providing multi-lake validation of paleogenetics to reconstruct past ecosystem dynamics. In both monitoring and sediment records, nutrients were strong and significant drivers of algal turnover in experimentally fertilized lakes compared with unmanipulated references. However, climate change effects were most evident from the 120-year perspective opened by sediment records. The algal assemblages in continuously nutrient-enriched L227 experienced the effects of warming a decade earlier than the re-oligotrophied or unmanipulated systems, pointing to the synergism between eutrophication and climate change. Notably, this climate change effect was apparent a decade earlier in sediment records than in monitoring. Overall, our study demonstrates the reliability of paleogenetic analysis for reconstructing historical algal community dynamics and the early detection of multiple stressors.

5.2 Introduction

The effects of eutrophication on freshwater ecosystems are being intensified by climate change (Jeppesen *et al.*, 2014; Rodgers, 2021). Eutrophication is a pervasive water quality issue promoting algal blooms and subsequent hypoxia, acidification, fish and benthic invertebrate mortality, and greenhouse gas production (Huisman *et al.*, 2018). At the lake scale, one of the key ways in which climate change intensifies eutrophication is through warming surface water temperatures which reduce vertical mixing, including

oxygen transport (Woolway, Sharma & Smol, 2022). Under low-oxygen conditions prolonged by stronger stratification, reactive phosphorus is recycled from the sediments into the water column. Climate warming can thus indirectly prolong or intensify eutrophication via internal loading, even when fertilization from the catchment is reduced (Søndergaard, Jensen & Jeppesen, 2003). Lake warming also alters species thermal habitats (Kraemer *et al.*, 2021), leading to greater biodiversity loss in freshwater ecosystems already experiencing the sharpest declines (Albert *et al.*, 2021). Eutrophication is predicted to become more severe and widespread as excess fertilizers are applied to meet increasing global demands for food and biofuels. At the catchment scale, climate change may exacerbate terrestrial runoff through increased precipitations, with repercussions on how lakes function as sinks or sources of nutrients (Wu *et al.*, 2022) and carbon (Tranvik *et al.*, 2009). All together, these links point to reciprocal feedbacks between eutrophication and climate change (Hessen *et al.*, 2023).

Long-term perspectives on lake ecosystem dynamics are opened through the analysis of sediment DNA archives (Domaizon *et al.*, 2017; Capo *et al.*, 2021). The gradual effects of eutrophication and even slower effects of climate change are well suited to be studied from paleolimnological records, though their effects are often challenging to disentangle (Davidson & Jeppesen, 2013). Progress has been made in this area by applying a paleogenetic approach to study multiple-stressor responses of microeukaryotes (Keck *et al.*, 2020; Ibrahim *et al.*, 2021; Barouillet *et al.*, 2022), a group occupying diverse ecological niches (Caron *et al.*, 2009). For example, in the paleogenetic study of two re-oligotrophied lakes, microeukaryote assemblages responded more strongly to air temperature than to total phosphorus (TP) beginning in 1980s-era sediments (Capo *et al.*, 2017). While sediment DNA captures most of the microeukaryotic groups represented in the contemporary water column (Capo *et al.*, 2015), there are few studies which have investigated whether sediment records reflect the historical dynamics of lake ecosystems. A sediment trap experiment aiming to calibrate DNA deposition from the water column found strong congruence when only microeukaryotes common to both sources were considered (Gauthier *et al.*, 2021). A more direct approach is to compare sediment DNA and long-term monitoring records. However, systems with sustained biological monitoring are exceedingly rare and have mostly focused on Cyanobacteria (Monchamp *et al.*, 2016; Thorpe *et al.*, 2022; Mejbøl *et al.*, 2023). The one previous study on eukaryotic algae identified temporal disconnects between taxonomic turnover in long-term monitoring and sediments (Thorpe *et al.*, 2023).

A unique opportunity to compare monitoring and sediment records and their responses to historical nutrient and climate variation presents itself in the International Institute for Sustainable Development Experimental Lakes Area (IISD-ELA). Long-term nutrient addition experiments simulating anthropogenic eutrophication were conducted in Lake (L)227, a naturally low-carbon lake that has been continuously fertilized since 1969, and L226, which was divided by a curtain and fertilized with carbon (C) and nitrogen (N) in one basin and C, N, and phosphorus (P) in the other. While both basins of L226 saw large increases in phytoplankton biomass, the P-enriched northeast basin experienced even greater growth (Findlay & Kasian, 1987). In the years since the end of the two-basin experiment, L226 is considered to have

recovered from eutrophication. The observation that P was the limiting nutrient led to a broader awareness and legislation to manage P pollution (Schindler *et al.*, 2008). Despite discontinued N additions to L227 in 1990, artificial P loading, a large sediment P reservoir (O’Connell *et al.*, 2020), and biological N fixation have sustained elevated phytoplankton biomass (Higgins *et al.*, 2018). Warming temperatures were also linked to earlier and larger phytoplankton blooms in L227 (Salk *et al.*, 2022). Unmanipulated reference lakes in the IISD-ELA have been useful for isolating the effects of climate change, including shifts to mixotrophy in the phytoplankton (Findlay *et al.*, 2001). As phytoplankton monitoring efforts were initiated in 1969, more substantial pre-manipulation records are needed to anchor the trajectories of the experimental systems in baseline conditions. Paleolimnological reconstructions of pre-manipulation algal assemblages in these lakes have been based on preserved subfossils (Zeeb *et al.*, 1994; Enache, Paterson & Cumming, 2011) and pigments (Leavitt & Findlay, 1994; Cottingham, Rusak & Leavitt, 2000) limited to select groups.

In this study, we leveraged the multidecadal monitoring histories and 120-year sediment DNA records of experimentally fertilized and unmanipulated reference lakes in the IISD-ELA. Our first objective was to evaluate the congruence between the temporal dynamics of algal assemblages recorded in monitoring and sediment DNA time series to assess the reliability of paleogenetic archives. Second, we modeled the contributions of eutrophication and climate change to driving algal turnover across the different timescales presented by monitoring and paleogenetics. We hypothesized that the effects of both nutrients and climate would be evident in experimentally fertilized lakes (L226, L227), whereas unmanipulated lakes (L224, L373) would be influenced primarily by climate. Additionally, we hypothesized that the effects of climate would manifest earliest in continuously fertilized L227 due to synergies between eutrophication and warming. Overall, our study advances a clearer understanding of the relationship between sediment DNA archives and the historical diversity they are taken to represent. Our study further compares how monitoring and paleogenetics can each be used to detect the effects and timing of multiple stressors.

5.3 Methods

5.3.1 Lake experimental histories

Experimentally fertilized and unmanipulated reference lakes were sampled in the IISD-ELA in northwestern Ontario, Canada (Figure 5.1). L226 is divided by a plastic curtain; between 1973–1980, the southwest basin (L226S; maximum depth 11.6 m) was fertilized with C and N and the northeast basin (L226N; 14.7 m) with C, N, and P. Other experiments in L226 were radioisotope tracer additions (1977–1978; 1989) and water level drawdown (1995–1997) (Turner *et al.*, 2005). A tear in the curtain in 1977 leaked phosphorus into the southwest basin for up to one month (Findlay & Kasian, 1987). L226 is considered to have recovered to pre-fertilization water chemistry and phytoplankton composition (Turner *et al.*, 2005). Different stoichiometric ratios of N and P were loaded into L227 (maximum depth 10.0 m) over the years: an initial five-year experiment (1969–1974) loading N-to-P in a 12:1 ratio by weight was designed to test carbon limitation under N+P levels for balanced phytoplankton growth; in the following 14 years (1975–1989), N-to-P was loaded 4:1 to test whether nitrogen-fixing Cyanobacteria are enriched below the Redfield

ratio; since 1990, only P has been added to test the potential for controlling algal blooms by restricting N loading (Schindler *et al.*, 2008). Between 1993–1996, northern pike were added then removed from L227 to test the effects of a trophic cascade on phytoplankton biomass (Elser *et al.*, 2000). The hypolimnion of L227 has likely been anoxic for most of the year in the 300 years predating experimentation, as evidenced by seasonal biogenic varves typically formed under bottom water anoxia (Wolfe *et al.*, 1994). L224 and L373 (maximum depths 27.3 and 21.2 m, respectively) are designated Long-Term Ecological Research sites and were not manipulated except for metal radiotracer additions in L224 in 1976 (Hesslein, Broecker & Schindler, 1980).

5.3.2 Monitoring data

Monitoring data were obtained with permission from the IISD-ELA. Daily minimum and maximum air temperatures were measured at the L239 meteorological station beginning in 1969. We also accessed historical weather data (daily mean air temperature) from the Kenora (1900–1938) and Kenora A (1939–2018) stations (Environment Canada; climate.weather.gc.ca). Missing daily data were filled in with the monthly average based on available days; missing monthly data were linearly interpolated between the same month over different years. Mean annual air temperature data from L239 and Kenora (A) stations were strongly correlated for the years 1970–2017 (mean vs. maximum air temperature $r = 0.97$), therefore climate data since 1900 from Kenora (A) stations were used in all statistical analyses.

Epilimnetic total nitrogen (TN) and TP were each calculated as the sum of the total dissolved nutrient and suspended nutrient concentrations. Mean annual nutrient concentrations were calculated based on data that were linearly interpolated for missing spring, summer, and fall seasons across years (winter data were sparse and therefore excluded).

Phytoplankton biomass was monitored using morphological identification by light microscopy. Winter and fall samples were removed for lack of consistently sampled data. Total annual spring-summer biomass was calculated for each taxon, which was generally resolved to species. The phytoplankton monitoring taxonomy was harmonized to the Protist Ribosomal Reference database (PR²) v.5.0.0 (Guillou *et al.*, 2012) to enable comparison with the molecular genetic data and to assign trophic functions according to the method described below. Only taxa that were assigned to phototrophs or mixotrophs were retained as algae.

5.3.3 Lake sampling

Surface water and sediment cores were collected between August 11–16, 2018 by the Natural Sciences and Engineering Research Council (NSERC) Canadian Lake Pulse Network (Huot *et al.*, 2019). Sampling was conducted above the deepest point of the lake (NSERC Canadian Lake Pulse Network, 2021). Surface water was collected over an integrated depth of twice the Secchi disk depth up to 2 m. Onshore, up to 500 mL of surface water from each lake were filtered on 0.22- μ m Durapore membranes or until the filter was nearly clogged. Sediments were collected using a NLA gravity corer. Most of the overlying water was siphoned off

from each core tube and the remainder was jellified through the incremental addition of sodium polyacrylate (Zorbitrol). Core tubes were wrapped in black plastic bags to prevent light infiltration and shipped to Université Laval in Quebec City, Canada, where they were stored horizontally at 4 °C.

5.3.4 Sediment dating and subsectioning

Sediment cores were split into vertical halves. Age-dating via gamma spectrometry measuring lead-210, lead-214, and cesium-137 was performed on one core half as described by Baud *et al.* (2023). Dating models for each core were systematically selected from among the constant flux and constant sedimentation, constant initial concentration, and constant rate of supply methods (Baud *et al.*, 2022).

The second core halves were subsectioned at 0.5-cm intervals starting with the oldest sediments at the bottom. Subsectioning was performed in a dedicated room using precautions to avoid contamination with modern DNA, including the use of sterile equipment. Subsamples for genetic analysis were collected from the centre of intervals into sterile 5-mL tubes to avoid smeared or contaminated outer sediments in contact with the core tube. Subsamples were stored at -20 °C freezer.

5.3.5 IISD-ELA filter archives

To assess the microeukaryotic diversity along different water column depths, GF/C (1.2-µm) filters originally collected for chlorophyll *a* analysis were accessed from the IISD-ELA freezer archive. Water was filtered from L227 at 4-, 6-, 8-, and 10-m depths on July 10, August 14, and October 10, 2018 and from L373 at 5-, 10-, 15-, and 20-m depths on July 9, August 13, and October 9, 2018.

5.3.6 DNA extraction, amplification, and sequencing

DNA extraction from sediments was performed in a dedicated facility using NucleoSpin Soil kits (Macherey-Nagel) with cell lysis SL1 and enhancer SX buffers. DNA extraction from water filters was performed using DNeasy PowerWater kits (QIAGEN). GF/C filters were first cut in half with metal scissors using aseptic technique and only one filter half was used in the extraction to preserve the remainder for the IISD-ELA archive. DNA was quantified in a Qubit 2.0 fluorometer (Thermo Fisher Scientific).

A ~265-bp fragment of the 18S rRNA gene V7 region was amplified using the 960F (5'-GGCTTAATTTGACTCAACRCG-3') (Gast, Dennett & Caron, 2004) and NSR1438 (5'-GGGCATCACAGACCTGTTAT-3') (Van de Peer, 2000) primer pair. Each 25-µL PCR mixture contained 14.25 mL Milli-Q water, 5 µL High-Fidelity buffer, 1.25 µL of each primer (10 µM), 0.5 mL dNTPs (10 mM), 0.25 µL dimethyl sulfoxide, 0.5 µL Phusion polymerase (Thermo Fisher Scientific), and 2 ng DNA. PCR conditions comprised a 1-min initial denaturation at 98 °C, 30 cycles of 10 s at 98 °C, 30 s at 60 °C, and 20 s at 72 °C, and a final 5-min extension at 72 °C. PCR products were electrophoresed in 2% agarose gel for 100 min at 40 V then extracted at the target band length using QIAquick Gel Extraction kits (QIAGEN) modified by final elution into Milli-Q. PCR products were submitted to Genome Quebec for library preparation and sequencing on an Illumina MiSeq.

5.3.7 Bioinformatic analysis of 18S rRNA amplicons

Primer sequences were trimmed from demultiplexed reads in Cutadapt v.3.1 (Martin, 2011). The DADA2 v.1.16 pipeline was used to quality-control reads, infer amplicon sequence variants (ASVs) across pooled samples, merge reads, remove chimeric contigs, and assign taxonomy (Callahan *et al.*, 2016). Taxonomy was assigned with PR² v.5.0.0 (Guillou *et al.*, 2012). To further quality-control the ASV dataset, poorly aligned ASVs were removed based on a *de novo* alignment performed in MAFFT (Katoh *et al.*, 2019). Remaining primer sequences were trimmed based on the alignment of ASVs against the SILVA 99% non-redundant small subunit rRNA database v.138.1 performed in SINA v.1.7.2 (Pruesse *et al.*, 2012; Quast *et al.*, 2012). ASVs assigned to Metazoa (animals) or Embryophyceae (land plants) were removed to retain a dataset of protists and fungi. Taxa were assigned to trophic functional groups based on nutritional modes reviewed by Adl *et al.* (2019) and FUNGuild, a database specific to fungi (Nguyen *et al.*, 2016). ASVs of phototrophic and mixotrophic taxa were categorized as algae.

5.3.8 Statistical analyses

Data analysis and visualization were performed in R v.4.2.0. Principal component analyses (PCAs) of the taxonomic variation among Hellinger-transformed monitoring and sediment assemblages were performed using the functions `decostand()` and `rda()` in `vegan` v.2.6.4 (Oksanen *et al.*, 2020).

Multivariate correlations among algal monitoring and sediment DNA data were calculated using RV coefficients in FactoMineR v.2.3 (Lê *et al.*, 2008). Assemblages were Hellinger-transformed prior to computing RV coefficients. To compare annually resolved monitoring assemblages with multiyear sediment assemblages, we averaged the phytoplankton biomass within the timeframes captured in sediment intervals. Assemblage comparisons were made at the level of monitoring “taxon codes” – the finest level of taxonomic resolution by microscopy – and sediment ASVs. Correlations among monitoring and sediment data were calculated for each lake and separately for epilimnion and metalimnion monitoring assemblages.

Generalized additive models (GAMs) were applied to assess the timing of nutrient and climate effects on algal time series. Because nutrient and climate data were available for different timeframes (after 1969 for nutrients and since 1900 for climate), separate GAMs were constructed using nutrient (epilimnetic TP and TN) and climate (air temperature) variables. To compare the contributions of nutrients and climate over a common timeframe (after 1969), a third set of GAMs were constructed based on nutrient and climate variables. Response data were the first principal component dimension (PC1) for algal assemblages in individual lakes. GAMs were constructed in `mgcv` v.1.8.31 using restricted maximum likelihood estimation for model smoothing. When the estimated degrees of freedom were within 0.5–1.5 for a particular variable, the GAM was recomputed with the linear term. Because smoothing functions are limited by the number of response data points, the number of knots used to smooth terms was iteratively decreased until convergence was reached; if decreasing the number of knots to 1 did not suffice to reach convergence, linear terms were

used. To pinpoint where in time variables had nonzero effects, we plotted the model regression slopes over time using the function `predict.gam(type = "terms")`.

5.3.9 Code availability

Scripts associated with this study are available at <https://github.com/rebeccagarner/paleoELA>.

5.3.10 Data availability

Amplicon reads were deposited at the European Nucleotide Archive under study accession PRJEB47328 (<https://www.ebi.ac.uk/ena/browser/view/PRJEB47328>). Monitoring data are available upon request from the IISD-ELA (<https://www.iisd.org/ela/science-data/our-data/data-requests/>).

5.4 Results

5.4.1 Phytoplankton monitoring

Phytoplankton monitoring was summarized for the years of available data in experimental lakes L226 (L226N 1971–2004; L226S 1973–2004) and L227 (1970–2018) and reference lakes L224 (1974–2018) and L373 (1983–2018). Algal biomass and composition identified by light microscopy varied by lake and through time (Figure 5.2A). L227, which was continuously fertilized under decreasing N:P regimes, experienced the largest changes in total annual spring-summer biomass (672–19,537 $\mu\text{g/L}$). Algal assemblages in L227 were characterized by streptophytes (e.g., *Spondylosium*, *Staurodesmus*), chlorophytes, chrysophytes, dinoflagellates, and cryptophytes. Chlorophyte biomass peaked in the early 1970s after initial fertilization, whereas streptophyte biomass peaked in 1988 during the 4N:P additions. Each of the L226 basins, which were fertilized between 1973–1980, also experienced pronounced changes in biomass (L226N 428–8,341 $\mu\text{g/L}$; L226S 441–10,274 $\mu\text{g/L}$). Algal assemblages in the two L226 basins were characterized by high diatom and chrysophyte biomass during the mid-1970s experimental era and early 1980s. Unmanipulated reference lakes had consistently low biomass (L224 211–924 $\mu\text{g/L}$; L373 171–1,003 $\mu\text{g/L}$) dominated by chrysophytes in L224 and characterized by dinoflagellates, haptophytes, and chrysophytes in L373.

5.4.2 Microeukaryotic diversity in lake sediments

Microeukaryotic diversity was investigated along 0.5-cm sediment core intervals (Figure 5.2B). Sediment stratigraphies from all lakes extended to the preindustrial background for the region (1881–1885) as evaluated by radionuclide dating models. Differences in depositional ages of sediments among lakes meant that cores were subsectioned to different depths and intervals captured variable timeframes. At one extreme, the L373 core was subsectioned into 18 intervals to reach preindustrial background and captured 7–8 years per slice, while at the other extreme, the L224 core had 39 intervals with 2–7 years captured per slice.

In total, 3,953 18S rRNA gene ASVs encompassing 6,345,459 sequences were inferred across 135 sediment samples spanning the five cores (Figure 5.2B). Gyrista, which contain the chrysophytes and

diatoms, were by far the most abundant eukaryotic subdivision in terms of total sequences and ASVs across samples (61.2%; 754 ASVs). Apicomplexa (12.6% of sequences) followed by Fungi (10.3%) were the next most abundant subdivisions, particularly in the sediments of fertilized lakes. Fungi were the next most taxonomically rich subdivision (735 ASVs). Each of the 135 sediment layers contained 1,645–54,473 sequences assigned to Chrysophyceae, comprising 2.4–92.0% of the sequence space in each sample. Trophic functions were assigned to 76% of ASVs encompassing 95% of sequences. Assigning trophic functional diversity served a dual purpose: to identify preserved groups whose nutritional modes rely on access to sunlight and are therefore unlikely to grow in the sediment column *in situ*, and to enable comparison between algal assemblages in monitoring and sediment assemblages.

5.4.3 Water column diversity not captured in sediment DNA records

Most taxonomic groups detected in water column samples via amplicon sequencing and light microscopy were captured in sediment DNA archives, with some exceptions. A qualitative comparison of depth-discrete microeukaryotic water column diversity identified on 1.2- μ m filters from L227 and L373 showed relatively high proportions of ASVs assigned to Cryptophyta and Kathablepharida (Figure S1A) that were mostly absent from the sediment assemblages. The presence of Centroplasthelida and Choanoflagellata in addition to Cryptophyta and Kathablepharida were also noted from surface water assemblages identified on 0.22- μ m filters from L226S, L227, L224, and L373 (Figure S1B) but lacking from sediments. Cryptophyta and phototrophic Euglenozoa were evident in phytoplankton monitoring based on microscopy (Figure 5.2A) but absent from sediments.

5.4.4 Taxonomic turnover in monitoring and sediment records

Considering the temporal dynamics evident in the PCAs (Figure S2A), 1970s experimental-era L226 monitoring assemblages from both basins were the most distinguished along PC axes 1 and 2, whereas post-manipulation assemblages exhibited more limited taxonomic variation. The first two decades of L227 monitoring assemblages were distinguished along PC2, while assemblages from the next three decades were mostly chronologically arranged along PC1. L224 assemblages were likewise chronologically arranged along PC axes 1 then 2. In contrast, L373 phytoplankton assemblages did not follow a coherent temporal trajectory but were dispersed in the PCA space. Among lakes (Figure 5.3A), L227 assemblages were notably distinct, though pre-manipulation assemblages were not available for comparison. Fertilization-era L226 assemblages from both basins were distinguished from the later recovery-era assemblages, which were taxonomically similar to assemblages from unmanipulated L224 and L373.

In contrast to the patterns in the monitoring time series, sediment algal assemblages were the most taxonomically distinguished among the most recent layers in all lakes (Figure S2B). Among lakes however (Figure 5.3B), while assemblages from recent layers diverged, pre-manipulation-era assemblages from experimental lakes (L226 and L227) were taxonomically similar. Likewise, there was high taxonomic overlap between sediment assemblages in unmanipulated L224 and L373.

5.4.5 Correlations between monitoring and sediment records

We compared the temporal dynamics of algal assemblages in monitoring and sediment records by calculating multivariate correlations (RV coefficients) between data matrices (Figure 5.4). The RV analyses required synchronizing monitoring and DNA datasets to common timeframes, which meant summarizing annually resolved monitoring assemblages within the estimated years captured by each sediment interval. Monitoring and sediment time series in both L226 basins, L227, and L373 had strong or very strong RV coefficients (epilimnion 0.69–0.90; metalimnion 0.65–0.87). L224 time series had the lowest RV coefficients, yet these were still moderately strong (epilimnion 0.53; metalimnion 0.56). There were marginal differences in RV coefficients when correlating either epilimnion or metalimnion monitoring biomass to sediment assemblages, except in L373 which had >10% higher RV coefficient when using the metalimnion assemblages. All other lakes had slightly greater RV coefficients when using epilimnion monitoring assemblages.

5.4.6 Effects of nutrients and climate on algal diversity

We next assessed the effects and timing of nutrients and climate on algal assemblages in monitoring and paleogenetic time series. To this end, we constructed GAMs using PC1 dimensions describing taxonomic variation through time as the response variables. Nutrients (epilimnetic TP and TN), climate (air temperature), and a combination of nutrients and climate were the selected predictor variables across separate GAMs. Like many regions of the world, air temperature trended upward over the last hundred years (Figure S3). Nutrient concentrations varied strongly as a function of experimental manipulation within and among individual lakes (Figure S3). Models were constrained by the years of available data, therefore GAMs using nutrient predictors (or in combination with climate) leveraged data from the start of monitoring (1969 or after), whereas GAMs using air temperature only were based on data from 1969 or after to fit phytoplankton monitoring or since 1900 to fit sediment records.

In GAMs fitting the PC1 axis of annually resolved algal monitoring, models constructed with nutrient predictors alone explained high deviance for experimentally fertilized lakes (35–73%) and relatively low deviance for unmanipulated reference lakes (4–6%) (Figure 5.5). GAMs based on air temperature alone did not explain large deviance for any of the lakes (<1–13%) and were significant only for L226N, L227, and L224 (in descending order of deviance; $p < 0.05$). GAMs combining nutrient and climate predictors again explained high deviance for fertilized lakes (36–74%) and lower deviance for unmanipulated lakes (6–13%), however, not all variables were significant across lakes. Only nutrient variables were significant for L226N and L226S, only air temperature was significant for L224, and both a nutrient and climate variable were significant for L227. In all models where nutrient variables were significant, TP was significant for L226N, while TN was significant for L226S and L227; nutrient variables were never significant in L224 or L373.

In GAMs fitting the PC1 axis of sediment algal assemblages, models constructed with nutrients alone explained high deviance for all fertilized lakes (62–100%) and unmanipulated L373 (99%), though nutrient variables were significant only for the experimental lakes (Figure 5.5). The L224 nutrient GAM explained 14% deviance without significant variables. For all lakes, GAMs based on climate alone explained moderate to high deviance (36–74%; led by L227) and air temperature was always significant. Across the GAMs based on nutrients and climate, only the L227 model explained higher deviance (87%) than its separate nutrient or climate GAMs and had significant variables from both nutrients (TN) and climate. L226N and L226S GAMs explained the same deviance (62–100%) as their nutrient-only GAMs. For unmanipulated L224 and L373, combined nutrient and climate GAMs were weaker than GAMs based on nutrients or climate alone, and neither model contained significant variables. The underperformances of the reference lake GAMs may be a consequence of inadequate smoothing of model terms, as the number of knots permitted in smoothing functions is limited by the number of response data points; this issue was exacerbated when GAMs incorporated multiple predictor variables.

The modeled effects of each predictor on algal monitoring PC1 over time are depicted in GAM contribution plots (Figure S4A). In L226N, the effect of TP switched from a positive to negative slope in the early 1980s (no effect of TN) and was always stronger than the effect size of air temperature. In L226S, TN had a stronger and opposite effect to that of air temperature (no effect of TP). In L227, while the effect of TP was consistently negative, the effect of TN exhibited a complex trend and several years without effect. The effect of air temperature tracked with that of TP and surpassed the strength of the effect of TN in the mid-1990s. In L224, the effects of TP and air temperature were both negative, overshadowing the small effect of TN, but did not generally change within four decades of data. In L373, the effects of TP and TN were both positive but stronger for TN; there was no substantial effect of air temperature.

The GAMs fitting L226N sediment assemblages showed mirroring effects of TN and TP and a very marginal effect of air temperature toward 2000 (Figure S4B). In L226S, TP had a positive effect, whereas a negative effect of air temperature emerged in the late 1990s. In L227, the effect of TP was always negative and peaked in strength around 1980, while TN exhibited a complex trend. Air temperature became the strongest driver in the early 1980s. In L224, the effect of TP was negative and stronger than that of air temperature, which had a marginal effect that became negative starting in the 1990s. In L373, TP and air temperature were both negative and opposite to TN, with air temperature having the strongest effect.

5.5 Discussion

Here, we leveraged a collection of rare multidecadal monitoring and sediment paleogenetic archives from the IISD-ELA to examine the coherence among time series and investigate the effects of two intensifying and interacting freshwater stressors: eutrophication and climate change. We evaluated the temporal dynamics of algal communities in lakes with a broad spectrum of eutrophication histories, including a continuously fertilized lake under varying N:P regimes, a two-basin fertilization experiment that has since recovered to oligotrophic conditions, and two unmanipulated reference lakes.

5.5.1 Strong congruence between monitoring and paleogenetic records

RV coefficients measured strong correlations between contemporaneous algal monitoring by light microscopy and algal paleogenetic archives. This is impressive considering that the time series were constructed using two very different analytical techniques that measure different breadths of microbial diversity and at different taxonomic resolutions. Furthermore, whereas monitoring captures discrete water column samples, sediments integrate microeukaryotic diversity throughout the year and among years from across the water column, other parts of the lake, indigenous sediment microbiota, and the catchment. RV coefficients in our analysis were higher than a monthly time series spanning three years that correlated shared microeukaryotic ASVs present in the water column and sediments traps designed to track DNA burial (Gauthier *et al.*, 2021). The strength of the RV coefficients in the present study is likely due to the multiyear dynamics captured in each sediment interval (2–8 years), which represented smoothed averages reducing the noise that is sampled by more sharply resolved time series. By necessity of having to summarize monitoring and paleogenetic time series within the same matrix dimensions to calculate correlations, the bimonthly monitoring was smoothed to average multiyear biomass. The RV coefficients were likely additionally strengthened by the comparisons of sufficiently long time series.

We expected metalimnion assemblages to be more substantially represented in sediment DNA because their closer physical proximity to sediments might be positioned for preferential preservation. Contrary to our expectation, and perhaps because the taxonomic and temporal distinctions between epilimnetic and metalimnetic algal assemblages were not generally large, correlations between assemblages from either stratum with sediment records were highly comparable. This demonstrates that sediments in these small hemi-boreal lakes integrate the temporal dynamics along different water column depths equally reliably.

To our knowledge, the results here represent the first demonstration of strong congruence between long-term monitoring and the paleogenetic records of microeukaryotes in multiple lakes; a critical validation step in the analysis of lake sediment DNA to reconstruct historical ecosystem dynamics. Coupled with the taxonomic scope enabled by molecular genetic techniques, lake sediments represent a trove of data that reliably reflect the temporal dynamics of the overlying water column. However, some phytoplankton groups are consistently underrepresented in sediment archives (discussed below).

5.5.2 Effects and timing of nutrient and climate drivers in monitoring and paleogenetic time series

We analyzed the effects and timing of eutrophication and climate on lake algal assemblages in two types of time series differing dramatically in their temporal resolution and scale. As expected, nutrients were strong and statistically significant drivers of monitored assemblages in lakes with histories of experimental fertilization. The significant nutrient variables aligned with the experimental nutrient manipulations of each system: TP in L226N, which was fertilized with C, N, P and in which P was the limiting nutrient; TN in L226S, which was fertilized with C and N only; and TN in L227, in which N was added in decreasingly

N:P ratios until N addition was discontinued entirely. The effect of air temperature on monitored time series was not discernable, to the exception of L227 in which climate effects were detected beginning in the mid-1990s.

As observed in the monitoring time series, nutrients were strong drivers of algal assemblages in the sediment records of the experimentally fertilized lakes. Unexpectedly, L373 sediment assemblages were also strongly explained by nutrients, but unlike the fertilized lakes, did not have significant variables in its model. This may be a consequence of insufficient response data points (six sediment intervals with corresponding physicochemical monitoring data) leading to model overfitting. Thus, it may be harder to extract coherent environmental signals from lakes with low sedimentation rates like L373 which compress their histories.

Climate was a strong and significant driver of sediment assemblage dynamics across all lakes, regardless of trophic state history. The effects of climate were considerably stronger when tested over the 120-year perspective enabled by the pairing of paleolimnology and historical weather data, compared to the shorter (<50-year) timescale presented by monitoring. The effect of air temperature emerged earliest in L227 (early 1980s), a decade before the effect became apparent in the other systems and a decade before the effect was detected by the monitoring time series. Remarkably, this observation supports our hypothesis that the effects of climate change should appear earlier in eutrophic lakes, as the food web is primed to respond to warming temperatures. In the immediate responses of sediment assemblages to nutrients and their gradual responses to air temperature, we see the contrasting actions of fast (nutrient) and slow (climate) drivers on lake ecosystems (Carpenter *et al.*, 2001). Our findings based on algal groups are complemented by those on the broader microeukaryotic assemblages in two re-oligotrophied European lakes, in which the effects of air temperature surpassed those of *Daphnia*-inferred P in the 1980s (Capo *et al.*, 2017).

5.5.3 Microeukaryotic diversity in the paleogenetic record

Despite taxonomic biases and blind spots associated with PCR-amplification, this study succeeded in expanding the historical taxonomic diversity of algae and other microeukaryotes. The analysis of sediment DNA identified a broad microeukaryotic diversity beyond the diatoms and chrysophytes identified through subfossil microscopy (Zeeb *et al.*, 1994; Enache *et al.*, 2011) and the broader algal pigment groups in these lakes (Leavitt & Findlay, 1994). Paleogenetic analysis also increased the taxonomic resolution of chrysophytes, often to genus level, as subfossil scales are challenging to associate to specific taxa (Zeeb *et al.*, 1994). The chrysophyte ASVs dominating the sediment assemblages are likely true paleogenetic signals as this group was prevalent in the monitoring and more generally is common across freshwater lakes (Bock *et al.*, 2022). Dinoflagellates and other mixotrophs were abundant in the most recently deposited sediment surface layers, aligning with the shift to mixoplankton observed in reference lake monitoring (Findlay *et al.*, 2001). As mixotrophic protists are typically associated with the plankton and not the benthos, the mixotrophs in the sediment surface layers may point to paleogenetic signals of recent climate-driven shifts.

The paleogenetic techniques employed in this study detected a broad but incomplete fraction of the microeukaryotic diversity found at different water column depths. Other studies employing the same and other primer sets have also noted the underrepresentation of cryptophytes in sediment DNA archives (Capo *et al.*, 2015; Gauthier *et al.*, 2021; Thorpe *et al.*, 2023). The consistent absence of cryptophytes is not primer bias, as the same primer set applied to the water filter DNA analyzed in this study and larger-scale lake analyses have identified large abundances of Cryptophyta (Garner *et al.*, 2022). Instead, the absence has been suggested to be a diagenetic effect due to their being soft-bodied and the preferred prey of many zooplankton (Gauthier *et al.*, 2021).

The sediment assemblages of the experimentally fertilized lakes contained high abundances of ASVs assigned to Fungi, Apicomplexa, and other parasitic alveolates. Many of these groups are associated with anoxic or microaerophilic environments (Adl *et al.*, 2019), and thus are likely to thrive in oxygen-deplete sediments. The predominance of Fungi and Apicomplexa in sediments deposited after L226 was reported to have recovered to oligotrophic conditions suggests that the sediment microeukaryotic community is altered by eutrophication. Therefore, high relative abundances of these groups in the sediment column could be an indicator for past eutrophication, even for re-oligotrophied lakes.

5.5.4 Conclusion

In conclusion, we leveraged one of the world's rare multidecadal lake monitoring programs to investigate the reliability of paleogenetic records and compare how these time series can be used to clarify the effects and timing of eutrophication and climate change. The strong correlations of algal turnover in monitoring and sediment DNA archives provide a direct demonstration of microeukaryotic community temporal dynamics reflected in the paleogenetic record. These comparisons were carried out in five systems with extremely divergent histories of eutrophication (continuously fertilized, re-oligotrophied, and unimpacted), and thus provide a broad picture on the chronologies being compiled in sediment DNA archives. In testing the effects of eutrophication and climate in the monitoring and paleogenetic time series, we found support for our hypotheses in both record types that nutrients were strong drivers in fertilized lakes. Furthermore, based on the 120-year perspective enabled by paleogenetics, we found support for our hypothesis that the effect of rising air temperatures occurs earlier, possibly by ten years, in eutrophic lakes than in oligotrophic lakes. The clear advantage of paleolimnology over water column monitoring is that it can be performed retrospectively: in the absence of monitoring programs, sediment archives can be used to access a lake's history. In addition to this convenience, paleogenetics presents a reliable early detection system for multiple lake stressors.

5.6 Acknowledgements

This research was made possible through the dedication of the IISD-ELA scientists who initiated and maintained the whole-lake experiments and monitoring programs over 50 years. We thank the LakePulse Network, its director Yannick Huot, and in particular, Alexandre Baud and Maxime Fradette for collecting

the sediment cores and other sampling data. We acknowledge Chris Grooms at the Paleoecological Environmental Assessment and Research Lab, Queen's University for the sediment age-dating analysis. Thank you to Paul MacKeigan and Michelle Gros for assisting in the subsectioning of sediment cores. Funding was provided through a *Fonds de recherche du Québec – Nature et technologie* Doctoral Research Scholarship to R.E.G. and Canada Research Chairs held by D.A.W. and I.G.E. R.E.G. extends special thanks to Stephan Bronfman for awarding the Scholarship in Graduate Environmental Studies to support this project.

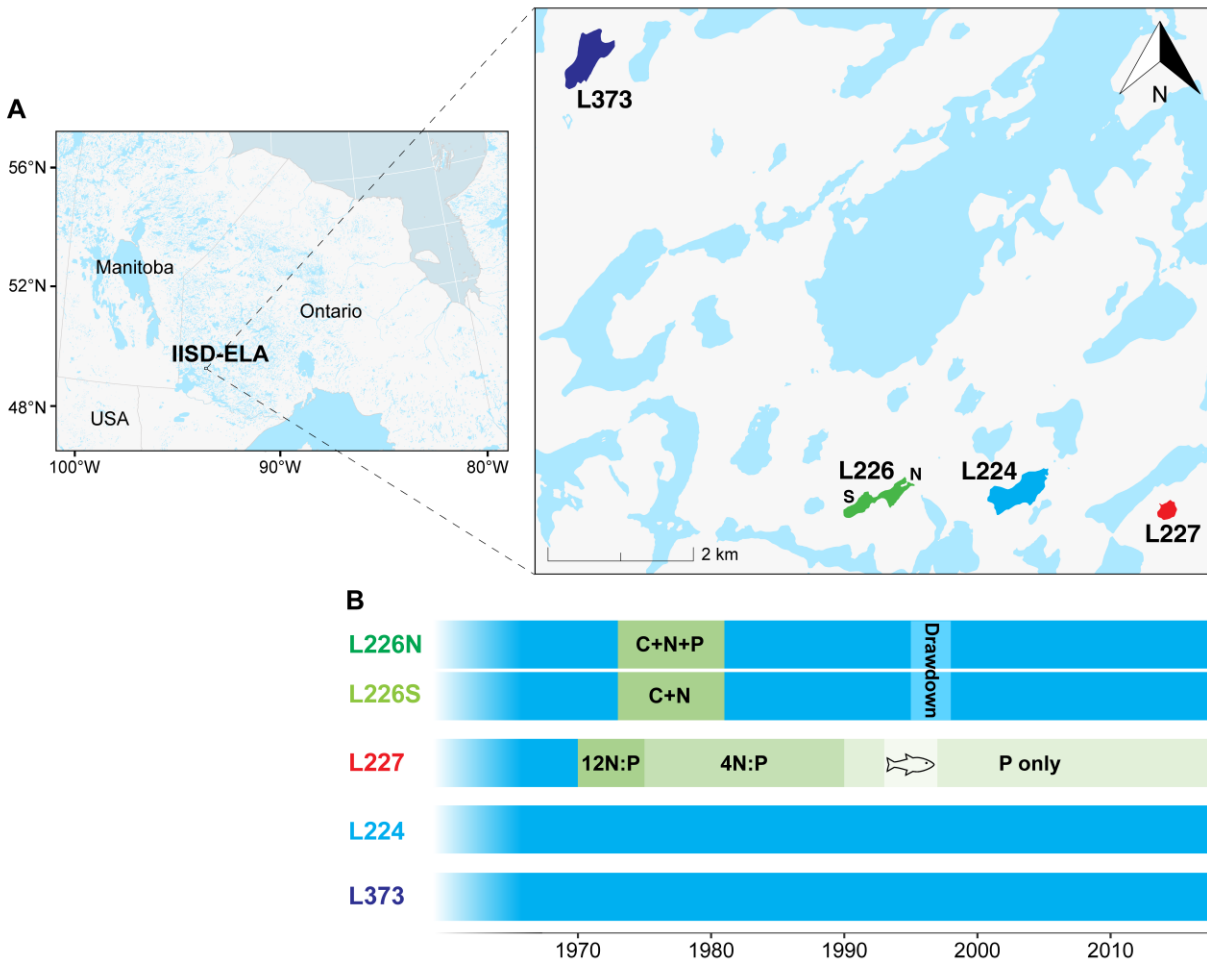


Figure 5.1 An overview of the study systems shown through (A) maps of the region and IISD-ELA lakes and (B) lake experimental histories. Lake timelines in blue indicated eras without experimental manipulation.

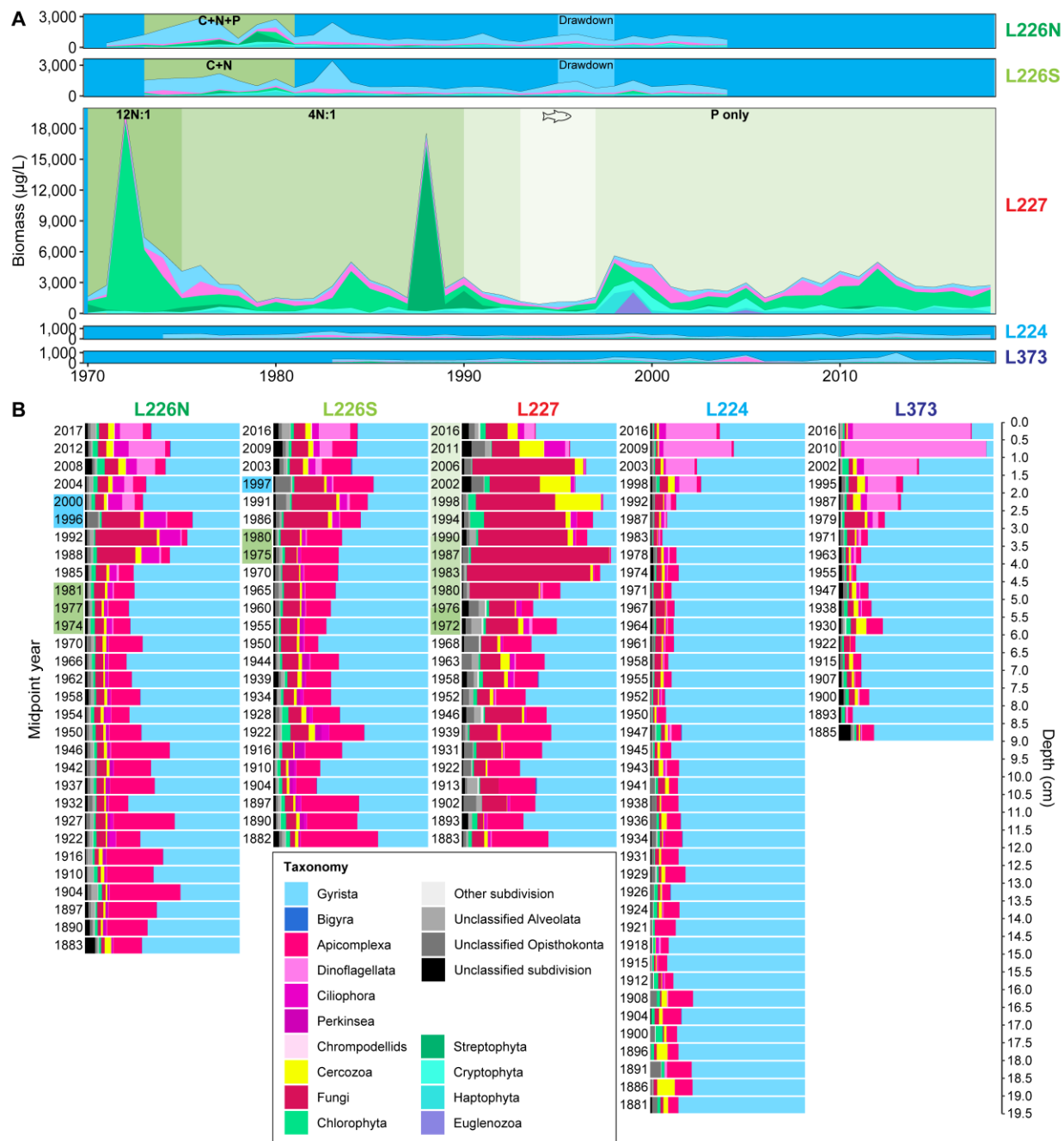


Figure 5.2 Microeukaryotic diversity in experimental and unmanipulated lakes recorded in (A) microscopy-based biomass monitoring and (B) ASV assemblages along 0.5-cm sediment intervals. Colours in the area and bar plots indicate taxonomic composition at the level of eukaryotic subdivisions (legend in panel B).

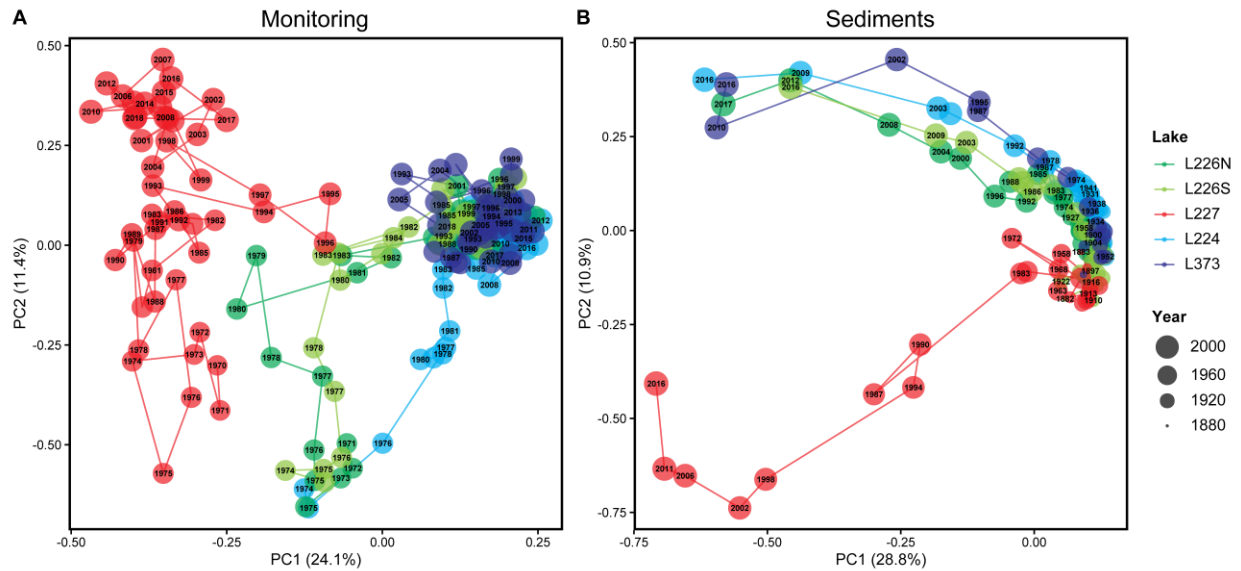


Figure 5.3 PCAs showing the taxonomic variation among algal assemblages in (A) monitoring and (B) sediment layers across all lakes. Paths between points link adjacent monitoring years or estimated sediment midpoint chronologies. The size of points indicate chronology; years of monitoring or modeled sediment midpoints are shown.

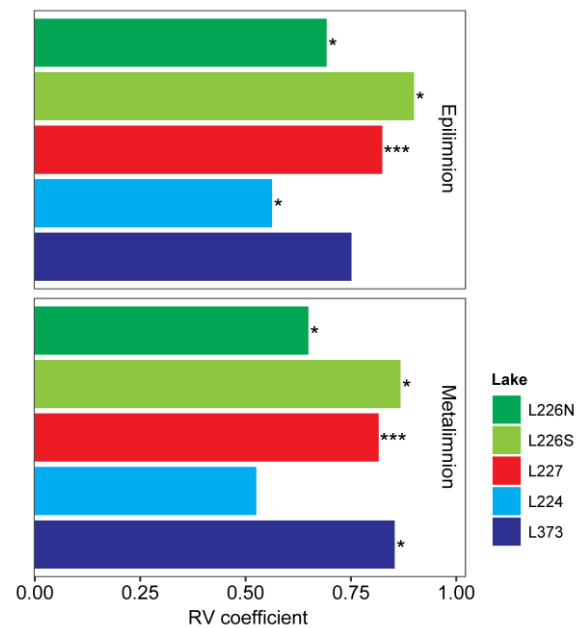


Figure 5.4 RV coefficients correlating the temporal dynamics of algal assemblages in monitoring and sediment records. Correlations are shown between monitoring assemblages from the epilimnion or metalimnion and sediment assemblages. Asterisks indicate the level of statistical significance: $p < 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***).

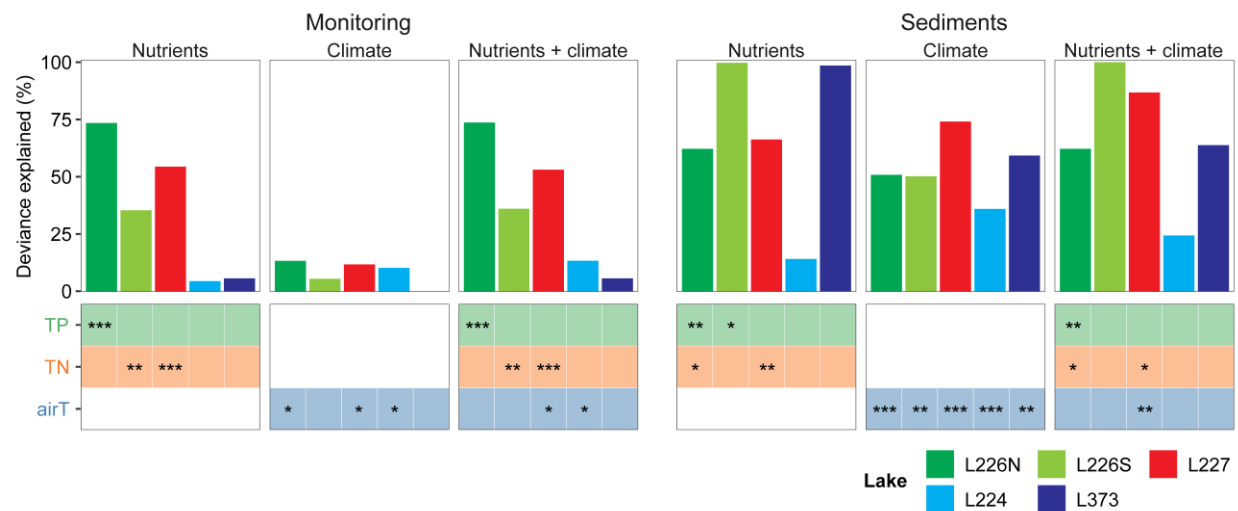


Figure 5.5 Deviance explained by GAMs fitting algal taxonomic variation in monitoring or sediment time series using nutrient (TP and TN) and/or climate (air temperature) predictors. Asterisks indicate the level of statistical significance for predictor variables included in models: $p < 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***).

Supplemental materials

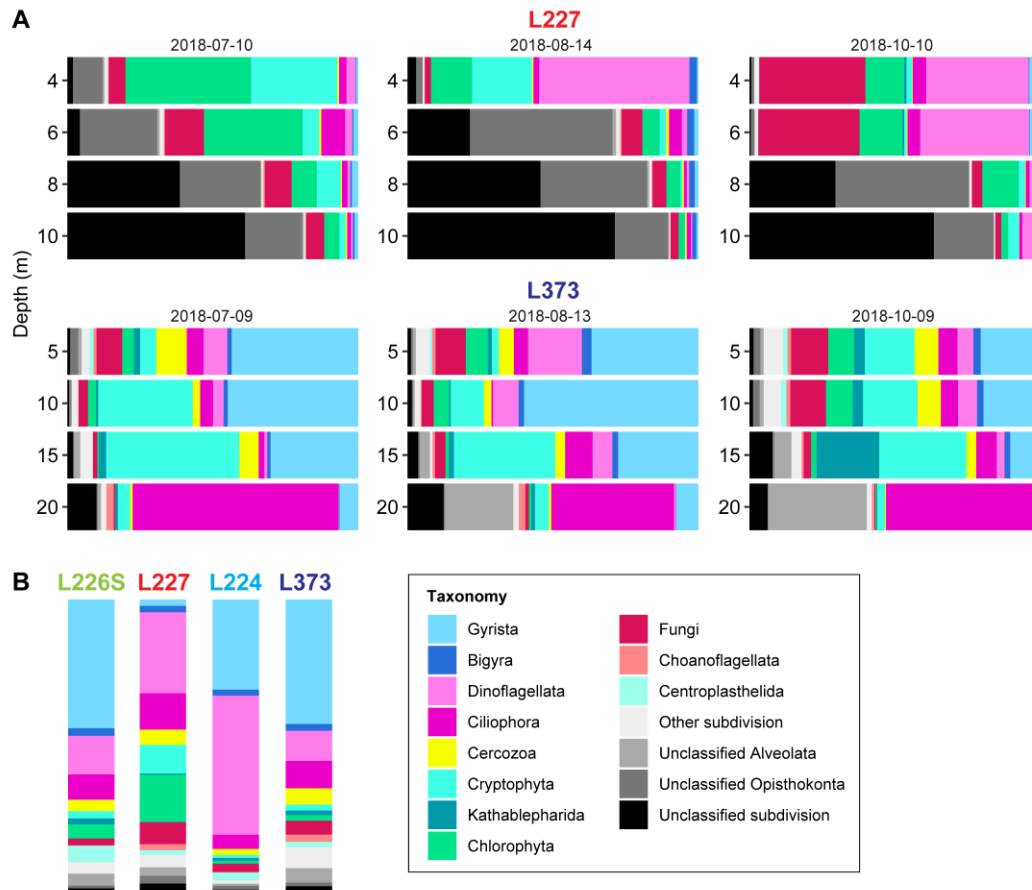


Figure S1. Taxonomic composition of microeukaryote assemblages captured on (A) depth-discrete 1.2- μ m GF/C filters and (B) integrated surface water 0.22- μ m GF/F filters in 2018.

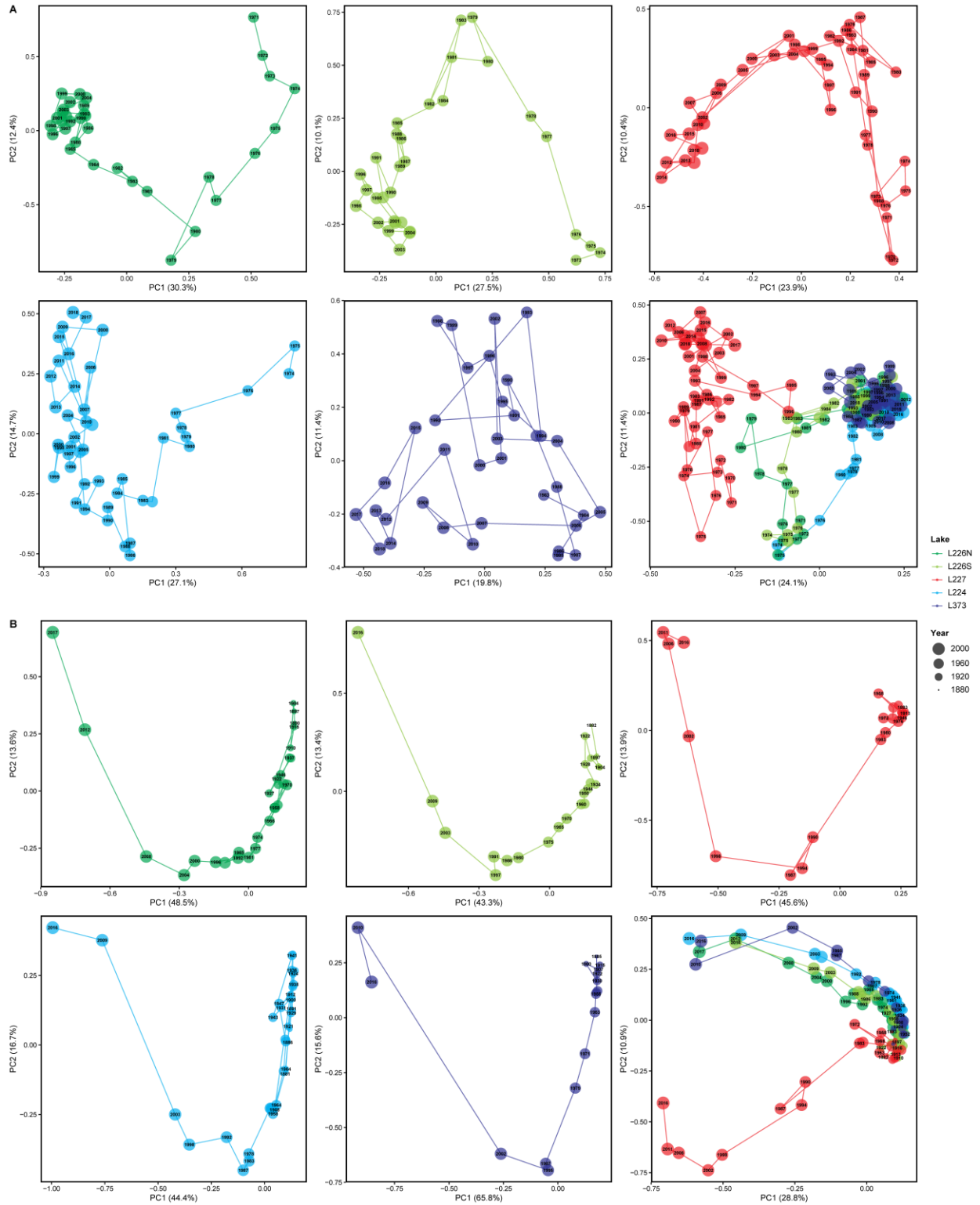


Figure S2. PCAs showing the taxonomic variation among algal assemblages in (A) monitoring and (B) sediment layers within and among lakes. Paths between points link adjacent monitoring years or estimated sediment midpoint chronologies. The size of points indicate chronology; years of monitoring or modeled sediment midpoints are shown.

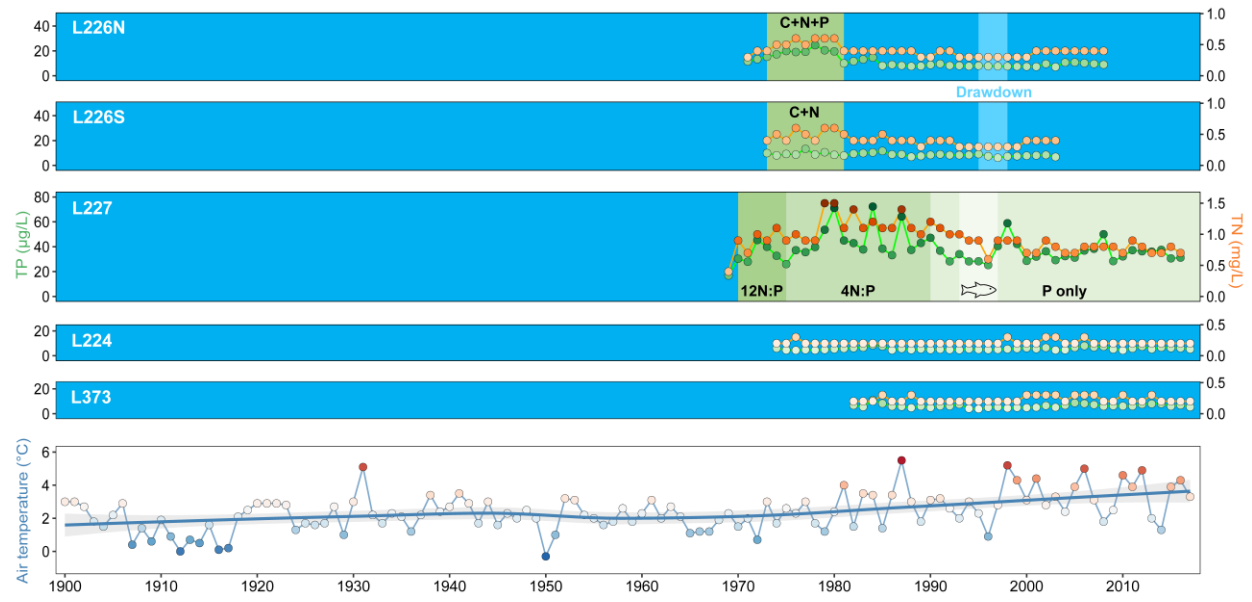


Figure S3. Nutrient (epilimnetic TP and TN) and air temperature trends summarized from IISD-ELA and meteorological station monitoring data. Background colours in nutrient plots show the experimental timelines (blue indicates eras without manipulation).



Figure S4. GAM contribution plots showing the modeled effects of nutrient (TP in green and TN in orange) and/or climate (air temperature in blue) predictors on algal time series. Points above zero indicate positive effects on the response variable (PC1) and points below zero indicate negative effects. Statistically significant variables ($p < 0.05$) are denoted by solid lines and non-significant variables are denoted by dashed lines.

Chapter 6

Discussion

Through their immense environmental heterogeneity at multiple spatial and temporal scales (Kratz, MacIntyre & Webster, 2005), lakes and their watersheds comprise a mosaic with few repeating pieces. This dissertation has shown that the lake microbiome represents a correspondingly diverse component of lake ecosystems, yet one that is predictable even at large spatial and temporal scales.

In Chapters 2 and 3, microbiome datasets were generated from the continental-scale sampling of >300 lakes of different sizes and under different levels of land use impact in 12 Canadian ecozones. Microorganisms were sampled across a large size range (0.22–100 μm) to capture a broad phylogenetic, morphological, and lifestyle diversity of bacteria and protists. In Chapter 2, the application of genome-resolved metagenomics to ecozone co-assemblies yielded a dataset of medium- to high-quality bacterial draft genomes representing a tenfold expansion in the number and heterogeneity of lakes integrated into the global genomic catalogue. In Chapter 3, protist diversity was assessed through 18S rRNA amplicon analysis. In both studies, lake trophic state and the terrestrial influences of soils and land use were significant drivers of taxonomic or phylogenetic turnover, while differences in effect size were observed across various functional components of the microbiome. Among protist nutritional groups, mixotrophs were the least predictable, compared with phototrophs and heterotrophs. Physicochemical conditions were always the leading drivers of bacterial and protist community turnover, often followed by soil characteristics. In both studies, agriculture was the land use type driving the strongest effects on turnover, and human population density was additionally a leading predictor in bacteria. While the taxonomic diversity and functional potential of metagenome-assembled genome (MAG) assemblages each responded to land use gradients, xenobiotics metabolism encoded by MAGs was even more strongly predicted by land use. These models underscore that the lake microbiome does not function in isolation, but rather is strongly impacted by the watershed and the human activities within it at continental scale. Moreover, land use variation may implicate lake bacteria in the degradation of aromatic pollutants; however, this remains to be validated experimentally.

In Chapters 4 and 5, historical contexts to the contemporary lake microbiome were reconstructed through paleogenetic analyses of sediment DNA archives. In Chapter 4, a very basic yet technically challenging question was asked: what microbial groups inhabiting historical water columns are preserved in lake sediment DNA records? A metagenome fragment recruitment experiment applied to three lakes detected DNA from historical Cyanobacteria, bacterioplankton, viruses, and eukaryotic algae in preindustrial-age sediments. We compared what appeared to be the preferential preservation of captured

groups with the notable absence of common lineages of freshwater ultramicrobacteria. Informed by the experimental metagenome capture, bacterioplankton genera with well-characterized ecotype diversity (e.g., *Polynucleobacter* and *Limnohabitans*) were also identified. As bacterial ecotypes are well defined in terms of their ecological niches, we proposed how recovering groups with ecotypic variation from the sediment record could potentially be used to indicate past lake conditions. In total, DNA from captured surface water-associated microbes contributed <1% of sediment metagenomes, prompting future work to target preserved groups using PCR amplification strategies.

In Chapter 5, paleogenetic analyses were conducted on algal assemblages in experimentally fertilized and unmanipulated lakes with the exceptionally rare quality of having been monitored over multiple decades. Having broadened the taxonomic scope of microorganisms available to paleogenetic analysis in Chapter 4, this chapter evaluated whether sediment DNA archives reflect the temporal dynamics in the historical water column through comparison with the monitoring time series. Strong congruence was observed between algal community turnover in the monitoring and sediment records, even when considering monitoring assemblages from different parts of the water column (epilimnion or metalimnion). This finding provides validation for paleogenetic analyses of lake sediments, where very little direct comparative evidence has previously existed (Thorpe *et al.*, 2023). We furthermore evaluated the effects of eutrophication and climate warming on algal assemblages as detected in monitoring or sediment time series. While both types of records showed that nutrients were strong drivers in fertilized lakes, the 120-year perspective enabled by paleogenetics signaled the effects of rising air temperatures a decade earlier in continuously fertilized L227 than in the re-oligotrophied or unmanipulated systems. This finding aligns with the synergism between eutrophication and warming and reinforces paleogenetics as a viable method for detecting the effects of multiple stressors.

Spatial and temporal perspectives on the lake microbiome are paired in this dissertation because they mutually inform each other. The contemporary microbiome is explained through its historical context, which, in this dissertation, is reconstructed through paleogenetic analyses of sediment DNA. The chronological deposition of lake sediments over decades and centuries enables paleogenetic analysis to access past microbial assemblages along the arrow of time. Perhaps less evident than the value of informing the contemporary microbiome through its historical context is that paleogenetic techniques for reconstructing historical microbiomes rely on adequate knowledge of contemporary diversity. A prime example is that the detection of preserved lake microbes in sediment archives in Chapter 4 required the use of reference sequences representing the surface water biome. Since the completion of the MAG catalogue in Chapter 2, an expanded set of bacterial reference genomes is now available which could enhance the detection of preserved groups.

Insights into the structure and function of the lake microbiome were achieved through the application of culture-independent molecular genetic and computational techniques, many of which were developed in this dissertation. In Chapters 2 and 3, generalized dissimilarity models (GDMs) were employed to fit the relationships between compositional turnover and environmental gradients. GDMs are powerful for predicting nonlinear patterns in compositional turnover across large spatial extents (Ferrier *et*

al., 2007) and therefore were suited to the continental-scale analyses. Furthermore, GDMs represent a novel statistical approach as they have rarely been applied in microbial ecology and – with the exception of Chapter 2 – have yet to be adopted in metagenomics.

In Chapter 2, a unique approach to analyzing the patterns and drivers of microbial functional diversity was developed by investigating the metabolic potential encoded by MAG assemblages. Genome-centric metagenomics resolves the taxonomic and functional diversity in environmental samples, and the standard approach is to link taxa and functions. While we followed the standard approach to analyze the functions encoded by diverse taxa (e.g., polysaccharide utilization by Bacteroidota), our unique approach considered the totalities of metabolic function detached from their taxonomic affiliations and investigated how these responded to environmental pressures. This novel framework uncovered significant associations between environmental predictors and different components of microbial metabolism within a group of genomically defined and ecologically relevant bacteria. An important validation of this method was the comparison with the gene catalogue constructed across the metagenome assemblies.

In Chapter 3, we assigned trophic functional diversity to amplicon data, an analysis that required pulling functional annotations from the literature (Adl *et al.*, 2019). While this approach lacked direct observation, it provided a best estimate for microeukaryotic functions based on the available data. An alternative, to construct eukaryotic MAGs, engenders complex and distinct challenges (Delmont & Eren, 2016). Currently, the generation of eukaryotic MAGs may be intractable for analyzing more than a small handful of species (Kavagutti *et al.*, 2023), and functional analysis may reach another level of complexity.

Along with generating the microbial datasets, this dissertation leveraged extensive morphometric and physicochemical data for each of the lakes in the continental-scale studies. In addition, geomatics data on watershed characteristics were made available through the LakePulse Network (Huot *et al.*, 2019), allowing us to assess the terrestrial influences of watershed size and slope, soils (e.g., texture and chemistry), and the proportions of watersheds occupied by various land use types (e.g., agriculture). These analyses pointed to significant effects on the lake microbiome across the land-water interface, which for lakes is an important conduit for inputs of terrestrial carbon, nutrients, and other forms of pollution. While terrestrial effects on microbial taxonomic diversity in lakes have previously been observed (Kraemer *et al.*, 2020; Sperlea *et al.*, 2022), studies on microbiome function are rarer and Chapter 2 may represent the first demonstration of soil and land use effects at large spatial scale.

6.1 Future directions

In this dissertation, the contemporary lake microbiome was sampled across >300 lakes spread across >6.5 million km² of Canada. Encompassed by this continental-scale sampling were lakes from 12 ecozones which displayed a degree of region-specific variation in their morphometric and physicochemical properties in addition to their geographic and climatic settings. The identification of significant – and sometimes strong – environmental drivers of the lake microbiome at the continental scale describes patterns which can be generalized to a large collection of Canadian lakes. Yet, future work could aim to clarify regional-scale

patterns that may be obscured in the continental-scale analyses. For example, lakes in the Prairies are impacted by intense agricultural land use and on average have higher nutrient statuses than other ecozones (Schacht *et al.*, 2023). Investigating how the lake microbiome varies among the eutrophic lakes of the region could lead to insights into how these communities respond to other stressors.

This dissertation has opened up several areas to advance taxon-focused research on the lake microbiome. The recovery of >1,000 quality-controlled MAGs presents low-hanging fruit to clarify the ecology and evolution of bacterial species. The MAG catalogue is exceptional not only in its representation of lakes, but also because it was generated through the analysis of biomass collected across a large size range capturing microorganisms from across free-living, particle-associated, and potentially host-associated lifestyles. While Chapters 2 and 3 considered the distributions of bacteria and protists in separate studies, it would be informative to evaluate the links between bacterial and protist diversity, as these groups may even interact directly in lake food webs. The fungal diversity captured in the contemporary 18S rRNA amplicon analyses was surprisingly rich in its collection of amplicon sequence variants, and may address the emerging interest in the ecology of freshwater fungi (Grossart *et al.*, 2019). While a rich diversity of parasitic protists and fungi were identified in the amplicon analyses, no links were made to hosts. Targeted analyses of these parasites may be well timed as parasitic diseases (e.g., chytridiomycosis) are impacting freshwater species and parasites are among the species experiencing the highest biodiversity declines through the loss of their hosts. The 18S rRNA amplicon data have already been leveraged in a focused epidemiological analysis of the distributions of putative microeukaryotic pathogens across Canadian lakes (Oliva *et al.*, 2022).

In exploring the paleogenetic signals in lake sediments, it was clear that relic DNA constituted tiny fractions of larger DNA pools that can be attributed to indigenous sediment microbiota. While these microorganisms were themselves seeded from historical water columns, they survived burial and propagated in the sediment subsurface (Vuillemin *et al.*, 2023), unlike the plankton groups analyzed in this dissertation. Extant (or recently extant) sediment communities are increasingly the subjects of paleogenetic analyses, which interpret vertical variation in their taxonomic and functional composition as reflective of past lake dynamics. More broadly, further work is needed to elucidate the extant protist and fungal diversity contained in these extreme environments. Underestimating the potential of microeukaryotic lineages to survive in the sediment subsurface may confound the ability to capture clear paleogenetic signals.

Finally, the next major step of this research is to situate it within the context of the global microbiome. By making the metagenome and amplicon datasets publicly available, this may already be underway, as other researchers are now able to map their microbial data to a continental-scale sampling of lakes. Overall, this dissertation underscores the dynamic nature of the lake microbiome in both space and time. The analyses herein may one day serve as a valuable temporal snapshot of the lake microbiome as it was in 2017–2019, when freshwater stressors were already driving effects, but not as severe as in the projected future. A more joyful final note is that the study of the lake microbiome – along the diverse habitats of the water and sediment columns – is an exercise in observing microbial diversity in its many forms thriving and adapting under the myriad conditions it encounters.

References

- Adl S.M., Bass D., Lane C.E., Lukeš J., Schoch C.L., Smirnov A., *et al.* (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology* **66**, 4–119. <https://doi.org/10.1111/jeu.12691>
- Adrian R., O'Reilly C.M., Zagarese H., Baines S.B., Hessen D.O., Keller W., *et al.* (2009). Lakes as sentinels of climate change. *Limnology and Oceanography* **54**, 2283–2297. https://doi.org/10.4319/lo.2009.54.6_part_2.2283
- Aguilar P. & Sommaruga R. (2020). The balance between deterministic and stochastic processes in structuring lake bacterioplankton community over time. *Molecular Ecology* **29**, 3117–3130. <https://doi.org/10.1111/mec.15538>
- Ahmed E., Parducci L., Unneberg P., Ågren R., Schenk F., Rattray J.E., *et al.* (2018). Archaeal community changes in Lateglacial lake sediments: Evidence from ancient DNA. *Quaternary Science Reviews* **181**, 19–29. <https://doi.org/10.1016/j.quascirev.2017.11.037>
- Albert J.S., Destouni G., Duke-Sylvester S.M., Magurran A.E., Oberdorff T., Reis R.E., *et al.* (2021). Scientists' warning to humanity on the freshwater biodiversity crisis. *Ambio* **50**, 85–94. <https://doi.org/10.1007/s13280-020-01318-8>
- Albertsen M., Hugenholtz P., Skarshewski A., Nielsen K.L., Tyson G.W. & Nielsen P.H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* **31**, 533–538. <https://doi.org/10.1038/nbt.2579>
- Alishum A. (2020). DADA2 formatted 16S rRNA gene sequences for both bacteria & archaea. <https://doi.org/10.5281/ZENODO.3951383>
- Almeida A., Mitchell A.L., Boland M., Forster S.C., Gloor G.B., Tarkowska A., *et al.* (2019). A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504. <https://doi.org/10.1038/s41586-019-0965-1>
- Anantharaman K., Brown C.T., Hug L.A., Sharon I., Castelle C.J., Probst A.J., *et al.* (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications* **7**, 13219. <https://doi.org/10.1038/ncomms13219>
- Anderson N.J., Heathcote A.J. & Engstrom D.R. (2020). Anthropogenic alteration of nutrient supply increases the global freshwater carbon sink. *Science Advances* **6**, eaaw2145. <https://doi.org/10.1126/sciadv.aaw2145>
- Aramaki T., Blanc-Mathieu R., Endo H., Ohkubo K., Kanehisa M., Goto S., *et al.* (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252. <https://doi.org/10.1093/bioinformatics/btz859>
- Arora-Williams K., Olesen S.W., Scandella B.P., Delwiche K., Spencer S.J., Myers E.M., *et al.* (2018). Dynamics of microbial populations mediating biogeochemical cycling in a freshwater lake. *Microbiome* **6**, 165. <https://doi.org/10.1186/s40168-018-0556-7>

- Azam F., Fenchel T., Field J., Gray J., Meyer-Reil L. & Thingstad F. (1983). The ecological role of water-column microbes in the sea. *Marine Ecology Progress Series* **10**, 257–263.
<https://doi.org/10.3354/meps010257>
- Azan S.S.E. & Arnott S.E. (2018). The impact of calcium decline on population growth rates of crustacean zooplankton in Canadian Shield lakes. *Limnology and Oceanography* **63**, 602–616.
<https://doi.org/10.1002/lno.10653>
- Bálint M., Pfenninger M., Grossart H.-P., Taberlet P., Vellend M., Leibold M.A., *et al.* (2018). Environmental DNA time series in ecology. *Trends in Ecology & Evolution* **33**, 945–957.
<https://doi.org/10.1016/j.tree.2018.09.003>
- Barak P., Jobe B.O., Krueger A.R., Peterson L.A. & Laird D.A. (1997). Effects of long-term soil acidification due to nitrogen fertilizer inputs in Wisconsin. *Plant and Soil* **197**, 61–69.
<https://doi.org/10.1023/A:1004297607070>
- Barouillet C., Monchamp M., Bertilsson S., Brasell K., Domaizon I., Epp L.S., *et al.* (2022). Investigating the effects of anthropogenic stressors on lake biota using sedimentary DNA. *Freshwater Biology* **00**, 1–19. <https://doi.org/10.1111/fwb.14027>
- Baud A., Aulard C., Ghanbari H., Fradette M., Antoniadou D., del Giorgio P., *et al.* (2022). A framework for 210Pb model selection and its application to 37 cores from Eastern Canada to identify the dynamics and drivers of lake sedimentation rates. *Earth Surface Processes and Landforms* **47**, 2518–2530. <https://doi.org/10.1002/esp.5391>
- Baud A., Jenny J.-P., Francus P. & Gregory-Eaves I. (2021). Global acceleration of lake sediment accumulation rates associated with recent human population growth and land-use changes. *Journal of Paleolimnology* **66**, 453–467. <https://doi.org/10.1007/s10933-021-00217-6>
- Baud A., Smol J.P., Meyer-Jacob C., Paterson M., Francus P. & Gregory-Eaves I. (2023). The impacts of whole-lake acidification and eutrophication on the accumulation of lead in sediments from manipulated lakes in the Experimental Lakes Area (IELA). *Environmental Pollution* **317**, 120829. <https://doi.org/10.1016/j.envpol.2022.120829>
- Beaulieu J.J., DelSontro T. & Downing J.A. (2019). Eutrophication will increase methane emissions from lakes and impoundments during the 21st century. *Nature Communications* **10**, 1375.
<https://doi.org/10.1038/s41467-019-09100-5>
- Becker R.A., Wilks A.R., Brownrigg R., Minka T.P. & Deckmyn A. (2018). maps: Draw geographical maps
- Belle S. & Parent C. (2019). Reconstruction of past dynamics of methane-oxidizing bacteria in lake sediments using a quantitative PCR method: Connecting past environmental changes and microbial community. *Geomicrobiology Journal* **36**, 570–579. <https://doi.org/10.1080/01490451.2019.1583698>
- Berg I.A. (2011). Ecological Aspects of the Distribution of Different Autotrophic CO₂ Fixation Pathways. *Applied and Environmental Microbiology* **77**, 1925–1936. <https://doi.org/10.1128/AEM.02473-10>

- Berlemont R. & Martiny A.C. (2016). Glycoside Hydrolases across Environmental Microbial Communities. *PLOS Computational Biology* **12**, e1005300.
<https://doi.org/10.1371/journal.pcbi.1005300>
- Berninger U.-G., Caron D.A. & Sanders R.W. (1992). Mixotrophic algae in three ice-covered lakes of the Pocono Mountains, U.S.A. *Freshwater Biology* **28**, 263–272. <https://doi.org/10.1111/j.1365-2427.1992.tb00583.x>
- Bertilsson S. & Mehrshad M. (2022). Diversity and Dynamics of Bacterial Communities in Freshwater Lakes. In: *Encyclopedia of Inland Waters*, 2nd edn. pp. 601–615. Elsevier.
- 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**, 59–74. <https://doi.org/10.3897/mbmg.6.79265>
- Bjorbækmo M.F.M., Evenstad A., Røsæg L.L., Krabberød A.K. & Logares R. (2020). The planktonic protist interactome: where do we stand after a century of research? *The ISME Journal* **14**, 544–559.
<https://doi.org/10.1038/s41396-019-0542-5>
- Bock C., Jensen M., Forster D., Marks S., Nuy J., Psenner R., *et al.* (2020). Factors shaping community patterns of protists and bacteria on a European scale. *Environmental Microbiology* **22**, 2243–2260.
<https://doi.org/10.1111/1462-2920.14992>
- Bock C., Olefeld J.L., Vogt J.C., Albach D.C. & Boenigk J. (2022). Phylogenetic and functional diversity of Chrysophyceae in inland waters. *Organisms Diversity & Evolution* **22**, 327–341.
<https://doi.org/10.1007/s13127-022-00554-y>
- Boenigk J., Wodniok S., Bock C., Beisser D., Hempel C., Grossmann L., *et al.* (2018). Geographic distance and mountain ranges structure freshwater protist communities on a European scale. *Metabarcoding and Metagenomics* **2**, e21519. <https://doi.org/10.3897/mbmg.2.21519>
- Boere A.C., Sinninghe Damsté J.S., Rijpstra W.I.C., Volkman J.K. & Coolen M.J.L. (2011). Source-specific variability in post-depositional DNA preservation with potential implications for DNA based paleoecological records. *Organic Geochemistry* **42**, 1216–1225.
<https://doi.org/10.1016/j.orggeochem.2011.08.005>
- Bolger A.M., Lohse M. & Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Boundary files, 2011 census *Statistics Canada Catalogue no. 92-160-X*
- Bowers R.M., Kyrpides N.C., Stepanauskas R., Harmon-Smith M., Doud D., Reddy T.B.K., *et al.* (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725–731.
<https://doi.org/10.1038/nbt.3893>
- Brown C.T., Hug L.A., Thomas B.C., Sharon I., Castelle C.J., Singh A., *et al.* (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211.
<https://doi.org/10.1038/nature14486>

- Buck M., Garcia S.L., Fernandez L., Martin G., Martinez-Rodriguez G.A., Saarenheimo J., *et al.* (2021). Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. *Scientific Data* **8**, 131. <https://doi.org/10.1038/s41597-021-00910-1>
- Bushnell B. (2015). BBMap. <https://sourceforge.net/projects/bbmap/>
- Cabello-Yeves P.J., Zemskaya T.I., Rosselli R., Coutinho F.H., Zakharenko A.S., Blinov V. V., *et al.* (2018). Genomes of Novel Microbial Lineages Assembled from the Sub-Ice Waters of Lake Baikal. *Applied and Environmental Microbiology* **84**, 1–21. <https://doi.org/10.1128/AEM.02132-17>
- 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**, 581–583. <https://doi.org/10.1038/nmeth.3869>
- Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421. <https://doi.org/10.1186/1471-2105-10-421>
- 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*. Winnipeg.
- Canchaya C., Proux C., Fournous G., Bruttin A. & Brüßow H. (2003). Prophage Genomics. *Microbiology and Molecular Biology Reviews* **67**, 473–473. <https://doi.org/10.1128/MMBR.67.3.473.2003>
- Capo E., Debroas D., Arnaud F. & Domaizon I. (2015). Is planktonic diversity well recorded in sedimentary DNA? Toward the reconstruction of past protistan diversity. *Microbial Ecology* **70**, 865–875. <https://doi.org/10.1007/s00248-015-0627-2>
- Capo E., Debroas D., Arnaud F., Guillemot T., Bichet V., Millet L., *et al.* (2016). Long-term dynamics in microbial eukaryotes communities: a palaeolimnological view based on sedimentary DNA. *Molecular Ecology* **25**, 5925–5943. <https://doi.org/10.1111/mec.13893>
- Capo E., Debroas D., Arnaud F., Perga M.E., Chardon C. & Domaizon I. (2017). Tracking a century of changes in microbial eukaryotic diversity in lakes driven by nutrient enrichment and climate warming. *Environmental Microbiology* **19**, 2873–2892. <https://doi.org/10.1111/1462-2920.13815>
- Capo E., Giguët-Covex C., Rouillard A., Nota K., Heintzman P.D., Vuillemin A., *et al.* (2021). Lake Sedimentary DNA Research on Past Terrestrial and Aquatic Biodiversity: Overview and Recommendations. *Quaternary* **4**, 2–61. <https://doi.org/10.3390/quat4010006>
- Caporaso J.G., Lauber C.L., Walters W.A., Berg-Lyons D., Lozupone C.A., Turnbaugh P.J., *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* **108**, 4516–4522. <https://doi.org/10.1073/pnas.1000080107>
- Caron D.A., Alexander H., Allen A.E., Archibald J.M., Armbrust E.V., Bachy C., *et al.* (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology* **15**, 6–20. <https://doi.org/10.1038/nrmicro.2016.160>

- Caron D.A., Worden A.Z., Countway P.D., Demir E. & Heidelberg K.B. (2009). Protists are microbes too: A perspective. *ISME Journal* **3**, 4–12. <https://doi.org/10.1038/ismej.2008.101>
- Carpenter S., Walker B., Anderies J.M. & Abel N. (2001). From Metaphor to Measurement: Resilience of What to What? *Ecosystems* **4**, 765–781. <https://doi.org/10.1007/s10021-001-0045-9>
- Castro J.C., Rodriguez-R L.M., Harvey W.T., Weigand M.R., Hatt J.K., Carter M.Q., *et al.* (2018). imGLAD: accurate detection and quantification of target organisms in metagenomes. *PeerJ* **6**, e5882. <https://doi.org/10.7717/peerj.5882>
- Chaumeil P.-A., Mussig A.J., Hugenholtz P. & Parks D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927. <https://doi.org/10.1093/bioinformatics/btz848>
- Chen J., Bittinger K., Charlson E.S., Hoffmann C., Lewis J., Wu G.D., *et al.* (2012). Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28**, 2106–2113. <https://doi.org/10.1093/bioinformatics/bts342>
- Coenye T. (2014). The Family Burkholderiaceae. In: *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria*. (Eds E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt & F. Thompson), pp. 759–776. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Cohan F.M. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **361**, 1985–1996. <https://doi.org/10.1098/rstb.2006.1918>
- Coolen M.J.L. (2011). 7000 Years of Emiliania huxleyi Viruses in the Black Sea. *Science* **333**, 451–452. <https://doi.org/10.1126/science.1200072>
- De Corte D., Srivastava A., Koski M., Garcia J.A.L., Takaki Y., Yokokawa T., *et al.* (2018). Metagenomic insights into zooplankton-associated bacterial communities. *Environmental Microbiology* **20**, 492–505. <https://doi.org/10.1111/1462-2920.13944>
- Cottingham K.L., Rusak J.A. & Leavitt P.R. (2000). Increased ecosystem variability and reduced predictability following fertilisation: Evidence from palaeolimnology. *Ecology Letters* **3**, 340–348. <https://doi.org/10.1046/j.1461-0248.2000.00158.x>
- Coy S.R., Gann E.R., Pound H.L., Short S.M. & Wilhelm S.W. (2018). Viruses of eukaryotic algae: diversity, methods for detection, and future directions. *Viruses* **10**, 487. <https://doi.org/10.3390/v10090487>
- Davidson T.A. & Jeppesen E. (2013). The role of palaeolimnology in assessing eutrophication and its impact on lakes. *Journal of Paleolimnology* **49**, 391–410. <https://doi.org/10.1007/s10933-012-9651-0>
- Debroas D., Domaizon I., Humbert J.-F.F., Jardillier L., Lepère C., Oudart A., *et al.* (2017). Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data. *FEMS Microbiology Ecology* **93**, fix023. <https://doi.org/10.1093/femsec/fix023>

- Delgado A. & Gómez J.A. (2016). The Soil. Physical, Chemical and Biological Properties. In: *Principles of Agronomy for Sustainable Agriculture*. (Eds F.J. Villalobos & E. Fereres), pp. 15–26. Springer International Publishing, Cham.
- Delgado-Baquerizo M., Oliverio A.M., Brewer T.E., Benavent-González A., Eldridge D.J., Bardgett R.D., *et al.* (2018). A global atlas of the dominant bacteria found in soil. *Science* **359**, 320–325. <https://doi.org/10.1126/science.aap9516>
- Dillon P.J., Jeffries D.S., Snyder W., Reid R., Yan N.D., Evans D., *et al.* (1978). Acidic precipitation in south-central Ontario: recent observations. *Journal of the Fisheries Research Board of Canada* **35**, 809–815. <https://doi.org/10.1139/f78-130>
- Dixit S.S., Smol J.P., Kingston J.C. & Charles D.F. (1992). Diatoms: powerful indicators of environmental change. *Environmental Science & Technology* **26**, 22–33. <https://doi.org/10.1021/es00025a002>
- Domaizon I., Winegardner A., Capo E., Gauthier J. & Gregory-Eaves I. (2017). DNA-based methods in paleolimnology: New opportunities for investigating long-term dynamics of lacustrine biodiversity. *Journal of Paleolimnology* **58**, 1–21. <https://doi.org/10.1007/s10933-017-9958-y>
- Dornelas M., Gotelli N.J., McGill B., Shimadzu H., Moyes F., Sievers C., *et al.* (2014). Assemblage time series reveal biodiversity change but not systematic loss. *Science* **344**, 296–299. <https://doi.org/10.1126/science.1248484>
- Dray S., Bauman D., Blanchet G., Borcard D., Clappe S., Guenard G., *et al.* (2021). adespatial: Multivariate multiscale spatial analysis
- Dubois N., Saulnier-Talbot É., Mills K., Gell P., Battarbee R., Bennion H., *et al.* (2018). First human impacts and responses of aquatic systems: A review of palaeolimnological records from around the world. *The Anthropocene Review* **5**, 28–68. <https://doi.org/10.1177/2053019617740365>
- Dugan H.A., Skaff N.K., Doubek J.P., Bartlett S.L., Burke S.M., Krivak-Tetley F.E., *et al.* (2020). Lakes at Risk of Chloride Contamination. *Environmental Science & Technology* **54**, 6639–6650. <https://doi.org/10.1021/acs.est.9b07718>
- Eloe-Fadrosh E.A., Ivanova N.N., Woyke T. & Kyrpides N.C. (2016). Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology* **1**, 15032. <https://doi.org/10.1038/nmicrobiol.2015.32>
- Elser J.J., Sterner R.W., Galford A.E., Chrzanowski T.H., Findlay D.L., Mills K.H., *et al.* (2000). Pelagic C:N:P Stoichiometry in a Eutrophied Lake: Responses to a Whole-Lake Food-Web Manipulation. *Ecosystems* **3**, 293–307. <https://doi.org/10.1007/s100210000027>
- Enache M.D., Paterson A.M. & Cumming B.F. (2011). Changes in diatom assemblages since pre-industrial times in 40 reference lakes from the Experimental Lakes Area (northwestern Ontario, Canada). *Journal of Paleolimnology* **46**, 1–15. <https://doi.org/10.1007/s10933-011-9504-2>
- Ezzedine J.A., Desdevises Y. & Jacquet S. (2022). Bdellovibrio and like organisms: current understanding and knowledge gaps of the smallest cellular hunters of the microbial world. *Critical Reviews in Microbiology* **48**, 428–449. <https://doi.org/10.1080/1040841X.2021.1979464>

- Falkowski P.G., Fenchel T. & Delong E.F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science* **320**, 1034–1039. <https://doi.org/10.1126/science.1153213>
- Fernández-Gómez B., Richter M., Schöler M., Pinhassi J., Acinas S.G., González J.M., *et al.* (2013). Ecology of marine Bacteroidetes: a comparative genomics approach. *The ISME Journal* **7**, 1026–1037. <https://doi.org/10.1038/ismej.2012.169>
- 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**, 252–264. <https://doi.org/10.1111/j.1472-4642.2007.00341.x>
- Fierer N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature Reviews Microbiology* **15**, 579–590. <https://doi.org/10.1038/nrmicro.2017.87>
- Filker S., Sommaruga R., Vila I. & Stoeck T. (2016). Microbial eukaryote plankton communities of high-mountain lakes from three continents exhibit strong biogeographic patterns. *Molecular Ecology* **25**, 2286–2301. <https://doi.org/10.1111/mec.13633>
- Findlay D.L. & Kasian S.E.M. (1987). Phytoplankton community responses to nutrient addition in Lake 226, Experimental Lakes Area, northwestern Ontario. *Canadian Journal of Fisheries and Aquatic Sciences* **44**, 35–46. <https://doi.org/10.1139/f87-278>
- Findlay D.L., Kasian S.E.M., Stainton M.P., Beaty K. & Lyng M. (2001). Climatic influences on algal populations of boreal forest lakes in the Experimental Lakes Area. *Limnology and Oceanography* **46**, 1784–1793. <https://doi.org/10.4319/lo.2001.46.7.1784>
- Finn R.D., Clements J. & Eddy S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**, W29–W37. <https://doi.org/10.1093/nar/gkr367>
- Fischer R., Giebel H.-A., Hillebrand H. & Ptacnik R. (2017). Importance of mixotrophic bacterivory can be predicted by light and loss rates. *Oikos* **126**, 713–722. <https://doi.org/10.1111/oik.03539>
- Fitzpatrick M.C., Mokany K., Manion G., Lisk M., Ferrier S. & Nieto-Lugilde D. (2021). gdm: Generalized dissimilarity modeling
- Foley J.A., DeFries R., Asner G.P., Barford C., Bonan G., Carpenter S.R., *et al.* (2005). Global Consequences of Land Use. *Science* **309**, 570–574. <https://doi.org/10.1126/science.1111772>
- Fuhrman J.A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541–548. <https://doi.org/10.1038/21119>
- Gadd G.M. & Raven J.A. (2010). Geomicrobiology of eukaryotic microorganisms. *Geomicrobiology Journal* **27**, 491–519. <https://doi.org/10.1080/01490451003703006>
- Garner R., Kraemer S., Onana V., Fradette M., Varin M.-P., Huot Y., *et al.* (2023). The LakePulse Metagenome-Assembled Genome catalogue. *Dryad*. <https://doi.org/10.5061/dryad.zkh1893fs>
- Garner R.E., Gregory-Eaves I. & Walsh D.A. (2020). Sediment Metagenomes as Time Capsules of Lake Microbiomes. *mSphere* **5**, e00512-20. <https://doi.org/10.1128/mSphere.00512-20>

- 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* **7**. <https://doi.org/10.1128/msystems.00316-22>
- Garron M.-L. & Henrissat B. (2019). The continuing expansion of CAZymes and their families. *Current Opinion in Chemical Biology* **53**, 82–87. <https://doi.org/10.1016/j.cbpa.2019.08.004>
- Gast R.J., Dennett M.R. & Caron D.A. (2004). Characterization of Protistan Assemblages in the Ross Sea, Antarctica, by Denaturing Gradient Gel Electrophoresis. *Applied and Environmental Microbiology* **70**, 2028–2037. <https://doi.org/10.1128/AEM.70.4.2028-2037.2004>
- Gast R.J., Sanders R.W. & Caron D.A. (2009). Ecological strategies of protists and their symbiotic relationships with prokaryotic microbes. *Trends in Microbiology* **17**, 563–569. <https://doi.org/10.1016/j.tim.2009.09.001>
- Gauthier J., Walsh D., Selbie D.T., Bourgeois A., Griffiths K., Domaizon I., *et al.* (2021). Evaluating the congruence between DNA-based and morphological taxonomic approaches in water and sediment trap samples: Analyses of a 36-month time series from a temperate monomictic lake. *Limnology and Oceanography* **66**, 3020–3039. <https://doi.org/10.1002/lno.11856>
- Giner C.R., Balagué V., Krabberød A.K., Ferrera I., Reñé A., Garcés E., *et al.* (2019). Quantifying long-term recurrence in planktonic microbial eukaryotes. *Molecular Ecology* **28**, 923–935. <https://doi.org/10.1111/mec.14929>
- Giner C.R., Pernice M.C., Balagué V., Duarte C.M., Gasol J.M., Logares R., *et al.* (2020). Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean. *The ISME Journal* **14**, 437–449. <https://doi.org/10.1038/s41396-019-0506-9>
- Giosan L., Orsi W.D., Coolen M., Wuchter C., Dunlea A.G., Thirumalai K., *et al.* (2018). Neoglacial climate anomalies and the Harappan metamorphosis. *Climate of the Past* **14**, 1669–1686. <https://doi.org/10.5194/cp-14-1669-2018>
- Glasl B., Webster N.S. & Bourne D.G. (2017). Microbial indicators as a diagnostic tool for assessing water quality and climate stress in coral reef ecosystems. *Marine Biology* **164**, 91. <https://doi.org/10.1007/s00227-017-3097-x>
- Griffiths K., Jeziorski A., Antoniadis D., Beaulieu M., Smol J.P. & Gregory-Eaves I. (2022). Pervasive changes in algal indicators since pre-industrial times: A paleolimnological study of changes in primary production and diatom assemblages from ~200 Canadian lakes. *Science of The Total Environment* **838**, 155938. <https://doi.org/10.1016/j.scitotenv.2022.155938>
- Grossart H., Massana R., McMahon K.D. & Walsh D.A. (2020). Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnology and Oceanography* **65**, 1–19. <https://doi.org/10.1002/lno.11382>
- Grujic V., Nuy J.K., Salcher M.M., Shabarova T., Kasalicky V., Boenigk J., *et al.* (2018). Cryptophyta as major bacterivores in freshwater summer plankton. *ISME Journal* **12**, 1668–1681. <https://doi.org/10.1038/s41396-018-0057-5>

- Guillou L., Bachar D., Audic S., Bass D., Berney C., Bittner L., *et al.* (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* **41**, D597–D604. <https://doi.org/10.1093/nar/gks1160>
- Hahn M.W., Huemer A., Pitt A. & Hoetzinger M. (2021). Opening a next-generation black box: Ecological trends for hundreds of species-like taxa uncovered within a single bacterial >99% 16S rRNA operational taxonomic unit. *Molecular Ecology Resources* **21**, 2471–2485. <https://doi.org/10.1111/1755-0998.13444>
- Hahn M.W., Jezberová J., Koll U., Saueressig-Beck T. & Schmidt J. (2016a). Complete ecological isolation and cryptic diversity in Polynucleobacter bacteria not resolved by 16S rRNA gene sequences. *The ISME Journal* **10**, 1642–1655. <https://doi.org/10.1038/ismej.2015.237>
- Hahn M.W., Schmidt J., Pitt A., Taipale S.J. & Lang E. (2016b). Reclassification of four Polynucleobacter necessarius strains as representatives of Polynucleobacter asymbioticus comb. nov., Polynucleobacter duraquae sp. nov., Polynucleobacter yangtzensis sp. nov. and Polynucleobacter sinensis sp. nov., and emended des. *International Journal of Systematic and Evolutionary Microbiology* **66**, 2883–2892. <https://doi.org/10.1099/ijsem.0.001073>
- Handelsman J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669–685. <https://doi.org/10.1128/MMBR.68.4.669-685.2004>
- Hansson T.H., Grossart H., Giorgio P.A., St-Gelais N.F. & Beisner B.E. (2019). Environmental drivers of mixotrophs in boreal lakes. *Limnology and Oceanography* **64**, 1688–1705. <https://doi.org/10.1002/lno.11144>
- Harding C.J., Huwiler S.G., Somers H., Lambert C., Ray L.J., Till R., *et al.* (2020). A lysozyme with altered substrate specificity facilitates prey cell exit by the periplasmic predator Bdellovibrio bacteriovorus. *Nature Communications* **11**, 4817. <https://doi.org/10.1038/s41467-020-18139-8>
- He C., Keren R., Whittaker M.L., Farag I.F., Doudna J.A., Cate J.H.D., *et al.* (2021). Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nature Microbiology* **6**, 354–365. <https://doi.org/10.1038/s41564-020-00840-5>
- He S., Stevens S.L.R., Chan L.-K., Bertilsson S., Glavina del Rio T., Tringe S.G., *et al.* (2017). Ecophysiology of Freshwater Verrucomicrobia Inferred from Metagenome-Assembled Genomes. *mSphere* **2**, e00277-17. <https://doi.org/10.1128/mSphere.00277-17>
- Heino J., Alahuhta J., Bini L.M., Cai Y., Heiskanen A., Hellsten S., *et al.* (2021). Lakes in the era of global change: moving beyond single-lake thinking in maintaining biodiversity and ecosystem services. *Biological Reviews* **96**, 89–106. <https://doi.org/10.1111/brv.12647>
- Hengl T., Mendes de Jesus J., Heuvelink G.B.M., Ruiperez Gonzalez M., Kilibarda M., Blagotić A., *et al.* (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE* **12**, e0169748. <https://doi.org/10.1371/journal.pone.0169748>

- Henson M.W., Lanclos V.C., Faircloth B.C. & Thrash J.C. (2018). Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. *ISME Journal* **12**, 1846–1860. <https://doi.org/10.1038/s41396-018-0092-2>
- Hessen D.O., Andersen T., Armstrong McKay D., Kosten S., Meerhoff M., Pickard A., *et al.* (2023). Lake ecosystem tipping points and climate feedbacks. *Earth System Dynamics Discussions*, 1–34. <https://doi.org/10.5194/esd-2023-22>
- Hesslein R.H., Broecker W.S. & Schindler D.W. (1980). Fates of Metal Radiotracers Added to a Whole Lake: Sediment–Water Interactions. *Canadian Journal of Fisheries and Aquatic Sciences* **37**, 378–386. <https://doi.org/10.1139/f80-052>
- Higgins S.N., Paterson M.J., Hecky R.E., Schindler D.W., Venkiteswaran J.J. & Findlay D.L. (2018). Biological Nitrogen Fixation Prevents the Response of a Eutrophic Lake to Reduced Loading of Nitrogen: Evidence from a 46-Year Whole-Lake Experiment. *Ecosystems* **21**, 1088–1100. <https://doi.org/10.1007/s10021-017-0204-2>
- Hitchman R.B. & Jones H.L.J. (2000). The role of mixotrophic protists in the population dynamics of the microbial food web in a small artificial pond. *Freshwater Biology* **43**, 231–241. <https://doi.org/10.1046/j.1365-2427.2000.00541.x>
- Hug L.A., Baker B.J., Anantharaman K., Brown C.T., Probst A.J., Castelle C.J., *et al.* (2016). A new view of the tree of life. *Nature Microbiology* **1**, 16048. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Huisman J., Codd G.A., Paerl H.W., Ibelings B.W., Verspagen J.M.H. & Visser P.M. (2018). Cyanobacterial blooms. *Nature Reviews Microbiology* **16**, 471–483. <https://doi.org/10.1038/s41579-018-0040-1>
- Huntemann M., Ivanova N.N., Mavromatis K., Tripp H.J., Paez-Espino D., Tennessen K., *et al.* (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., Antoniadou D., Baulch H.M., Beisner B.E., *et al.* (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>
- Husnik F., Tashyreva D., Boscaro V., George E.E., Lukeš J. & Keeling P.J. (2021). Bacterial and archaeal symbioses with protists. *Current Biology* **31**, R862–R877. <https://doi.org/10.1016/j.cub.2021.05.049>
- Hyatt D., Chen G.-L., LoCascio P.F., Land M.L., Larimer F.W. & Hauser L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119. <https://doi.org/10.1186/1471-2105-11-119>
- Ibarbalz F.M., Henry N., Brandão M.C., Martini S., Busseni G., Byrne H., *et al.* (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**, 1084–1097. <https://doi.org/10.1016/j.cell.2019.10.008>

- Ibrahim A., Capo E., Wessels M., Martin I., Meyer A., Schleheck D., *et al.* (2021). Anthropogenic impact on the historical phytoplankton community of Lake Constance reconstructed by multimarker analysis of sediment-core environmental DNA. *Molecular Ecology* **30**, 3040–3056. <https://doi.org/10.1111/mec.15696>
- Jain C., Rodriguez-R L.M., Phillippy A.M., Konstantinidis K.T. & Aluru S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114. <https://doi.org/10.1038/s41467-018-07641-9>
- Jane S.F., Hansen G.J.A., Kraemer B.M., Leavitt P.R., Mincer J.L., North R.L., *et al.* (2021). Widespread deoxygenation of temperate lakes. *Nature* **594**, 66–70. <https://doi.org/10.1038/s41586-021-03550-y>
- Jaworski G.H.M., Talling J.F. & Heaney S.I. (2003). Potassium dependence and phytoplankton ecology: an experimental study. *Freshwater Biology* **48**, 833–840. <https://doi.org/10.1046/j.1365-2427.2003.01051.x>
- Jeppesen E., Kronvang B., Meerhoff M., Søndergaard M., Hansen K.M., Andersen H.E., *et al.* (2009). Climate change effects on runoff, catchment phosphorus loading and lake ecological state, and potential adaptations. *Journal of Environmental Quality* **38**, 1930–1941. <https://doi.org/10.2134/jeq2008.0113>
- Jeppesen E., Meerhoff M., Davidson T.A., Trolle D., Søndergaard M., Lauridsen T.L., *et al.* (2014). Climate change impacts on lakes: An integrated ecological perspective based on a multi-faceted approach, with special focus on shallow lakes. *Journal of Limnology* **73**, 88–111. <https://doi.org/10.4081/jlimnol.2014.844>
- Jones D.T., Taylor W.R. & Thornton J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8**, 275–282. <https://doi.org/10.1093/bioinformatics/8.3.275>
- Kalvari I., Nawrocki E.P., Ontiveros-Palacios N., Argasinska J., Lamkiewicz K., Marz M., *et al.* (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* **49**, D192–D200. <https://doi.org/10.1093/nar/gkaa1047>
- Kang D.D., Li F., Kirton E., Thomas A., Egan R., An H., *et al.* (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359. <https://doi.org/10.7717/peerj.7359>
- Katechakis A. & Stibor H. (2006). The mixotroph *Ochromonas tuberculata* may invade and suppress specialist phago- and phototroph plankton communities depending on nutrient conditions. *Oecologia* **148**, 692–701. <https://doi.org/10.1007/s00442-006-0413-4>
- Katoh K., Rozewicki J. & Yamada K.D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**, 1160–1166. <https://doi.org/10.1093/bib/bbx108>
- Kaushal S.S., Likens G.E., Pace M.L., Utz R.M., Haq S., Gorman J., *et al.* (2018). Freshwater salinization syndrome on a continental scale. *Proceedings of the National Academy of Sciences* **115**, E574–E583. <https://doi.org/10.1073/pnas.1711234115>

- Kaushal S.S., Mayer P.M., Vidon P.G., Smith R.M., Pennino M.J., Newcomer T.A., *et al.* (2014). Land Use and Climate Variability Amplify Carbon, Nutrient, and Contaminant Pulses: A Review with Management Implications. *JAWRA Journal of the American Water Resources Association* **50**, 585–614. <https://doi.org/10.1111/jawr.12204>
- Kavagutti V.S., Andrei A.-Ş., Mehrshad M., Salcher M.M. & Ghai R. (2019). Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. *Microbiome* **7**, 135. <https://doi.org/10.1186/s40168-019-0752-0>
- Kavagutti V.S., Bulzu P.-A., Chiriac C.M., Salcher M.M., Mukherjee I., Shabarova T., *et al.* (2023). High-resolution metagenomic reconstruction of the freshwater spring bloom. *Microbiome* **11**, 15. <https://doi.org/10.1186/s40168-022-01451-4>
- Keck F., Millet L., Debroas D., Etienne D., Galop D., Rius D., *et al.* (2020). Assessing the response of micro-eukaryotic diversity to the Great Acceleration using lake sedimentary DNA. *Nature Communications* **11**, 3831. <https://doi.org/10.1038/s41467-020-17682-8>
- Keeling P.J. & Campo J. del (2017). Marine Protists Are Not Just Big Bacteria. *Current Biology* **27**, R541–R549. <https://doi.org/10.1016/j.cub.2017.03.075>
- Kembel S.W., Cowan P.D., Helmus M.R., Cornwell W.K., Morlon H., Ackerly D.D., *et al.* (2010). Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464. <https://doi.org/10.1093/bioinformatics/btq166>
- Khomich M., Kauserud H., Logares R., Rasconi S. & Andersen T. (2017). Planktonic protistan communities in lakes along a large-scale environmental gradient. *FEMS Microbiology Ecology* **93**, fiw231. <https://doi.org/10.1093/femsec/fiw231>
- Kraemer B.M., Pilla R.M., Woolway R.I., Anneville O., Ban S., Colom-Montero W., *et al.* (2021). Climate change drives widespread shifts in lake thermal habitat. *Nature Climate Change* **11**, 521–529. <https://doi.org/10.1038/s41558-021-01060-3>
- 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**, 1–13. <https://doi.org/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**, 3011–3023. <https://doi.org/10.1038/s41396-020-0733-0>
- Kratz T.K., MacIntyre S. & Webster K.E. (2005). Causes and Consequences of Spatial Heterogeneity in Lakes. In: *Ecosystem Function in Heterogeneous Landscapes*. (Eds G.M. Lovett, C. Jones, M.G. Turner & K.C. Weathers), pp. 329–347. Springer New York, New York, NY.
- Kumar S., Stecher G., Li M., Knyaz C. & Tamura K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **35**, 1547–1549. <https://doi.org/10.1093/molbev/msy096>
- Lakes, Rivers and Glaciers in Canada - CanVec Series - Hydrographic Features (2019).

- Lammers Y., Heintzman P.D. & Alsos I.G. (2021). Environmental palaeogenomic reconstruction of an Ice Age algal population. *Communications Biology* **4**, 220. <https://doi.org/10.1038/s42003-021-01710-4>
- Lê S., Josse J. & Husson F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **25**, 253–258. <https://doi.org/10.18637/jss.v025.i01>
- Leavitt P.R. & Findlay D.L. (1994). Comparison of Fossil Pigments with 20 Years of Phytoplankton Data from Eutrophic Lake 227, Experimental Lakes Area, Ontario. *Canadian Journal of Fisheries and Aquatic Sciences* **51**, 2286–2299. <https://doi.org/10.1139/f94-232>
- Li D., Luo R., Liu C.-M., Leung C.-M., Ting H.-F., Sadakane K., *et al.* (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 H. & Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Liang Z., Soranno P.A. & Wagner T. (2020). The role of phosphorus and nitrogen on chlorophyll a: evidence from hundreds of lakes. *Water Research* **185**, 116236. <https://doi.org/10.1016/j.watres.2020.116236>
- Lima-Mendez G., Faust K., Henry N., Decelle J., Colin S., Carcillo F., *et al.* (2015). Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073. <https://doi.org/10.1126/science.1262073>
- Linz A.M., He S., Stevens S.L.R., Anantharaman K., Rohwer R.R., Malmstrom R.R., *et al.* (2018). Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *PeerJ* **6**, e6075. <https://doi.org/10.7717/peerj.6075>
- Logares R., Deutschmann I.M., Junger P.C., Giner C.R., Krabberød A.K., Schmidt T.S.B., *et al.* (2020). Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* **8**, 55. <https://doi.org/10.1186/s40168-020-00827-8>
- MacKeigan P.W., Garner R.E., Monchamp M.-È., Walsh D.A., Onana V.E., Kraemer S.A., *et al.* (2022). Comparing microscopy and DNA metabarcoding techniques for identifying cyanobacteria assemblages across hundreds of lakes. *Harmful Algae* **113**, 102187. <https://doi.org/10.1016/j.hal.2022.102187>
- Mahé F., de Vargas C., Bass D., Czech L., Stamatakis A., Lara E., *et al.* (2017). Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology & Evolution* **1**, 0091. <https://doi.org/10.1038/s41559-017-0091>
- Maileht K., Nõges T., Nõges P., Ott I., Mischke U., Carvalho L., *et al.* (2013). Water colour, phosphorus and alkalinity are the major determinants of the dominant phytoplankton species in European lakes. *Hydrobiologia* **704**, 115–126. <https://doi.org/10.1007/s10750-012-1348-x>

- Maloufi S., Catherine A., Mouillot D., Louvard C., Couté A., Bernard C., *et al.* (2016). Environmental heterogeneity among lakes promotes hyper β -diversity across phytoplankton communities. *Freshwater Biology* **61**, 633–645. <https://doi.org/10.1111/fwb.12731>
- Maniloff J. & Ackermann H.-W. (1998). Taxonomy of bacterial viruses: establishment of tailed virus genera and the other Caudovirales. *Archives of Virology* **143**, 2051–2063. <https://doi.org/10.1007/s007050050442>
- Marmen S., Blank L., Al-Ashhab A., Malik A., Ganzert L., Lalzar M., *et al.* (2020). The Role of Land Use Types and Water Chemical Properties in Structuring the Microbiomes of a Connected Lake System. *Frontiers in Microbiology* **11**, 89. <https://doi.org/10.3389/fmicb.2020.00089>
- Martin M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12. <https://doi.org/10.14806/ej.17.1.200>
- Martin S.L. & Soranno P.A. (2006). Lake landscape position: relationships to hydrologic connectivity and landscape features. *Limnology and Oceanography* **51**, 801–814. <https://doi.org/10.4319/lo.2006.51.2.0801>
- Martínez de la Escalera G., Antoniadis D., Bonilla S. & Piccini C. (2014). Application of ancient DNA to the reconstruction of past microbial assemblages and for the detection of toxic cyanobacteria in subtropical freshwater ecosystems. *Molecular Ecology* **23**, 5791–5802. <https://doi.org/10.1111/mec.12979>
- Massana R., Castresana J., Balagué V., Guillou L., Romari K., Groisillier A., *et al.* (2004). Phylogenetic and ecological analysis of novel marine Stramenopiles. *Applied and Environmental Microbiology* **70**, 3528–3534. <https://doi.org/10.1128/AEM.70.6.3528-3534.2004>
- Massana R., Gobet A., Audic S., Bass D., Bittner L., Boutte C., *et al.* (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology* **17**, 4035–4049. <https://doi.org/10.1111/1462-2920.12955>
- McKee L.S., La Rosa S.L., Westereng B., Eijsink V.G., Pope P.B. & Larsbrink J. (2021). Polysaccharide degradation by the Bacteroidetes: mechanisms and nomenclature. *Environmental Microbiology Reports* **13**, 559–581. <https://doi.org/10.1111/1758-2229.12980>
- Mehrshad M., Salcher M.M., Okazaki Y., Nakano S., Šimek K., Andrei A.-S., *et al.* (2018). Hidden in plain sight—highly abundant and diverse planktonic freshwater Chloroflexi. *Microbiome* **6**, 176. <https://doi.org/10.1186/s40168-018-0563-8>
- Mejbel H.S., Irwin C.L., Dodsworth W., Higgins S.N., Paterson M.J. & Pick F.R. (2023). Long-term cyanobacterial dynamics from lake sediment DNA in relation to experimental eutrophication, acidification and climate change. *Freshwater Biology*, 1–19. <https://doi.org/10.1111/fwb.14074>
- Melton E.D., Stief P., Behrens S., Kappler A. & Schmidt C. (2014). High spatial resolution of distribution and interconnections between Fe- and N-redox processes in profundal lake sediments. *Environmental Microbiology* **16**, 3287–3303. <https://doi.org/10.1111/1462-2920.12566>

- Messenger 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**, 13603. <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**, 1763–1778. <https://doi.org/10.1139/F08-110>
- Mitchell P. & Prepas E.E. (1990). *Atlas of Alberta lakes*. (Eds P. Mitchell & E.E. Prepas), University of Alberta Press, Edmonton, Alberta.
- Monchamp M.-E., Spaak P., Domaizon I., Dubois N., Bouffard D. & Pomati F. (2018). Homogenization of lake cyanobacterial communities over a century of climate change and eutrophication. *Nature Ecology & Evolution* **2**, 317–324. <https://doi.org/10.1038/s41559-017-0407-0>
- Monchamp M.-E., Spaak P. & Pomati F. (2019). High dispersal levels and lake warming are emergent drivers of cyanobacterial community assembly in peri-Alpine lakes. *Scientific Reports* **9**, 7366. <https://doi.org/10.1038/s41598-019-43814-2>
- Monchamp M.-E., Walser J.-C., Pomati F. & Spaak P. (2016). Sedimentary DNA Reveals Cyanobacterial Community Diversity over 200 Years in Two Perialpine Lakes. *Applied and Environmental Microbiology* **82**, 6472–6482. <https://doi.org/10.1128/AEM.02174-16>
- Monteil C.L., Vallenet D., Menguy N., Benzerara K., Barbe V., Fouteau S., *et al.* (2019). Ectosymbiotic bacteria at the origin of magnetoreception in a marine protist. *Nature Microbiology* **4**, 1088–1095. <https://doi.org/10.1038/s41564-019-0432-7>
- More K.D., Giosan L., Grice K. & Coolen M.J.L. (2019). Holocene paleodepositional changes reflected in the sedimentary microbiome of the Black Sea. *Geobiology* **17**, 436–448. <https://doi.org/10.1111/gbi.12338>
- Muñoz Sabater J. (2019). ERA5-Land hourly data from 1981 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*
- Nawrocki E.P. & Eddy S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>
- Nayfach S. & Pollard K.S. (2015). Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biology* **16**, 51. <https://doi.org/10.1186/s13059-015-0611-7>
- Nayfach S., Roux S., Seshadri R., Udworthy D., Varghese N., Schulz F., *et al.* (2021). A genomic catalog of Earth’s microbiomes. *Nature Biotechnology* **39**, 499–509. <https://doi.org/10.1038/s41587-020-0718-6>
- Nayfach S., Shi Z.J., Seshadri R., Pollard K.S. & Kyrpides N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510. <https://doi.org/10.1038/s41586-019-1058-x>
- Nearing M.A., Pruski F.F. & O’Neal M.R. (2004). Expected climate change impacts on soil erosion rates: a review. *Journal of Soil and Water Conservation* **59**, 43–50

- Neuenschwander S.M., Ghai R., Pernthaler J. & Salcher M.M. (2018). Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. *The ISME Journal* **12**, 185–198.
<https://doi.org/10.1038/ismej.2017.156>
- 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* **75**, 14–49.
<https://doi.org/10.1128/MMBR.00028-10>
- Ng T.F.F., Chen L.-F.F., Zhou Y., Shapiro B., Stiller M., Heintzman P.D., *et al.* (2014). Preservation of viral genomes in 700-y-old caribou feces from a subarctic ice patch. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 16842–16847.
<https://doi.org/10.1073/pnas.1410429111>
- Nguyen N.H., Song Z., Bates S.T., Branco S., Tedersoo L., Menke J., *et al.* (2016). FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology* **20**, 241–248. <https://doi.org/10.1016/j.funeco.2015.06.006>
- NSERC Canadian Lake Pulse Network 1 (2021). *NSERC Canadian Lake Pulse Network field manual 2017 - 2018 - 2019 surveys*. (Eds M.-P. Varin, M.-L. Beaulieu & Y. Huot), Université de Sherbrooke.
- Nurk S., Meleshko D., Korobeynikov A. & Pevzner P.A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**, 824–834. <https://doi.org/10.1101/gr.213959.116>
- Nuy J.K., Hoetzing M., Hahn M.W., Beisser D. & Boenigk J. (2020). Ecological differentiation in two major freshwater bacterial taxa along environmental gradients. *Frontiers in Microbiology* **11**, 154. <https://doi.org/10.3389/fmicb.2020.00154>
- O’Connell D.W., Ansems N., Kukkadapu R.K., Jaisi D., Orihel D.M., Cade-Menun B.J., *et al.* (2020). Changes in Sedimentary Phosphorus Burial Following Artificial Eutrophication of Lake 227, Experimental Lakes Area, Ontario, Canada. *Journal of Geophysical Research: Biogeosciences* **125**.
<https://doi.org/10.1029/2020JG005713>
- Okazaki Y., Nakano S., Toyoda A. & Tamaki H. (2022). Long-Read-Resolved, Ecosystem-Wide Exploration of Nucleotide and Structural Microdiversity of Lake Bacterioplankton Genomes. *mSystems*, 2022.03.23.485478. <https://doi.org/10.1128/msystems.00433-22>
- Oksanen J., Blanchet F.G., Kindt R., Legendre P., Minchin P.R., Hara R.B.O., *et al.* (2020). vegan: an R package for community ecologists
- 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., *et al.* (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>

- Oliverio A.M., Geisen S., Delgado-Baquerizo M., Maestre F.T., Turner B.L. & Fierer N. (2020). The global-scale distributions of soil protists and their contributions to belowground systems. *Science Advances* **6**, eaax8787. <https://doi.org/10.1126/sciadv.aax8787>
- Olm M.R., Brown C.T., Brooks B. & Banfield J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal* **11**, 2864–2868. <https://doi.org/10.1038/ismej.2017.126>
- Ontl T.A. & Schulte L.A. (2012). Soil carbon storage. *Nature Education Knowledge* **3**, 35. <https://doi.org/10.1016/c2016-0-03949-9>
- O'Reilly C.M., Sharma S., Gray D.K., Hampton S.E., Read J.S., Rowley R.J., *et al.* (2015). Rapid and highly variable warming of lake surface waters around the globe. *Geophysical Research Letters* **42**, 10773–10781. <https://doi.org/10.1002/2015GL066235>
- Orsi W.D., Coolen M.J.L., Wuchter C., He L., More K.D., Irigoien X., *et al.* (2017). Climate oscillations reflected within the microbiome of Arabian Sea sediments. *Scientific Reports* **7**, 6040. <https://doi.org/10.1038/s41598-017-05590-9>
- Ortiz-Álvarez R., Triadó-Margarit X., Camarero L., Casamayor E.O. & Catalan J. (2018). High planktonic diversity in mountain lakes contains similar contributions of autotrophic, heterotrophic and parasitic eukaryotic life forms. *Scientific Reports* **8**, 4457. <https://doi.org/10.1038/s41598-018-22835-3>
- Paquette C., Gregory-Eaves I. & Beisner B.E. (2021). Multi-scale biodiversity analyses identify the importance of continental watersheds in shaping lake zooplankton biogeography. *Journal of Biogeography* **48**, 2298–2311. <https://doi.org/10.1111/jbi.14153>
- Paradis E. & Schliep K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. <https://doi.org/10.1093/bioinformatics/bty633>
- Parducci L., Alsos I.G., Unneberg P., Pedersen M.W., Han L., Lammers Y., *et al.* (2019). Shotgun environmental DNA, pollen, and macrofossil analysis of lateglacial lake sediments from southern Sweden. *Frontiers in Ecology and Evolution* **7**, 189. <https://doi.org/10.3389/fevo.2019.00189>
- Parfrey L.W., Walters W.A., Lauber C.L., Clemente J.C., Berg-Lyons D., Teiling C., *et al.* (2014). Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Frontiers in Microbiology* **5**, 298. <https://doi.org/10.3389/fmicb.2014.00298>
- Parks D.H., Imelfort M., Skennerton C.T., Hugenholtz P. & Tyson G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**, 1043–1055. <https://doi.org/10.1101/gr.186072.114>
- Pasolli E., Asnicar F., Manara S., Zolfo M., Karcher N., Armanini F., *et al.* (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20. <https://doi.org/10.1016/j.cell.2019.01.001>
- Payne R.J. (2013). Seven reasons why protists make useful bioindicators. *Acta Protozoologica* **52**, 105–113. <https://doi.org/10.4467/16890027AP.13.0011.1108>

- Pedersen M.W., Ruter A., Schweger C., Friebe H., Staff R.A., Kjeldsen K.K., *et al.* (2016). Postglacial viability and colonization in North America's ice-free corridor. *Nature* **537**, 45–49.
<https://doi.org/10.1038/nature19085>
- Van de Peer Y. (2000). The European Small Subunit Ribosomal RNA database. *Nucleic Acids Research* **28**, 175–176. <https://doi.org/10.1093/nar/28.1.175>
- Perga M.-E., Frossard V., Jenny J.-P., Alric B., Arnaud F., Berthon V., *et al.* (2015). High-resolution paleolimnology opens new management perspectives for lakes adaptation to climate warming. *Frontiers in Ecology and Evolution* **3**, 72. <https://doi.org/10.3389/fevo.2015.00072>
- Pernthaler J. (2017). Competition and niche separation of pelagic bacteria in freshwater habitats. *Environmental Microbiology* **19**, 2133–2150. <https://doi.org/10.1111/1462-2920.13742>
- Phale P.S., Sharma A. & Gautam K. (2019). Microbial degradation of xenobiotics like aromatic pollutants from the terrestrial environments. In: *Pharmaceuticals and Personal Care Products: Waste Management and Treatment Technology*. pp. 259–278. Elsevier.
- Pilon S., Zastepa A., Taranu Z.E., Gregory-Eaves I., Racine M., Blais J.M., *et al.* (2019). Contrasting histories of microcystin-producing cyanobacteria in two temperate lakes as inferred from quantitative sediment DNA analyses. *Lake and Reservoir Management* **35**, 102–117.
<https://doi.org/10.1080/10402381.2018.1549625>
- Piwosz K., Mukherjee I., Salcher M.M., Grujić V. & Šimek K. (2021). CARD-FISH in the sequencing era: opening a new universe of protistan ecology. *Frontiers in Microbiology* **12**, 640066.
<https://doi.org/10.3389/fmicb.2021.640066>
- Piwosz K., Shabarova T., Pernthaler J., Posch T., Šimek K., Porcal P., *et al.* (2020). Bacterial and eukaryotic small-subunit amplicon data do not provide a quantitative picture of microbial communities, but they are reliable in the context of ecological interpretations. *mSphere* **5**, e00052-20.
<https://doi.org/10.1128/mSphere.00052-20>
- Poulain A.J., Aris-Brosou S., Blais J.M., Brazeau M., Keller W. & Paterson A.M. (2015). Microbial DNA records historical delivery of anthropogenic mercury. *The ISME Journal* **9**, 2541–2550.
<https://doi.org/10.1038/ismej.2015.86>
- Price M.N., Dehal P.S. & Arkin A.P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490. <https://doi.org/10.1371/journal.pone.0009490>
- Props R. & Denef V.J. (2020). Temperature and nutrient levels correspond with lineage-specific microdiversification in the ubiquitous and abundant freshwater genus *Limnohabitans*. *Applied and Environmental Microbiology* **86**, 1–17. <https://doi.org/10.1128/AEM.00140-20>
- Pruesse E., Peplies J. & Glöckner F.O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829.
<https://doi.org/10.1093/bioinformatics/bts252>

- Ptacnik R., Lepistö L., Willén E., Brettum P., Andersen T., Rekolainen S., *et al.* (2008). Quantitative responses of lake phytoplankton to eutrophication in Northern Europe. *Aquatic Ecology* **42**, 227–236. <https://doi.org/10.1007/s10452-008-9181-z>
- Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., *et al.* (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**, D590–D596. <https://doi.org/10.1093/nar/gks1219>
- Quinlan A.R. & Hall I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Ramond P., Sourisseau M., Simon N., Romac S., Schmitt S., Rigaut-Jalabert F., *et al.* (2019). Coupling between taxonomic and functional diversity in protistan coastal communities. *Environmental Microbiology* **21**, 730–749. <https://doi.org/10.1111/1462-2920.14537>
- R Core Team (2021). R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*
- Reid A.J., Carlson A.K., Creed I.F., Eliason E.J., Gell P.A., Johnson P.T.J., *et al.* (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews* **94**, 849–873. <https://doi.org/10.1111/brv.12480>
- Rocap G., Larimer F.W., Lamerdin J., Malfatti S., Chain P., Ahlgren N.A., *et al.* (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**, 1042–1047. <https://doi.org/10.1038/nature01947>
- Rodgers E.M. (2021). Adding climate change to the mix: responses of aquatic ectotherms to the combined effects of eutrophication and warming. *Biology Letters* **17**. <https://doi.org/10.1098/rsbl.2021.0442>
- Rodriguez-R L.M., Tsementzi D., Luo C. & Konstantinidis K.T. (2020). Iterative subtractive binning of freshwater chronoserries metagenomes identifies over 400 novel species and their ecologic preferences. *Environmental Microbiology* **22**, 3394–3412. <https://doi.org/10.1111/1462-2920.15112>
- 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**, 21–32. <https://doi.org/10.1111/j.1600-0587.2013.00466.x>
- Roux S., Enault F., Robin A., Ravet V., Personnic S., Theil S., *et al.* (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* **7**, e33641. <https://doi.org/10.1371/journal.pone.0033641>
- Saad J.F., Unrein F., Tribelli P.M., López N. & Izaguirre I. (2016). Influence of lake trophic conditions on the dominant mixotrophic algal assemblages. *Journal of Plankton Research* **38**, 818–829. <https://doi.org/10.1093/plankt/fbw029>
- Sagova-Mareckova M., Boenigk J., Bouchez A., Cermakova K., Chonova T., Cordier T., *et al.* (2021). Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. *Water Research* **191**, 116767. <https://doi.org/10.1016/j.watres.2020.116767>

- Saier M.H., Reddy V.S., Tsu B. V., Ahmed M.S., Li C. & Moreno-Hagelsieb G. (2016). The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Research* **44**, D372–D379. <https://doi.org/10.1093/nar/gkv1103>
- Salcher M.M., Pernthaler J. & Posch T. (2011). Seasonal bloom dynamics and ecophysiology of the freshwater sister clade of SAR11 bacteria ‘that rule the waves’ (LD12). *The ISME Journal* **5**, 1242–1252. <https://doi.org/10.1038/ismej.2011.8>
- Salcher M.M., Schaeffe D., Kaspar M., Neuenschwander S.M. & Ghai R. (2019). Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in Methylophilaceae. *The ISME Journal* **13**, 2764–2777. <https://doi.org/10.1038/s41396-019-0471-3>
- Salk K.R., Venkiteswaran J.J., Couture R., Higgins S.N., Paterson M.J. & Schiff S.L. (2022). Warming combined with experimental eutrophication intensifies lake phytoplankton blooms. *Limnology and Oceanography* **67**, 147–158. <https://doi.org/10.1002/lno.11982>
- Salmaso N. & Tolotti M. (2021). Phytoplankton and anthropogenic changes in pelagic environments. *Hydrobiologia* **848**, 251–284. <https://doi.org/10.1007/s10750-020-04323-w>
- Sánchez Schacht J.R., MacKeigan P.W., Taranu Z.E., Huot Y. & Gregory-Eaves I. (2022). Agricultural and urban land use explain substantial variation in lake water quality across national-scale physiographic and climatic gradients. *bioRxiv*, 1–46
- Schelske C.L., Peplow A., Brenner M. & Spencer C.N. (1994). Low-background gamma counting: applications for 210Pb dating of sediments. *Journal of Paleolimnology* **10**, 115–128. <https://doi.org/10.1007/BF00682508>
- Schindler D.W., Armstrong F.A.J., Holmgren S.K. & Brunskill G.J. (1971). Eutrophication of lake 227, Experimental Lakes Area, northwestern Ontario, by addition of phosphate and nitrate. *Journal of the Fisheries Research Board of Canada* **28**, 1763–1782. <https://doi.org/10.1139/f71-261>
- Schindler D.W., Hecky R.E., Findlay D.L., Stainton M.P., Parker B.R., Paterson M.J., *et al.* (2008). Eutrophication of lakes cannot be controlled by reducing nitrogen input: Results of a 37-year whole-ecosystem experiment. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 11254–11258. <https://doi.org/10.1073/pnas.0805108105>
- Schulte L., Bernhardt N., Stoof-Leichsenring K., Zimmermann H.H., Pestryakova L.A., Epp L.S., *et al.* (2021). Hybridization capture of larch (*Larix* Mill.) chloroplast genomes from sedimentary ancient DNA reveals past changes of Siberian forest. *Molecular Ecology Resources* **21**, 801–815. <https://doi.org/10.1111/1755-0998.13311>
- Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
- Sheridan E.A., Fonvielle J.A., Cottingham S., Zhang Y., Dittmar T., Aldridge D.C., *et al.* (2022). Plastic pollution fosters more microbial growth in lakes than natural organic matter. *Nature Communications* **13**, 4175. <https://doi.org/10.1038/s41467-022-31691-9>

- Sherr E.B. & Sherr B.F. (2002). Significance of predation by protists in aquatic microbial food webs. *Antonie van Leeuwenhoek* **81**, 293–308. <https://doi.org/10.1023/A:1020591307260>
- Simpson A.G.B. & Eglit Y. (2016). Protist diversification. In: *Encyclopedia of Evolutionary Biology*, Academic P. (Ed. R.M. Kliman), pp. 344–360. Elsevier.
- Singer D., Seppey C.V.W., Lentendu G., Dunthorn M., Bass D., Belbahri L., *et al.* (2021). Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. *Environment International* **146**, 106262. <https://doi.org/10.1016/j.envint.2020.106262>
- Smith V.H. & Schindler D.W. (2009). Eutrophication science: where do we go from here? *Trends in Ecology & Evolution* **24**, 201–207. <https://doi.org/10.1016/j.tree.2008.11.009>
- Smol J.P. (1988). Chrysophycean microfossils in paleolimnological studies. *Palaeogeography, Palaeoclimatology, Palaeoecology* **62**, 287–297. [https://doi.org/10.1016/0031-0182\(88\)90058-2](https://doi.org/10.1016/0031-0182(88)90058-2)
- Smol J.P. (2019). Under the radar: long-term perspectives on ecological changes in lakes. *Proceedings of the Royal Society B* **286**, 20190834. <https://doi.org/10.1098/rspb.2019.0834>
- Søndergaard M., Jensen J.P. & Jeppesen E. (2003). Role of sediment and internal loading of phosphorus in shallow lakes. *Hydrobiologia* **506–509**, 135–145. <https://doi.org/10.1023/B:HYDR.0000008611.12704.dd>
- Sperlea T., Schenk J.P., Dreßler H., Beisser D., Hattab G., Boenigk J., *et al.* (2022). The relationship between land cover and microbial community composition in European lakes. *Science of The Total Environment* **825**, 153732. <https://doi.org/10.1016/j.scitotenv.2022.153732>
- Stoecker D.K., Hansen P.J., Caron D.A. & Mitra A. (2017). Mixotrophy in the marine plankton. *Annual Review of Marine Science* **9**, 311–335. <https://doi.org/10.1146/annurev-marine-010816-060617>
- Stomp M., Huisman J., Mittelbach G.G., Litchman E. & Klausmeier C.A. (2011). Large-scale biodiversity patterns in freshwater phytoplankton. *Ecology* **92**, 2096–2107. <https://doi.org/10.1890/10-1023.1>
- Strayer D.L. & Dudgeon D. (2010). Freshwater biodiversity conservation: recent progress and future challenges. *Journal of the North American Benthological Society* **29**, 344–358. <https://doi.org/10.1899/08-171.1>
- Sunagawa S., Coelho L.P., Chaffron S., Kultima J.R., Labadie K., Salazar G., *et al.* (2015). Structure and function of the global ocean microbiome. *Science* **348**, 1261359–1261359. <https://doi.org/10.1126/science.1261359>
- Taberlet P., Coissac E., Pompanon F., Brochmann C. & Willerslev E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* **21**, 2045–2050. <https://doi.org/10.1111/j.1365-294X.2012.05470.x>
- Team Q.D. & Project O.S.G.F. (2020). QGIS Geographic Information System
- Terrapon N., Lombard V., Gilbert H.J. & Henrissat B. (2015). Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics* **31**, 647–655. <https://doi.org/10.1093/bioinformatics/btu716>

- Thorpe A., Mackay E., Goodall T., Bendle J., Thackeray S., Maberly S., *et al.* (2023). Evaluating the use of lake sedimentary DNA in palaeolimnology: A comparison with long-term microscopy-based monitoring of the phytoplankton community. *Authorea*, 1–31
- Thorpe A.C., Anderson A., Goodall T., Thackeray S.J., Maberly S.C., Bendle J.A., *et al.* (2022). Sedimentary DNA records long-term changes in a lake bacterial community in response to varying nutrient availability. *Environmental DNA*, 1–16. <https://doi.org/10.1002/edn3.344>
- Tian R., Ning D., He Z., Zhang P., Spencer S.J., Gao S., *et al.* (2020). Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* **8**, 51. <https://doi.org/10.1186/s40168-020-00825-w>
- Torti A., Jørgensen B.B. & Lever M.A. (2018). Preservation of microbial DNA in marine sediments: insights from extracellular DNA pools. *Environmental Microbiology* **20**, 4526–4542. <https://doi.org/10.1111/1462-2920.14401>
- Tran P.Q., Bachand S.C., McIntyre P.B., Kraemer B.M., Vadeboncoeur Y., Kimirei I.A., *et al.* (2021). Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *The ISME Journal* **15**, 1971–1986. <https://doi.org/10.1038/s41396-021-00898-x>
- Tranvik L.J., Downing J.A., Cotner J.B., Loiselle S.A., Striegl R.G., Ballatore T.J., *et al.* (2009). Lakes and reservoirs as regulators of carbon cycling and climate. *Limnology and Oceanography* **54**, 2298–2314. https://doi.org/10.4319/lo.2009.54.6_part_2.2298
- Tully B.J., Graham E.D. & Heidelberg J.F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data* **5**, 170203. <https://doi.org/10.1038/sdata.2017.203>
- Turner M.A., Huebert D.B., Findlay D.L., Hendzel L.L., Jansen W.A., Bodaly R. (Drew), *et al.* (2005). Divergent impacts of experimental lake-level drawdown on planktonic and benthic plant communities in a boreal forest lake. *Canadian Journal of Fisheries and Aquatic Sciences* **62**, 991–1003. <https://doi.org/10.1139/f05-003>
- Tyson G.W., Chapman J., Hugenholtz P., Allen E.E., Ram R.J., Richardson P.M., *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43. <https://doi.org/10.1038/nature02340>
- de Vargas C., Audic S., Henry N., Decelle J., Mahe F., Logares R., *et al.* (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605. <https://doi.org/10.1126/science.1261605>
- Vass M., Székely A.J., Lindström E.S. & Langenheder S. (2020). Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. *Scientific Reports* **10**, 2455. <https://doi.org/10.1038/s41598-020-59182-1>
- Vuillemin A., Horn F., Alawi M., Henny C., Wagner D., Crowe S.A., *et al.* (2017). Preservation and significance of extracellular DNA in ferruginous sediments from Lake Towuti, Indonesia. *Frontiers in Microbiology* **8**, 1440. <https://doi.org/10.3389/fmicb.2017.01440>

- Walsh E.A., Kirkpatrick J.B., Rutherford S.D., Smith D.C., Sogin M. & D'Hondt S. (2016). Bacterial diversity and community composition from seasurface to seafloor. *The ISME Journal* **10**, 979–989. <https://doi.org/10.1038/ismej.2015.175>
- Weyhenmeyer G.A., Hartmann J., Hessen D.O., Kopáček J., Hejzlar J., Jacquet S., *et al.* (2019). Widespread diminishing anthropogenic effects on calcium in freshwaters. *Scientific Reports* **9**, 10450. <https://doi.org/10.1038/s41598-019-46838-w>
- Wickham H., Averick M., Bryan J., Chang W., McGowan L., François R., *et al.* (2019). Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686. <https://doi.org/10.21105/joss.01686>
- Wight J., Varin M.-P., Robertson G.J., Huot Y. & Lang A.S. (2020). Microbiology in the Field: Construction and Validation of a Portable Incubator for Real-Time Quantification of Coliforms and Other Bacteria. *Frontiers in Public Health* **8**, 1–11. <https://doi.org/10.3389/fpubh.2020.607997>
- 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*. Ottawa.
- Wilken S., Choi C.J. & Worden A.Z. (2020). Contrasting mixotrophic lifestyles reveal different ecological niches in two closely related marine protists. *Journal of Phycology* **56**, 52–67. <https://doi.org/10.1111/jpy.12920>
- Wilken S., Soares M., Urrutia-Cordero P., Ratcovich J., Ekvall M.K., Van Donk E., *et al.* (2018). Primary producers or consumers? Increasing phytoplankton bacterivory along a gradient of lake warming and browning. *Limnology and Oceanography* **63**, S142–S155. <https://doi.org/10.1002/lno.10728>
- Willems A. (2014). The Family Comamonadaceae. In: *The Prokaryotes: Alphaproteobacteria and Betaproteobacteria*. (Eds E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt & F. Thompson), pp. 777–851. Springer Berlin Heidelberg, Berlin, Heidelberg.
- 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**, 247–254. <https://doi.org/10.1890/070140>
- Wolfe B., Kling H.J., Brunskill G.J. & Wilkinson P. (1994). Multiple Dating of a Freeze Core from Lake 227, an Experimentally Fertilized Lake with Varved Sediments. *Canadian Journal of Fisheries and Aquatic Sciences* **51**, 2274–2285. <https://doi.org/10.1139/f94-231>
- Woolway R.I., Sharma S. & Smol J.P. (2022). Lakes in Hot Water: The Impacts of a Changing Climate on Aquatic Ecosystems. *BioScience* **72**, 1050–1061. <https://doi.org/10.1093/biosci/biac052>
- Worden A.Z., Follows M.J., Giovannoni S.J., Wilken S., Zimmerman A.E. & Keeling P.J. (2015). Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594. <https://doi.org/10.1126/science.1257594>
- Wu Z., Li J., Sun Y., Peñuelas J., Huang J., Sardans J., *et al.* (2022). Imbalance of global nutrient cycles exacerbated by the greater retention of phosphorus over nitrogen in lakes. *Nature Geoscience* **15**, 464–468. <https://doi.org/10.1038/s41561-022-00958-7>

- Yu G., Smith D.K., Zhu H., Guan Y. & Lam T.T. (2017). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28–36. <https://doi.org/10.1111/2041-210X.12628>
- Zeeb B.A., Christie C.E., Smol J.P., Findlay D.L., Kling H.J. & Birks H.J.B. (1994). Responses of Diatom and Chrysophyte Assemblages in Lake 227 Sediments to Experimental Eutrophication. *Canadian Journal of Fisheries and Aquatic Sciences* **51**
- Zhang H., Yohe T., Huang L., Entwistle S., Wu P., Yang Z., *et al.* (2018). dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **46**, W95–W101. <https://doi.org/10.1093/nar/gky418>
- Ziesemer K.A., Mann A.E., Sankaranarayanan K., Schroeder H., Ozga A.T., Brandt B.W., *et al.* (2015). Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Scientific Reports* **5**, 16498. <https://doi.org/10.1038/srep16498>