Genome-resolved community structure and function of freshwater bacteria at a continental scale

Authors

Rebecca E. Garner[a,f]#, Susanne A. Kraemer[a,b,c,f,g], Vera E. Onana[a,f], Yannick Huot[d,f], David A. Walsh[a,f]#

Affiliations

[a]Department of Biology, Concordia University, Montreal, Quebec, Canada

[b]Genome Centre, Department of Microbiology & Immunology, McGill University, Montreal, Quebec, Canada

[c]Department of Civil Engineering, McGill University, Montreal, Quebec, Canada

[d]Département de géomatique appliquée, Université de Sherbrooke, Sherbrooke, Quebec, Canada

[e]Department of Biology, McGill University, Montreal, Quebec, Canada

[f]Groupe de recherche interuniversitaire en limnologie, Quebec, Canada

[g]Environment and Climate Change Canada, Canada

#Address correspondence to Rebecca E. Garner, rebecca.garner@mail.concordia.ca or David A. Walsh, david.walsh@concordia.ca.

Abstract

Lakes are highly heterogenous ecosystems inhabited by a rich microbiome whose genomic diversity remains poorly defined compared to other major biomes. Here, we present a continental-scale study of metagenomes collected across one of the most lake-rich landscapes on Earth. Analysis of 308 Canadian lakes resulted in a metagenome-assembled genome (MAG) catalogue of 1,008 bacterial genomospecies spanning a broad phylogenetic and metabolic diversity. Lake trophic state was a significant determinant of taxonomic and functional turnover of MAG assemblages. We detected a role for resource availability, particularly carbohydrate diversity, in driving biogeographic patterns. Coupling the MAG catalogue with geomatics information on watershed characteristics revealed an influence of soil properties and human land use on MAG assemblages. Agriculture and human population density were particularly influential on MAG functional turnover, signifying a detectable human footprint in lake bacterial communities. Overall, the Canadian lake MAG catalogue greatly expands the freshwater microbial genomic landscape, bringing us closer to an integrative view of bacterial genome diversity across Earth's biomes.

Introduction

Freshwater bacterial communities are a diverse component of lake ecosystems, and are central to biogeochemical cycles and the regulation of water quality (Newton *et al.*, 2011; Bertilsson & Mehrshad, 2022). Collectively, Earth's millions of lakes exhibit an immense environmental heterogeneity, reflecting a range of influences from global climate to regional variation in the terrestrial landscapes in which lakes are situated (Kratz, MacIntyre & Webster, 2005). The growing number of freshwater metagenomic studies is providing important insights into the metabolism, ecology, and evolution of lake bacteria in the context of changing environmental conditions (Grossart *et al.*, 2020). However, the majority of studies are focused on a limited number of lakes, restricting our understanding on how environmental variation influences the structure and function of bacterial communities broadly across lake ecosystems. Developing a comprehensive view of freshwater bacterial genomic diversity is critical given the myriad anthropogenic impacts on lakes (Reid *et al.*, 2019). Lake warming (O'Reilly *et al.*, 2015), eutrophication (Smith & Schindler, 2009), deoxygenation (Jane *et al.*, 2021), salinization (Dugan *et al.*, 2020), and chemical contaminant inputs are all lake stressors expected to be linked to shifts in bacterial community structure and function, but in ways that are only beginning to be deciphered.

Genome-resolved metagenomics is an approach that enables the reconstruction of composite genomes from microbial populations. From lakes studies, genome-resolved metagenomics has led to insights into the evolutionary dynamics (Cabello-Yeves *et al.*, 2018), metabolic capacities (Mehrshad *et al.*, 2018), and biogeography of freshwater bacteria (Rodriguez-R *et al.*, 2020). 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), and even across Earth's microbiomes (Nayfach *et al.*, 2021). As we continue towards a global view of microbial genome diversity, an essential next step is to expand the representation of freshwater lakes in the global genomic catalogue (Buck *et al.*, 2021).

Here we introduce a genome-resolved perspective of freshwater bacterial diversity across 6.5 million km^2 of Canada, one of the most lake-rich landscapes on Earth. The study is a component of the NSERC Canadian LakePulse Network (Huot *et al.*, 2019), which is a large-scale coordinated effort to assess the health status of lake ecosystems using standardized sampling and analysis methods with a major focus on microbial communities (Kraemer *et al.*, 2020; Garner *et al.*, 2022). Through a combination of lake metagenome co-assembly and binning, we reconstructed 1,184 high- and medium-quality metagenome-assembled genomes (MAGs) from 308 lake surface water samples. The LakePulse MAG catalogue was then used to investigate the continental-scale biogeography of lake bacteria from a community structure and function perspective. We reveal the effects of lake physicochemistry in shaping bacterial communities. We further report on factors influencing bacterial communities across the terrestrial-aquatic interface, including a detectable signal for human land use in lake microbiomes. The LakePulse MAG catalogue represents bacterial genomic diversity compiled under the largest multi lake sampling initiative and is an essential resource for investigation of freshwater, as well as global, bacterial diversity on a rapidly changing planet.

Results

The LakePulse MAG catalogue

To produce a catalogue of lake MAGs, we generated and analyzed metagenomes from the surface waters of 308 lakes in 12 Canadian ecozones (43 – 68 °N, 62 – 141 °W) (**Fig. 1**; **Fig. S1**). Lakes represented a wide range of physicochemical conditions, productivity, morphometry, climatic conditions, as well as human land use type and extent within watersheds (**Fig. S2*)***. At the continental scale, a full range of ultraoligotrophic to hypereutrophic lakes were present in the lake survey (**Fig. 1a**). Comparison of lake features by principal component analysis (PCA) revealed a large-scale spatial pattern in lake and watershed characteristics (**Fig. 1b**). Lakes within western Canada were highly heterogeneous, but were the deepest and most oligotrophic on average, and often set in watersheds with a high proportion of natural and harvested forest landscapes. Northern Canadian lakes were subject to the coldest climates and the lowest land use conversions. Central Canada is comprised of the agriculturally rich Prairies and Boreal Plains ecozones; compared to other regions of Canada, lakes of this region were generally shallow, alkaline, nutrient- and ion-rich, and highly productive. Lakes in eastern Canada had the warmest surface waters and on average the most extensive built environments in their watersheds. Agriculture

was also a common land use in eastern Canada, particularly within the Mixedwood Plains ecozone. Lakes in urban watersheds with the highest human population densities were located in the metropolitan areas of Vancouver and Toronto within the Pacific Maritime and Mixedwood Plains ecozones, respectively.

We aimed to capture a broad phylogenetic and phenotypic diversity (e.g., organism size and lifestyle strategy) of microorganisms in the LakePulse survey. To this end, metagenomes were generated from biomass collected across a wide size range (0.22 – 100 µm). From 308 metagenomes, we generated 11 metagenome co-assemblies representing the 12 ecozones (Boreal and Taiga Cordilleras lakes were combined). In total, 1,184 quality-controlled MAGs were recovered by contig binning of each co-assembly (**Table S1**). Between 28 – 233 MAGs were generated for each ecozone (**Fig. S3**). MAGs were dereplicated within and across ecozones (ANI ≥95%), resulting in a genomospecies-level set of 1,008 MAGs. MAGs were categorized as 136 high- and 872 medium-quality drafts based on MIMAG standards (Bowers *et al.*, 2017), excluding the requirement for rRNA gene presence (**Fig. 1c-d**). MAGs exhibited a wide range in size (0.35 – 10.39 Mbp), GC content (23.1 – 75.9%), and coding density (78.4 – 96.8%) (**Table S1**).

To begin mapping the biogeography of MAGs across lakes, we calculated the normalized central 80% truncated average depths (TAD_{80}) of coverage from fragment recruitment analyses (Rodriguez-R *et al.*, 2020). Based on positive TAD_{80} values, observed MAG 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 (**Fig. 1e**). A MAG accumulation curve showed that hundreds of lakes were required to adequately sample genomes across the diverse Canadian landscape using our assembly and binning approach (**Fig. 1f**). MAGs captured an average of 7.1% (range of 0.9 – 36.0%) of metagenome reads per lake (**Fig. 1g**), signifying that the MAG set is a reasonable but incomplete representation of bacterial genome diversity in the lakes. Still more MAGs remain to be recovered in unexplored sites and through a deeper assembly and binning effort of sampled metagenomes. Currently, the LakePulse MAG catalogue represents a significant expansion in the availability of freshwater genomes and represents the widest range of lake systems available to date.

Overview of phylogenetic and metabolic diversity across lakes

The LakePulse MAG catalogue represented a phylogenetically rich set of genomes (**Fig. 2a**). MAGs also exhibited high taxonomic novelty. For example, 96% of MAGs were novel candidate species (**Fig. 2b**). Over a third (349) belonged to novel genera, substantially expanding the known genomic diversity of lake bacteria. We recovered a surprising diversity of MAGs assigned to Bacteroidota, a group that plays a major role in organic particle attachment and degradation in marine settings (Fernández-Gómez *et al.*,

2013) but is much less studied in freshwater systems (Bertilsson & Mehrshad, 2022). A third of the Bacteroidota MAGs (80 of 275) represented new genera (**Table S2**), representing a rich genomic resource for future targeted research on understudied freshwater Bacteroidota ecology and metabolism, including potential symbiotic associations within the guts of zooplankton and other freshwater animals. Other taxa where genome and metabolic diversity have only recently been explored in freshwaters, but are well represented in the MAG set, include Verrucomicrobiota, Planctomycetota, Patescibacteria, and Myxococcota. Intriguingly, we identified 38 MAGs from the Bdellovibrionota, almost all of which represented new genera (**Table S2**). Bdellovibrionota are described as obligate predators of Gram-negative bacteria (Ezzedine, Desdevises & Jacquet, 2022) and their genomic diversity and role in freshwater food webs warrant further attention. Overall, these observations demonstrate that the LakePulse MAG data set represents a broad phylogenetic and lifestyle diversity which will serve as an indispensable resource for further taxon-focused research.

Freshwater bacteria contribute to lake ecosystem function through a diversity of energy-capturing and carbon cycling metabolisms that regulate biogeochemical cycles. To link bacterial taxa to functions, we explored the distributions of select metabolic modules across the MAG catalogue (**Fig. 2c**; **Fig. S4**). Oxygenic photosynthetic Cyanobacteria and anoxygenic photosynthetic Chlorobia were present, while bacteriorhodopsins were identified in the Thermoleophilia (Actinobacteriota). Photosystem II (PS-II) was the most widespread phototrophic module in the MAGs. We identified PS-II in numerous Alphaproteobacteria and Gammaproteobacteria MAGs as well as Gemmatimonadota MAGs. In combination, PS-II-encoding MAGs were identified in ~75% of lakes, indicating a diverse bacterial contribution to freshwater phototrophy. With respect to autotrophic carbon fixation, we identified three of the six known 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 was common in Proteobacteria but also phylogenetically widespread, demonstrating that non-algal carbon fixation is also common in lake ecosystems.

Heterotrophic degradation of complex algal and terrestrial organic matter is a central role of lake bacteria. Here, we found an abundance of Bacteroidota MAGs, as well as Gemmatimonadota MAGs equipped with SusCD-mediated glycan transport systems (**Figure 2c**; **Fig. S4**), suggesting a broad potential for polysaccharide transport and breakdown. Aromatic compound degradation pathways associated with the metabolism of plant-derived lignin were common within Alphaproteobacteria and Gammaproteobacteria, but also broadly distributed across the MAG phylogeny (**Figure 2c**; **Fig. S4**). In addition to heterotrophy, many Burkholderiales MAGs within Gammaproteobacteria appeared 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 were evident. Overall, these findings highlight that the

MAG catalogue captures a broad diversity of metabolisms for more in-depth comparative analysis on functional contributions to lake ecosystems, which integrate complex pools of substrates from both in-lake processes and the broader watershed.

Lake conditions shape the structure and function of bacterial assemblages

Mapping the biogeography of MAGs across Canadian lakes revealed that most MAGs occupied only a few lakes, reinforcing the importance of analyzing lake metagenomes across large geographic and environmental gradients to gain a more complete understanding for the role of local diversity in lake ecosystem function (**Fig. 3a**). Patescibacteria MAGs were among those with the most restricted range, suggesting a relatively specialist lifestyle, or perhaps input from groundwater environments where they thrive (Tian *et al.*, 2020). For most other phyla, especially the Bacteroidota, MAGs exhibited a gamut of range distributions. No MAGs were present in more than 75% of lakes, so we have little evidence for the occurrence of a "core" lake microbiome in our MAG catalogue. Notable biogeographic patterns were observed when MAGs were aggregated at the phylum level (**Fig. S5**; **Fig. S6**). Phyla previously determined to be common in lakes (e.g., Proteobacteria and Actinobacteria) were broadly distributed across the continent. But others were relatively restricted to distinct regions. For example, Firmicutes and Campylobacterota were more common in Prairies lakes of central Canada than elsewhere, perhaps in response to high nutrients or extensive agriculture within watersheds (**Fig. S5**; **Fig. S6**).

We next used the MAG catalogue to investigate biogeographic patterns across continental-scale environmental gradients from both a taxonomic and functional perspective. Taxonomic variation across lakes was visualized by principal coordinate analysis (PCoA) of Jaccard dissimilarities (**Fig. 3b**). Distributions of lake assemblages along PCoA axes 1 and 2 were both strongly related to lake trophic state (correlation with total phosphorus (TP) concentration: r_1 = -0.32, r_2 = 0.33). We suspected that the influence of trophic state may be related to functional differences among bacterial assemblages. To assess functional variation, we generated lake functional profiles by aggregating the collection of metabolic KEGG orthologs (KOs) encoded by the MAG assemblage within each lake. Bacterial taxonomic and functional dissimilarity exhibited a positive, but variable relationship with each other (**Fig. 3c**). PCoA of functional dissimilarities revealed a distribution of lake assemblages along PCoA axis 1 related to lake trophic state (correlation with TP: r_1 = 0.23). These results demonstrate that nutrient availability and overall lake productivity play a significant role in structuring the taxonomic and functional composition of bacterial assemblages at large geographic scale.

To further elucidate the factors influencing bacterial assemblage variation across lakes, we evaluated the importance of lake physicochemistry, morphology, geography, and climate conditions using Generalized Dissimilarity Models (GDM) (Ferrier *et al.*, 2007). The model based on lake physiochemistry

explained the largest amount of taxonomic and functional turnover (**Fig. 3e**; **Table S3**), reflecting the importance of environmental filtering in shaping lake communities. The influence of trophic state was evident in the physiochemistry model as taxonomic and functional turnover were similarly responsive to chlorophyll *a* (Chl a) concentration, a measure of algal biomass (**Fig. 3f**). Taxonomic turnover, but not functional turnover, was significantly related to TP (**Fig. 3f**), suggesting functional differences in bacterial assemblages are driven more by variation in organic matter composition produced by phytoplankton rather than directly by nutrient availability. Most rapid turnover was observed at the lowest end of both variables, suggesting rapid shifts in community composition across oligotrophic to mesotrophic lakes (**Fig. 3f**).

In addition to lake trophic state, the physicochemical GDM provided additional insight into the factors shaping bacterial assemblages across the continental landscape. For example, turnover was significantly associated with lake surface temperature and pH (**Fig. 3g**). Temperature and pH are often considered master variables in shaping 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 communities are influenced by a combination of the two. Remarkably, water calcium concentration was the strongest predictor of both taxonomic and functional turnover in the physiochemical GDM (**Fig. 3h**). A shift in bacterial ecotype diversity along a calcium gradient was previously reported (Hahn *et al.*, 2021), but how calcium directly influences community composition is unclear. Perhaps the effect is indirect and calcium influences ecological interaction in the plankton community, as low calcium levels were previously shown to limit zooplankton productivity and diversity (Azan & Arnott, 2018). The ecological shift in predation on phytoplankton could elicit trophic cascades through the lake food web. Further investigation of the mechanisms and implications of the relationship between calcium and bacterial communities is warranted, particularly since lake calcium decline is a legacy effect of emissions-linked acid rain (Weyhenmeyer *et al.*, 2019).

The demonstration that lake physicochemistry shapes the overall functional composition of bacterial assemblages prompted a deeper analysis focused on specific metabolic categories. To do so, we generated GDMs for different KEGG metabolic categories implicated in energy and nutrient cycling. In line with the full functional profiles, the physiochemistry model consistently explained more deviance than morphometry, geography, or climate models (**Fig 4a**). Interestingly, the amount of deviance explained varied with metabolic category, and variation within carbohydrate metabolism was the best predicted by lake physicochemistry (**Fig 4a**). Significant predictors of carbohydrate metabolism turnover included Chl a concentration and lake colour (**Fig 4b**). Lake colour reflects from the level of terrestrial organic matter input from their watersheds. 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 was the only additional metabolic category where turnover was predicted by Chl a concentration and lake colour (**Fig 4b**). Subsequent PCoAs of carbohydrate and lipid metabolisms in lakes showed strong relationships with trophic state and large amounts of variation across the eutrophic to hypereutrophic spectrum (**Fig 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 metabolic capacity to access complex carbohydrates of either aquatic or terrestrial origin depends on the initial degradation through carbohydrate-active enzymes (CAZymes). Here, we identified over 10,000 genes assigned to 129 different glycoside hydrolase (GH) gene families, which play key roles in cleaving the glycosidic bonds of diverse polysaccharides. Similar to carbohydrate metabolism (KEGG category), GH repertoires within bacterial assemblages were linked to lake physicochemistry (**Fig 4a**). Over half of GH genes were from Bacteroidota MAGs (**Fig. 4e**) and were often encoded in polysaccharide utilization loci (PULs) (**Fig. S7**), demonstrating that freshwater Bacteroidota are central to complex organic matter degradation comparable to their role in marine and human gut ecosystems (McKee *et al.*, 2021). Verrucomicrobiota MAGs were also replete with GH genes, 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 bacterial assemblages (**Fig. 4f**). Bacteroidota displayed the most distinct GH family repertoires, characterized by an enrichment in GH74 implicated in cellulose and xyloglucan degradation and GH20 implicated in anhydromuropeptides recycling and chitin degradation. Proteobacteria and Bdellovibrionota displayed overlapping GH repertoires distinct from other taxonomic groups and driven by GH23, GH102, and GH103, which are all lysozymes implicated in peptidoglycan degradation, in keeping with potential predatory activity against other bacteria (Harding *et al.*, 2020). The thousands of GH genes encoded in the LakePulse MAG catalogue not only represent a rich resource to develop a deeper understanding of organic matter degradation in freshwaters, but also to expand on previous global explorations of CAZyme diversity across ecosystems (Berlemont & Martiny, 2016; Garron & Henrissat, 2019).

Linking watershed soil and human land use characteristics to bacterial assemblages

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 geomatic descriptions of lake watersheds, including soil properties and human 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

communities across the terrestrial-aquatic interface. Soil properties explained a significant amount of GDM deviance for both taxonomic (12.1 %) and functional (16.3 %) turnover in bacterial assemblages (**Fig 5a**). In fact, of all GDMs developed in this study, the importance of soil was second only to lake physicochemistry, providing evidence for a significant terrestrial influence on lake bacterial assemblage structure and function. Turnover was most strongly associated with two soil characteristics: pH and the volumetric fraction of coarse fragments, a variable linked to soil texture (**Fig. 5b**). Soil texture governs water retention in surface soil, while pH determines the mobility of ions and nutrient input into lakes (Delgado & Gómez, 2016). Hence the degree of terrestrial input from soils appears to influence lake bacterial diversity in a moderately predictable manner at continental scale.

Next, we investigated if bacterial assemblages could be predicted by human land use. We were particularly interested in how agriculture (**Fig. 5c**) and human population density (**Fig 5d**) within watersheds may shape bacterial diversity, as both are recognized as influential on lake ecosystems through eutrophication and run off of chemical contaminants. GDMs based on land use explained a lesser, but still significant, amount of taxonomic (6.8 %) and functional turnover (5.4 %) in bacterial assemblages compared to soil properties. Of all human impact variables, agriculture exhibited a significant influence on both taxonomic and functional turnover, while human population density exhibited a significant effect on functional turnover (**Fig. 5e**). Intriguingly, when analyzing specific metabolic categories, xenobiotic metabolism was the one most strongly explained by a land use model (**Fig. 5e**). Consistent with general functional turnover (**Fig. 5f**), human population density followed by agriculture were the strongest variables explaining turnover within xenobiotic metabolism (**Fig. 5g**). In both cases, the most rapid turnover was at the high end of population density (1,000 – 3,000 individuals per km^2), which corresponds to the few lakes sampled within the highly urbanized regions (Vancouver and Toronto metropolitan areas) in the Pacific Maritimes and Mixedwood Plains ecozones (**Fig 5d**). The underlying effect of population density appeared complex; bacterial assemblages within highly populated landscapes were in fact characterized by a lower diversity of KOs linked to xenobiotic metabolism. Moreover, variation in community xenobiotic metabolism analyzed using PCoA showed that although assemblages in highly populated lakes were distinct from low population lakes, they were also largely dissimilar from each other (**Fig. S8)**. Many of the KOs within xenobiotic metabolism were involved in the degradation of aromatic compounds. Detecting this xenobiotic degradation capacity presents the possibility that lake bacteria are further capable of metabolising a wider suite of aromatic compounds including those considered environmental pollutants, such as agricultural pesticides or industrial wastes. Overall, these results suggest human population density and associated ecological influences have a detectable and significant influence on the functional diversity encoded in the LakePulse MAG catalogue. The detection of these potentially anthropogenic effects is notable considering they represent evidence of land-based pollution pervading the base of freshwater food webs.

Discussion

In constructing a MAG catalogue from metagenomes sampled from hundreds of lakes at a continental scale, this study has substantially increased the number and diversity of bacterial genomes from freshwater ecosystems. The restricted geographic distribution of individual MAGs reflected the immense environmental heterogeneity among lakes and demonstrates the value of sampling a large number of lakes using a standardized approach. Trophic state emerged as one of the most important lake characteristics driving the taxonomic and functional turnover of bacterial communities at large spatial scale and has implications in understanding bacterial responses to ongoing anthropogenic eutrophication of lakes. The detected effects of watershed characteristics, including human land use, on the functional composition of freshwater bacteria at the community level opens the door to more in-depth analyses into the specific metabolic pathways responsive across the terrestrial-aquatic interface. Future research will include in-depth MAG analysis of uncultivated groups to further clarify their ecologies and evolutionary relationships to bacteria in other biomes. Additional insights into lake microbiomes, especially the identification of taxonomic or functional genes that may serve as indicators of environmental change, will be gained by exploring the wealth of metagenomic information beyond discrete genomes.

Methods

Lake selection and sampling

Lake sampling was conducted by the Natural Sciences and Engineering Research Council (NSERC) Canadian Lake Pulse Network between 2017 – 2019 (Huot *et al.*, 2019). Sampling was conducted during the period of lake thermal stratification, where relevant, in July to September. To capture the natural and human-mediated heterogeneity in lake and watershed conditions, lakes were selected randomly across three lake area categories (small [0.1, 0.5) km^2, medium [0.5, 5), and large [5, 100]) 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 on-shore 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 in Hg. Up to 500 mL of water was filtered or until the filter nearly clogged. All sampling equipment was acid-washed and rinsed with lake water prior to use. Filters were stored in sterile cryovials at -80 °C.

Environmental data

Six categories of environmental explanatory variables were selected for analysis: (1) geography, (2) lake morphometry, (3) surface water physicochemical factors, (4) watershed surface soil properties, (5) land use, and (6) climate. Geography variables included latitude, longitude, and altitude. Lake morphometry and physical characteristics 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. Surface water physicochemical factors included surface water temperature, pH, colour, and concentrations of Chl a, DIC, DOC, TN, TP, calcium, chloride, magnesium, potassium, sodium, and sulfate. Watershed soil properties estimated for the top 0-5 cm soil depth interval included bulk density of the fine earth fraction, cation exchange capacity, volumetric fraction of coarse fragments, proportion of clay particles in the fine earth fraction, total nitrogen, soil pH, proportion of sand particles in the fine earth fraction, proportion of silt particles in the fine earth fraction, soil organic carbon content in the fine earth fraction, and organic carbon density. Land use variables were calculated as fractions of watershed area not covered by water and included crop agriculture, pasture, forestry, built development, human population density, livestock density, and poultry density. Climate variables measured over the seven days prior to lake sampling included mean air temperature, total precipitations, mean net solar radiation, mean wind speed, and ice disappearance day for the year of sampling. Lake trophic state was categorized by TP concentrations according to the Canadian Water Quality guidelines: ultraoligotrophic (TP <4 µg/L), oligotrophic (4 – 10 µg/L), mesotrophic (10 – 20 µg/L), mesoeutrophic (20 – 35 µg/L), eutrophic (35 – 100 µg/L), and hypereutrophic (>100 µg/L) (Canadian Council of Ministers of the Environment, 2004).

DNA extraction and metagenome sequencing

DNA was extracted from filters using the DNeasy PowerWater kit (QIAGEN) according to the manufacturer's instructions supplemented by optional steps (i.e. 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.

Metagenome assembly and binning

Adapter-clipping and quality-trimming of raw reads were performed in Trimmomatic v. 0.38 using default settings (Bolger, Lohse & Usadel, 2014). We used an assembly approach to maximize the recovery of genomes from the environment, but given the constraints of the immense scale of data and complexity of

the assembly, we generated combined-assemblies from metagenomes common to the same ecozone. Metagenomes were co-assembled in MEGAHIT v. 1.2.7 using kmer lengths 27, 37, 47, 57, 67, 77, 87 and a minimum count of two (Li *et al.*, 2016). The Boreal/Taiga Cordilleras co-assembly was exceptionally generated from the combination of metagenomes from two ecozones.

Binning was performed in MetaBAT v. 2.12.1 using the default minimum scaffold length of 2,500 bp and based on the differential coverage of each contig determined by the jgi_summarize_bam_contig_depths script (Kang *et al.*, 2019). Bin completeness, contamination, and strain heterogeneity were estimated in CheckM v. 1.0.7 based on single-copy marker genes (Parks *et al.*, 2015). Bins scoring ≥50% completeness, <10% contamination, and <10% strain heterogeneity were retained as MAGs. MAGs were dereplicated at ANI ≥95% and alignment coverage ≥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 subsequent analysis. MAGs were categorized according to MIMAG standards by which high-quality drafts display completeness >90%, contamination <5%, and ≥18/20 tRNA genes (16S, 23S, and 5S rRNA gene presence criteria were omitted) and medium-quality drafts display completeness ≥50% and contamination <10% (Bowers *et al.*, 2017).

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). Ribosomal 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 (Aramaki *et al.*, 2020). 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 composition analysis.

Functions were assigned based on 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, K08943). Bacteriorhodopsin assignment required evidence of bop (K04641) only. Autotrophic carbon fixation via three independent pathways was evidenced by rbcL (K01601) only for the CBB cycle, mct (K14470) only for the glyoxylate assimilation pathway, and abfD (K14534) only 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 funneling or ring fission pathway: evidence of catabolic funneling was identified by ferulate to vanillin conversion (evidenced by fcs (K12508) only), vanillin to protocatechuate conversion (evidenced by vanA (K03862) only), or protocatechuate ring opening (evidenced by pcaG (K00448) OR ligA (K04100)); ring fission was identified by gallate ring opening (evidenced by galA (K04099) only), catechol ring opening (evidenced by K00446 OR K03381), salicylate to gentisate (evidenced by K00480 OR K18242 OR K18243), or gentisate ring opening (evidenced by K00450 only). Methanol oxidation was evidenced by xoxF (K23995) only. Nitrogen fixation was evidenced by nifD (K02586) OR nifH (K02588) OR nifK (K02591). Dissimilatory nitrate reduction was evidenced by (narG AND narH) OR (napA AND napB) (narG = K00370, narH = K00371, napA = K02567, 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) only, or 4) nirS (K15864) only. Nitric oxide was evidenced by norB (K04561) AND norC (K02305). Nitrous oxide was evidenced by nosZ (K00376) only. Evidence of sulfur oxidation via SOX was determined by the presence of 1) soxA (K17222) AND soxB (K17224) AND soxX (K17223) AND 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 (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 by Terrapon *et al*. (2015) so as 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.

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) prior to 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 BBMap 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 of Rodriguez-R *et al.* (2020) as the average sequencing depth truncated to the central 80% of mapped positions normalized by the number of genome equivalents. Genome equivalents were estimated in MicrobeCensus v. 1.1.0 (Nayfach & Pollard, 2015).

Taxonomic classification and phylogenetic analysis

Taxonomic classification was performed in GTDB-Tk v. 1.3.0 by phylogenetic placement onto the bacterial reference tree from reference data 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 JTT model of evolution and subsequently plotted with the R package ggtree (Yu *et al.*, 2017).

Ecological analysis

Biogeographical analysis was conducted on MAG assemblages in 300 freshwater and oligosaline lakes, identified as having conductivity <8 mS/cm and total major ions <4,000 mg/L. To build community function matrices, we assumed that a function was implied when an encoding MAG was detected at a site. PCoA was performed on pairwise Jaccard dissimilarities between pairwise presence-absence community response data using the pcoa() function in the R package ape v. 5.4 (Paradis & Schliep, 2019). GDMs fitting community 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). LCBD analysis was performed on Jaccard dissimilarities using 999 permutations in the R package adespatial v. 0.3.8 (Dray *et al.*, 2021). PCA was performed on Hellinger-transformed data using the rda() function in the R package vegan v. 2.5.6 (Oksanen *et al.*, 2020). Data wrangling 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).

Data availability

Raw metagenome reads were archived in the European Nucleotide Archive under study accession PRJEB29238 (https://www.ebi.ac.uk/). Metagenome co-assemblies were deposited and annotated at the Joint Genome Institute Genomes OnLine Database under study accession Gs0136026. For peer-review purposes, the MAG data has been included as supplementary data. The data include the genomic scaffolds (lpmags_supplement-review1.zip), protein-coding sequences (lpmags_supplement-review2.zip), and annotations as general feature format files (lpmags_supplement-review3.zip).

Acknowledgements

This research was funded by the NSERC Canadian Lake Pulse Network (Strategic Network Grant NETGP-479720) and Canada Research Chairs held by DAW and YH. REG and VEO were supported by the NSERC CREATE ÉcoLac training program in lake and fluvial ecology. REG 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 thank the researchers in LakePulse who contributed to the data set. We also thank Wentworth Brookes, Compute Canada, and Shawn Simpson for their technical support in bioinformatic analyses.

References

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

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

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

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

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.

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

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

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

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.

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

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

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.

Dray S., Bauman D., Blanchet G., Borcard D., Clappe S., Guenard G., *et al.* (2021). adespatial: Multivariate multiscale spatial analysis

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

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

Fernández-Gómez B., Richter M., Schüler 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

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

Fitzpatrick M.C., Mokany K., Manion G., Lisk M., Ferrier S. & Nieto-Lugilde D. (2021). gdm: Generalized dissimilarity modeling

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*. 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

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

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

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 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

Huot Y., Brown C.A., Potvin G., Antoniades 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

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

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

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

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.

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., 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

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

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., Udwary 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

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

NSERC Canadian Lake Pulse Network (2021). *NSERC Canadian Lake Pulse Network field manual 2017 - 2018 - 2019 surveys*. (Eds M.-P. Varin, M.-L. Beaulieu & Y. Huot), Université de Sherbrooke.

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

Oksanen J., Blanchet F.G., Kindt R., Legen- P., Minchin P.R., Hara R.B.O., *et al.* (2020). vegan:

Community ecology package

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

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

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

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

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

Rodriguez-R L.M., Tsementzi D., Luo C. & Konstantinidis K.T. (2020). Iterative subtractive binning of freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic preferences. *Environmental Microbiology* **22**, 3394–3412. https://doi.org/10.1111/1462-2920.15112

Seemann T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153

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

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

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

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

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

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

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

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

Figure legends

Figure 1. Geographical and environmental scope of the LakePulse MAG catalogue, its 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)** Principal component analysis illustrating the variation in environmental features among sampled lakes. The letter symbols and colours of each point represent ecozones and geographic regions of Canada, respectively. Arrows show the contributions of environmental variables to PC loadings. Letters designating ecozones are presented in the legend key of panel A. **(C)** Proportions of MAGs categorized as high- or medium-quality drafts. **(D)** Distributions of MAG quality characteristics. **(E)** MAG richness within lakes. Point colours represent richness, which is the number of MAGs detected at each site based on positive TAD80 values. Ecozone colours shown on the base map are described in the legend key of panel a. **(F)** MAG accumulation curve (i.e., the number of new genomes detected) in a random ordering of lakes. **(G)** Fraction of LakePulse metagenomes captured within MAGs (measured as the percentage of each metagenome recruited to the MAG set). Point colours represent lake trophic states and letters represent ecozones (see legend of panel a).

Figure 2. Taxonomic, phylogenetic, and functional diversity of LakePulse MAGs. **(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. Bar plots are divided by colour to show the number of MAGs assigned to unclassified (i.e., novel) taxa from class to species level. **(C)** Functional diversity across taxonomic phyla or orders. The lower heat map shows the number of MAGs within each order containing genetic marker for a selection of metabolic functions (grey = no evidence). The upper heat map shows the number of lakes containing MAGs with evidence of metabolic functions.

Figure 3. Niche breadth and environmental filtering of MAGs based on taxonomic and functional biogeography. **(A)** Patterns of MAG incidence (i.e., the number of lakes occupied by each MAG). Colours show MAG phylum-level assignments. **(B)** Principal coordinate analysis (PCoA) showing the variation in MAG taxonomic composition among lakes based on Jaccard distance. 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 distances. Point colours and letter symbols are described in the legend of panel b. **(E)** Percent deviance explained by generalized dissimilarity models (GDMs) fitting the responses of taxonomic or functional assemblages to physicochemical, morphometric, geographic, or climatic gradients. (**F-H)** Partial effects of **(F)** total phosphorus (TP) and chlorophyll *a* (Chl a), **(G)** temperature and pH, and **(H)** calcium and lake colour on the taxonomic and functional turnover of MAG assemblages across lakes. Curve shapes represent the rate of turnover and the maximum heights of curves represent the magnitude of turnover across environmental gradients. Line colours represent different community response data (red = taxonomic assemblages, blue = functional potential) and line types (solid or dashed) represent different environmental explanatory variables (e.g., panel f, TP vs. chlorophyll *a*).

Figure 4. Diversity of specific bacterial functions among lake assemblages and individual MAGs. **(A)** Percent deviance explained by generalized dissimilarity models (GDMs) fitting the responses of specific community functions to physicochemical, morphometric, geographic, or climatic gradients. **(B)** Partial effects of chlorophyll *a* and lake colour on the carbohydrate and lipid metabolism turnover of MAG assemblages across lakes. **(C)** Principal coordinate analysis (PCoA) showing the variation in carbohydrate metabolism potential among MAG assemblages. **(D)**, Principal coordinate analysis (PCoA) showing the variation in lipid metabolism potential among MAG assemblages. Point colours in **C** and **D** represent lake trophic state and letter symbols represent ecozones. **(E)** Percentage and number of glycoside hydrolase (GH) genes distributed across MAG phyla. **(F)** Principal component analysis (PCA) showing the variation in GH gene families among MAGs. Point colours represent MAG phyla and letter symbols represent order-level taxonomy within groups.

Figure 5. Influences watershed characteristics and human land use on lake bacterial assemblages. **(A)** Percent deviance explained by generalized dissimilarity models (GDMs) fitting the responses of taxonomic or functional assemblages to watershed soil characteristics. **(B)** Partial effects of soil chemistry (pH) and particle composition (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 generalized dissimilarity models (GDMs) fitting the responses of taxonomic or functional assemblages to watershed

land use characteristics. **(F)** Partial effects of land use (crop agriculture and human population density) on the taxonomic and functional turnover of MAG assemblages across lakes. **(G)** Partial effects of land use (crop agriculture and human population density) on the xenobiotics biodegradation and metabolism potential of MAG assemblages across lakes.

Supplementary information

Figure S1. Project workflow illustrating the generation and analysis of the MAG data set.

Figure S2. Distributions of **a**, surface water physicochemistry, **b**, lake morphometry, **c**, climate, **d**, watershed land use, **e**, watershed surface soil, and **f**, geography variables. Colours represent ecozones and ecozone medians are indicated with dotted lines.

Figure S3. Number of MAGs generated within each ecozone co-assembly.

Figure S4. Functional diversity at the level of individual MAGs.

Figure S5. Phylum-level biogeographic distributions of MAGs.

Figure S6. Hierarchical clustering (Ward's linkage) of phylum-level taxonomic assemblages.

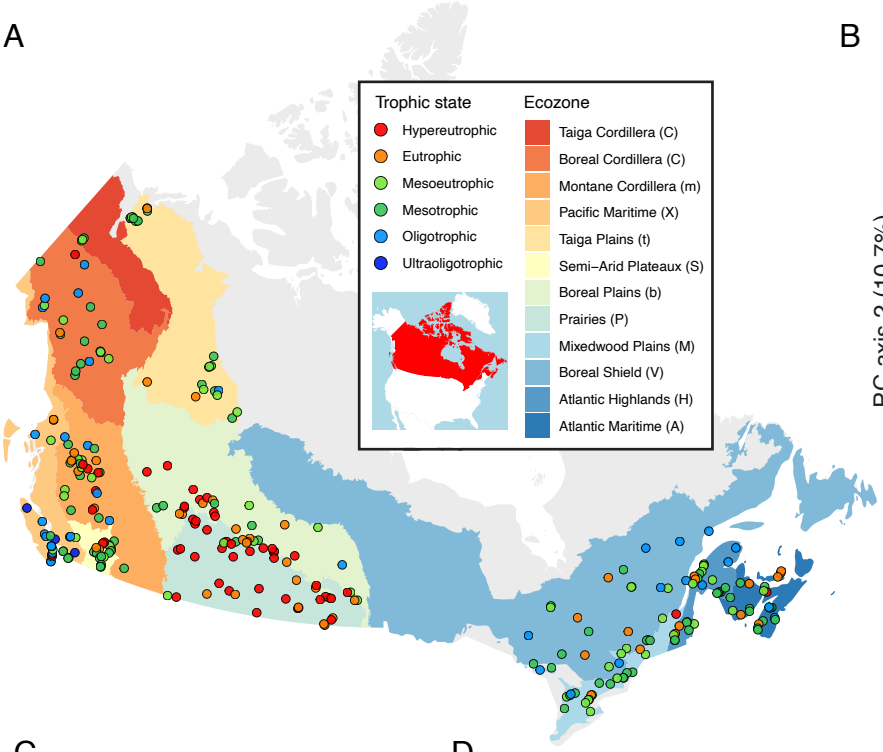
Figure S7. Gene maps of polysaccharide utilization loci (PULs) identified across Bacteroidota MAGs. Arrows indicate gene directions. Colours represent gene types (SusCD, CAZymes, tRNA genes, other KOs).

Figure S8. Principal coordinate analysis (PCoA) showing the variation in xenobiotic metabolism potential among MAG assemblages.

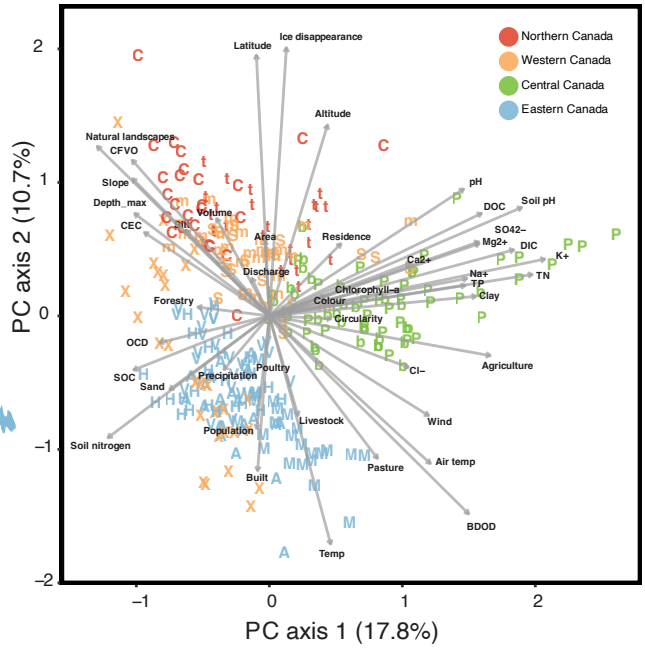
Table S1. Summary of MAG characteristics including MAG quality, genome characteristics, taxonomy, and associated file names.

Table S2. Number of novel MAGs within each phylum.

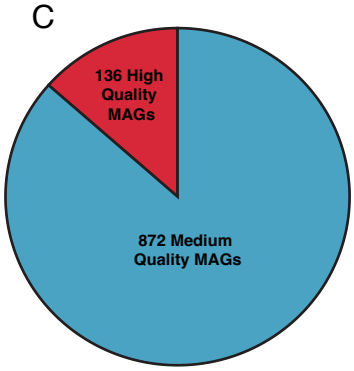
A



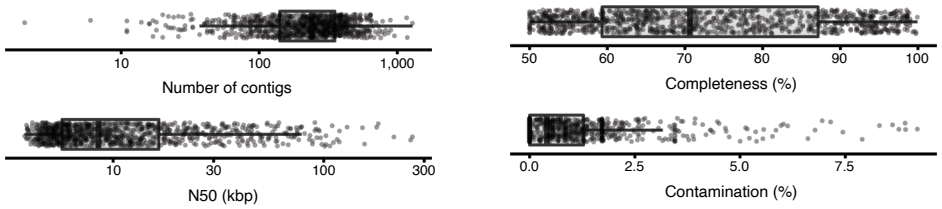
B



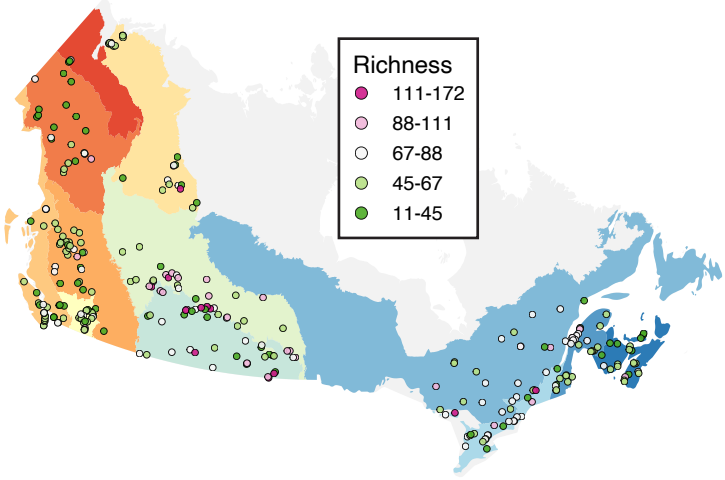
C



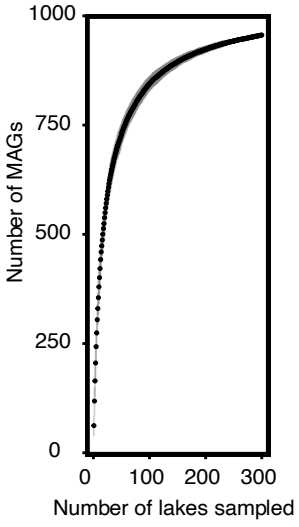
D



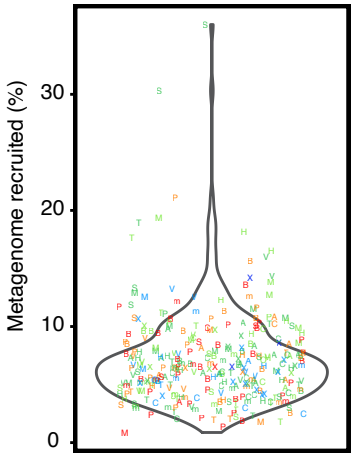
E

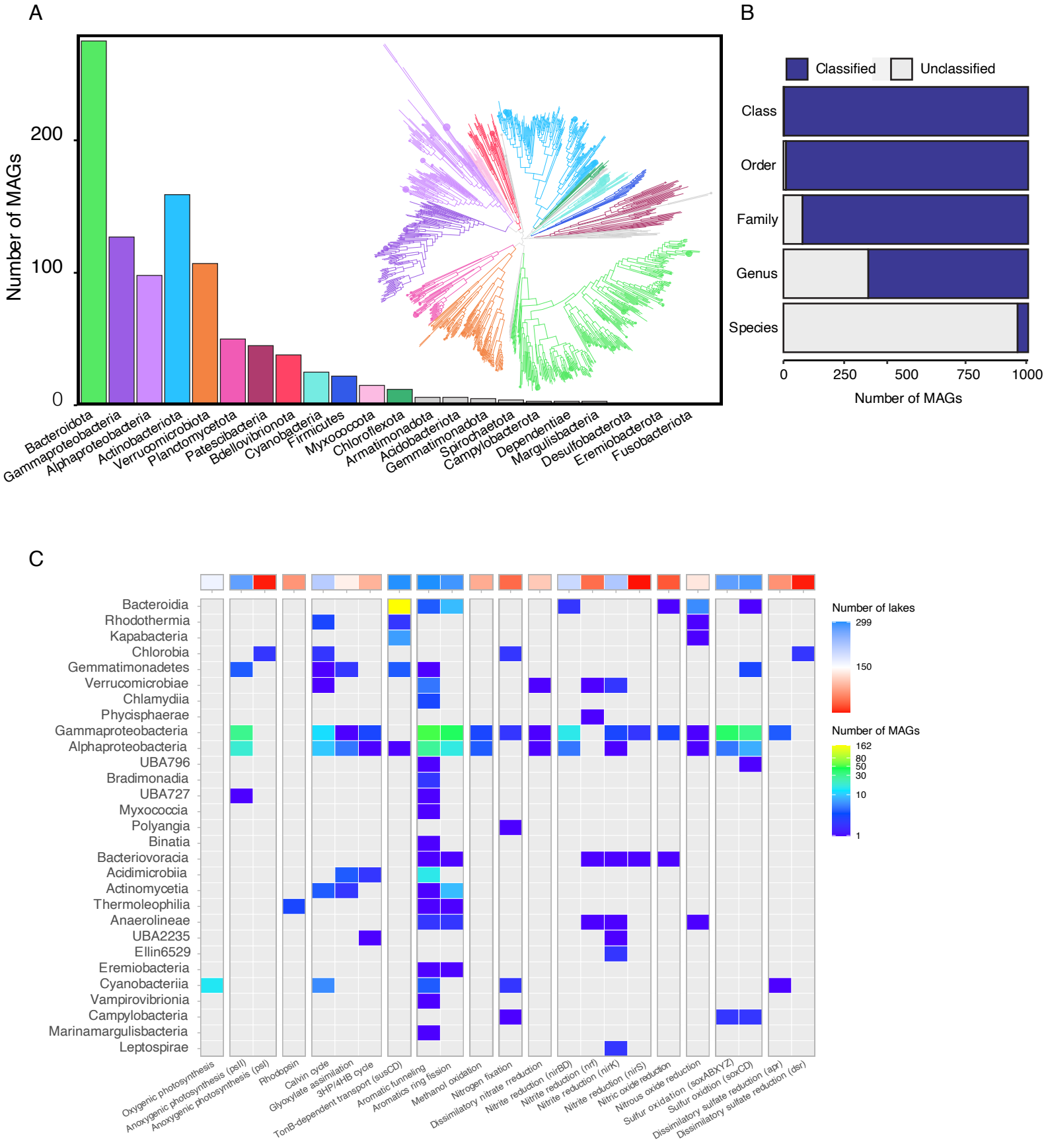


F

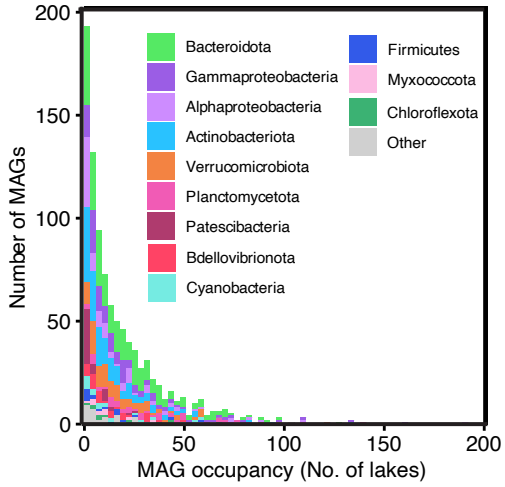


G

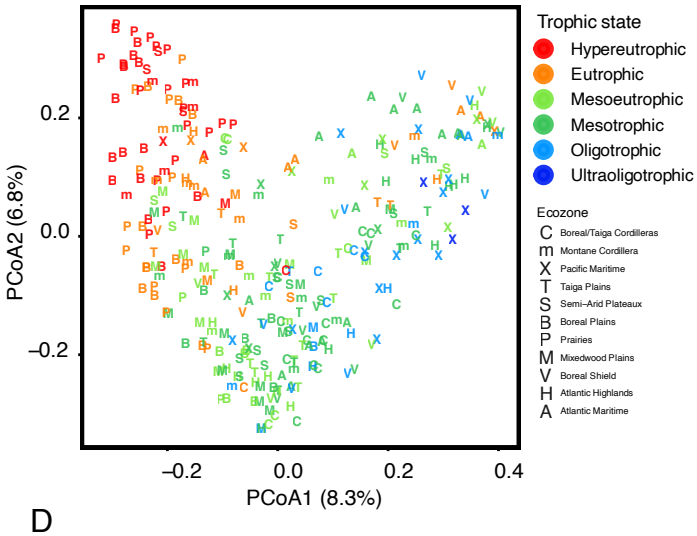




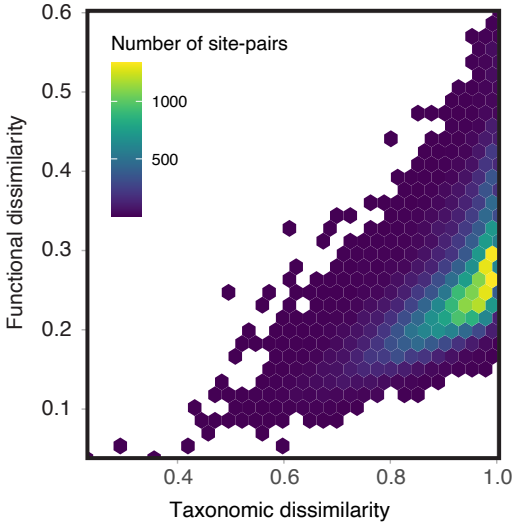
A



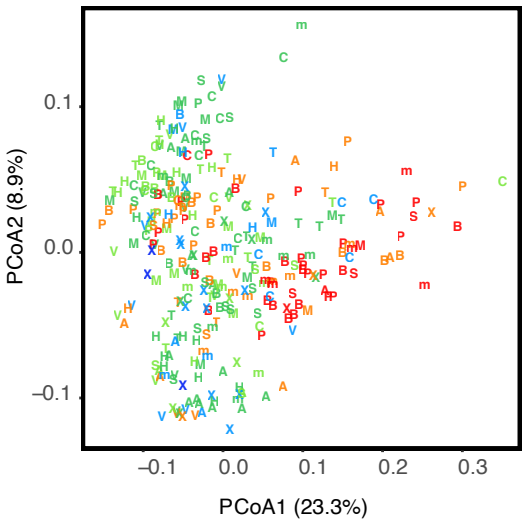
B



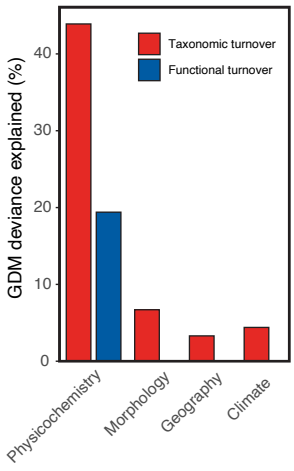
C



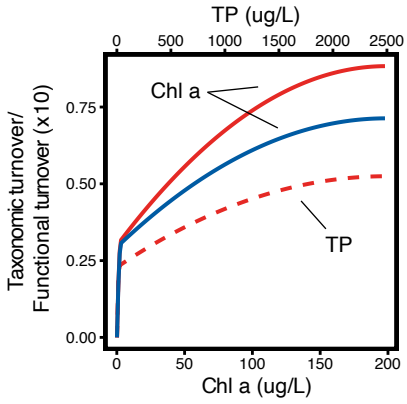
D



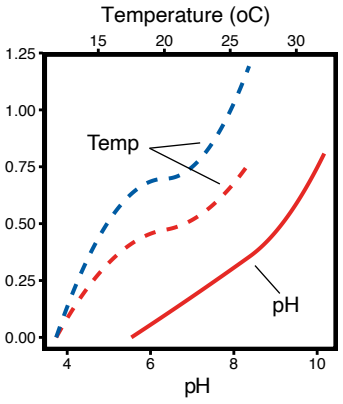
E



F



G



H