A genome catalogue of lake bacteria across watershed land use gradients at a continental scale

Authors

Rebecca E. Garner[1,4], Susanne A. Kraemer[2,4], Vera E. Onana[1,4], Maxime Fradette[3], Marie-Pierre Varin[3], Yannick Huot[3,4], David A. Walsh[1,4]#

Affiliations

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

[2]Environment and Climate Change Canada, Montreal, Quebec, Canada

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

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

#Address correspondence to David A. Walsh, david.walsh@concordia.ca.

Abstract

Lakes are heterogenous ecosystems inhabited by a rich microbiome whose genomic diversity is poorly defined. We present a continental-scale study of metagenomes representing 6.5 million km^2 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.

Main

Freshwater bacteria are a diverse component of lake ecosystems and are central to biogeochemical cycles and water quality[1–3]. 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 runoff[4]. 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[5] and sentinels of environmental change[6]. 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[7]. However, most studies are focused on a limited number of lakes, restricting our understanding on 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[8], eutrophication[9], deoxygenation[10], salinization[11], chemical contamination, and other emergent stressors[12] are expected to modify the freshwater microbiome, but in ways that we are only beginning to understand[13].

Genome-resolved metagenomics enables the reconstruction of composite genomes from microbial populations[14]. MAGs supply reference genomes for uncultivated taxa, revealing enormous microbial dark matter[15]. 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[16,17], metabolism[18–20], and succession[21,22] of freshwater bacteria. Recently, the approach has been applied at larger scale to generate genomic catalogues from marine[23], terrestrial[24], and host-

associated microbiomes[25–27], culminating in a genome catalogue of Earth's microbiomes (GEM)[28]. Although impressive in its collection of habitats, GEM encompasses very few freshwater lakes (<30), compared to the immense abundance and heterogeneity worldwide. Further efforts have begun to elucidate lake genome diversity in stratified water columns[29]. 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 km^2 of Canada, the most lake-rich landscape on Earth[30]. This study is a component of the NSERC Canadian LakePulse Network[31], 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[32–40]. 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 signaling 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.

Results

The LakePulse MAG catalogue

To produce the MAG catalogue, we generated metagenomes from the surface waters of 308 lakes in 12 Canadian ecozones (**Fig. 1**; **Fig. S1**; **Table S1**). Lakes represented a wide range of physicochemical conditions, productivity, morphometry, climate, and land use within watersheds[41] (**Fig. S2*)***. At the continental scale (43 – 68 °N, 62 – 141 °W), a full range of ultraoligotrophic to hypereutrophic lakes were present in the survey (**Fig. 1A**). Comparison of lake features by principal component analysis (PCA) revealed large-scale spatial patterns in lake and watershed characteristics (**Fig. 1B**). Western Canadian lakes were the deepest and most oligotrophic on average, and often set in watersheds with high proportions of natural landscapes and harvested forests. Northern Canadian lakes experienced the coldest climates and lowest land use pressures. Central Canada was comprised of the agriculturally intensive Prairies and Boreal Plains ecozones; compared to other Canadian regions, central Canadian lakes were generally shallow, alkaline, nutrient- and ion-rich, and highly productive. Eastern Canadian

lakes 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 watersheds with the largest population centres were 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 (e.g., 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). Though computationally intensive to produce, co-assemblies facilitated differential coverage binning, ensuring higher-quality MAGs[42]. In total, contig binning of each co-assembly recovered 1,184 MAGs exceeding MIMAG medium-quality standards[43] through an additional <10% strain heterogeneity criterion (**Table S2**). Between 28 – 233 MAGs were generated for each ecozone (**Fig. S3**). MAGs were dereplicated 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[43], excluding rRNA gene requirements (**Fig. 1C**). MAGs ranged widely in size (0.35 – 10.39 Mbp), GC content (23.1 – 75.9%), and coding density (78.4 – 96.8%) (**Table S2**). Single metagenome assemblies produced only 563 bins passing our quality control (**Table S1**); 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[21]. 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 (**Fig. 1D**). A species accumulation curve showed that hundreds of lakes were required to adequately sample genomes across the heterogenous Canadian landscape, highlighting the importance of analyzing lake metagenomes across large geographic and environmental gradients to capture the breadth of local lake bacterial diversity and ecosystem function (**Fig. 1E**). MAGs captured on average 7.1% (range 0.9 – 36.0%) of metagenome reads per lake (**Fig. 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 significant expansion in the availability of freshwater genomes and range of lake systems to date.

Phylogenetic and metabolic diversity of novel lake bacterial species

The LakePulse MAG catalogue represented a phylogenetically rich set of genomes (**Fig. 2A**; **Table S2**)

112 exhibiting high taxonomic novelty. Overall, 96% of MAGs were novel candidate species and over a third
113 (349) belonged to novel genera (**Fig. 2B**), substantially expanding the known genomic diversity of lake
114 bacteria. The largest collection of MAGs was recovered from *Bacteroidota*, a group that plays a major role
115 in organic particle degradation in marine settings[44] but is much less studied in freshwater systems[3]. A
116 third of *Bacteroidota* MAGs (80 of 275) represented new genera (**Table S3**), presenting a rich genomic
117 resource for future targeted research on freshwater *Bacteroidota* ecology and metabolism, including
118 potential zooplankton associations[45]. *Actinobacteriota* and *Gammaproteobacteria* were the next most
119 MAG-rich groups and included lineages of common and abundant freshwater bacteria (e.g., *Rhodoluna*,
120 *Polynucleobacter*, and *Limnohabitans*). Other freshwater taxa that were well represented in the MAG
121 catalogue included *Verrucomicrobiota*, *Planctomycetota*, *Patescibacteria*, and *Myxococcota*. Intriguingly,
122 we identified 38 MAGs from *Bdellovibrionota*, almost all of which represented new genera (**Table S3**).
123 *Bdellovibrionota* are described as obligate predators of Gram-negative bacteria[46] and their role in
124 freshwater food webs warrants further attention.
125
126 To assess whether the MAGs captured a representative phylogenetic diversity of lake bacteria,
127 we compared taxonomy between MAGs and 16S rRNA gene amplicon sequence variants (ASVs). MAGs
128 and ASVs overlapped in their 10 most diverse phyla, though ASVs exhibited an expected long tail of rare
129 diversity (**Fig. S4A**). Moreover, the three most MAG-rich orders (*Chitinophagales*, *Burkholderiales*, and
130 *Flavobacteriales*) were in the top four among ASVs (**Fig. S4B**). The largest differences in diversity across
131 datasets were within *Bacteroidota* and *Actinobacteriota*, which were slightly more represented as a
132 percentage of MAGs. In contrast, *Patescibacteria* exhibited higher ASV diversity and were potentially
133 underrepresented in the MAGs because of low or divergent marker gene content[47]. Overall, the
134 LakePulse MAG catalogue encompasses broad phylogenetic diversity reflective of 16S rRNA genes,
135 making it a useful resource for further taxon-focused research.
136
137 Bacteria contribute to lake ecosystem function through a diversity of energy-capturing and carbon
138 cycling metabolisms that regulate biogeochemical cycles. To provide a broad overview of MAG metabolic
139 diversity, we explored the distributions of select metabolic modules across the MAG catalogue (**Fig. 2C**;
140 **Table S2**). Oxygenic photosynthetic *Cyanobacteria* and anoxygenic photosynthetic *Chlorobia* were
141 present, while bacteriorhodopsins were identified in *Thermoleophilia*. Photosystem II (PS-II) was the most
142 widespread phototrophic module in the MAGs, including in numerous *Alphaproteobacteria*,
143 *Gammaproteobacteria*, and *Gemmatimonadota* MAGs. In combination, PS-II-encoding MAGs were
144 identified in 95% of lakes, indicating a diverse bacterial contribution to freshwater phototrophy. We
145 identified three of the six known autotrophic carbon fixation pathways[48], including the Calvin-Benson-
146 Bassham (CBB) cycle, the 3-hydroxypropionate/4-hydroxybutyrate (3HP/4HB) cycle, and the glyoxylate
147 assimilation pathway. The CBB cycle was linked to *Proteobacteria* and *Cyanobacteria* MAGs present in
148 127 lakes (42%), demonstrating that non-algal carbon fixation is also common in lakes.

149
150 Heterotrophic degradation of complex algal and terrestrial organic matter is a central ecological
151 function of lake bacteria. An abundance of *Bacteroidota* and *Gemmatimonadota* MAGs were equipped
152 with SusCD-mediated glycan transport systems (**Fig. 2C**; **Table S2**), suggesting broad potential for
153 polysaccharide transport and breakdown. Aromatic compound degradation pathways associated with the
154 metabolism of plant-derived lignin were common within *Alphaproteobacteria* and *Gammaproteobacteria*,
155 but also broadly distributed across the MAG phylogeny (**Fig. 2C**; **Table S2**). In addition to heterotrophy,
156 many *Burkholderiales* MAGs appeared to augment their energy metabolism through Sox-mediated sulfur
157 oxidation pathways. We did not find evidence for methane or ammonia oxidation within the MAG
158 catalogue. However, energy conservation via methanol oxidation and redox cycling of nitrogen oxides
159 was evident. Overall, these findings highlight that the MAG catalogue captures a broad diversity of
160 metabolisms for in-depth comparative analysis on functional contributions to lake ecosystems.
161
162 *Lake conditions shape the structure and function of bacterial assemblages*
163
164 Biogeographical mapping across Canadian lakes revealed that MAGs occupied on average 20 lakes (7%)
165 and as many as 190 lakes (63%) (**Fig. 3A**). Thus, the co-assembly-resolved MAG catalogue appeared to
166 comprise relatively widely distributed bacteria. *Patescibacteria* MAGs were among those with the most
167 restricted ranges, suggesting relatively specialist lifestyles, perhaps dependent on host ranges[49] or input
168 from groundwater environments where they thrive[50]. Notable biogeographic patterns were observed when
169 MAGs were aggregated at the class to phylum levels (**Fig. S5**; **Fig. S6**). Phyla previously determined to
170 be common in lakes (e.g., *Bacteroidota*, *Proteobacteria*, and *Actinobacteriota*) were broadly distributed
171 across the continent. Other groups were relatively restricted to distinct regions. For example, *Bacilli* and
172 *Campylobacteria* MAGs were more common in Prairies lakes than elsewhere, perhaps in response to
173 high nutrients or extensive agriculture within watersheds (**Fig. S5**; **Fig. S6**).
174
175 We next used the MAG catalogue to investigate biogeographic patterns of taxonomic and
176 functional diversity across continental-scale environmental gradients. Taxonomic variation among MAG
177 assemblages was correlated with lake trophic state, indexed by total phosphorus (TP) concentration
178 (correlations with principal coordinate analysis [PCoA] axes: r_1 = -0.32; r_2 = 0.33) (**Fig. 3B**). ASV
179 community taxonomic profiles were likewise distinguished by trophic state (r_1 = 0.41) (**Fig. S7A**),
180 demonstrating that MAGs were reasonably representative of bacterial community variation. To assess
181 functional variation, we generated lake functional profiles by aggregating the collection of metabolic
182 KEGG orthologs (KOs) encoded by the MAG assemblage within each lake. Bacterial taxonomic and
183 functional dissimilarities exhibited a positive, but variable relationship (**Fig. 3C**). PCoA of functional
184 dissimilarities revealed a distribution of MAG assemblages along axis 1 related to lake trophic state (r_1 =
185 0.23) (**Fig. 3D**) that reflected the relationship exhibited by metagenome functional profiles (r_1 = 0.40; r_2 = -

186 0.33; metagenome functional profiles form the basis of a dedicated study [Onana *et al*., in prep.]) (**Fig.**
187 **S7B**). These results suggest that nutrient availability and productivity play leading roles in structuring the
188 taxonomic and functional composition of lake bacterial assemblages at continental scale.
189
190 To further elucidate the factors influencing bacterial diversity across lakes, we evaluated the
191 importance of lake physicochemistry, morphometry, geography, and climate conditions using generalized
192 dissimilarity models (GDMs)[51]. GDMs of membership (Jaccard dissimilarities) explained more deviance
193 compared to structure (Bray-Curtis dissimilarities) (**Table S4**). As membership is more stable within lake
194 assemblages than structure due to the seasonal succession of bacterial populations, it follows that
195 membership is better predicted across large spatial scales. Lake physicochemistry explained the largest
196 amounts of taxonomic (MAGs 43.9%; ASVs 48.3%) and functional (MAGs 19.4%; metagenomes 32.6%)
197 variation among MAG assemblages and community profiles (**Fig. 3E**; **Table S4**). The consistency in
198 deviance explained by taxonomic or functional models showed that MAG assemblages reliably reflected
199 the significant environmental filtering of the broader community. The influence of trophic state was evident
200 as taxonomic and functional turnover across MAG and broader assemblages was consistently responsive
201 to chlorophyll-*a* (Chl-*a*) concentration, a measure of algal biomass (**Fig. 3F**; **Fig. S8A**). MAG taxonomic,
202 but not functional, turnover was significantly related to TP (**Fig. S8A**), suggesting metabolic differences
203 among bacterial assemblages are driven more by the composition of organic matter produced by
204 phytoplankton than by direct nutrient availability. Turnover was most rapid at low Chl-*a* and TP,
205 suggesting rapid shifts in bacterial assemblage composition across oligotrophic to mesotrophic lakes (**Fig.**
206 **3F**).
207
208 In addition to the effects of trophic state, GDMs provided insights into other physicochemical
209 factors shaping lake bacterial assemblages at continental scale. MAG and community taxonomic turnover
210 was significantly associated with pH, and MAG functional turnover was additionally predicted by lake
211 surface temperature (**Fig. 3G**; **Fig. S8A**). Temperature and pH are often considered master variables
212 driving large-scale spatial patterns in marine[52] and soil[53] communities, respectively. As aquatic systems
213 that are heavily influenced by their terrestrial surroundings, it follows that lake bacterioplankton
214 assemblages are influenced by a combination of temperature and pH.
215
216 Remarkably, calcium concentration was the strongest predictor of both taxonomic and functional
217 turnover across MAG assemblages and community-level taxonomy (**Fig. 3H**; **Fig. S8A**). A shift in
218 bacterial ecotype diversity along a freshwater calcium gradient was previously reported[54], but how
219 calcium directly influences assemblage composition is unclear. The effect is perhaps indirectly linked to
220 interactions among plankton, as low calcium was previously shown to limit zooplankton growth[55],
221 potentially eliciting trophic cascades through the lake food web. Further investigation of the mechanisms
222 and implications of the relationship between calcium and bacterial assemblages is warranted, particularly

223 as lake calcium decline is a legacy effect of emissions-linked acid rain[56], while calcium levels may be
224 increasing elsewhere through winter road salt application[57].
225
226 The demonstration that lake physicochemistry shapes the functional composition of MAG
227 assemblages prompted a deeper analysis focused on specific metabolic categories. We generated GDMs
228 for different KEGG metabolic categories implicated in energy and nutrient cycling. As with the full
229 functional profiles, lake physicochemistry consistently explained more deviance than morphometry,
230 geography, or climate (**Fig. 4A**). Interestingly, explained deviance varied by metabolic category, and
231 variation within carbohydrate metabolism was the best predicted by physicochemistry (**Fig. 4A**).
232 Significant predictors of carbohydrate metabolism included Chl-*a* concentration and lake colour (**Fig. 4B**;
233 **Fig. S8B**). Lake colour reflects terrestrial organic matter input from the watershed. Hence, variation in
234 carbohydrate metabolism across bacterial assemblages appears to be linked to the quantity and
235 composition of autochthonous (algal) and allochthonous (plant) organic matter in lakes. We note that lipid
236 metabolism was the only additional metabolic category where turnover was predicted by Chl-*a* and colour
237 (**Fig. 4B**; **Fig. S8B**). Subsequent PCoAs of carbohydrate and lipid metabolisms in lakes showed strong
238 relationships with trophic state and large variation across the eutrophic to hypereutrophic spectrum (**Fig.
239 4C-D**). Overall, these results suggest that a diverse and dynamic organic matter pool characterized by
240 punctuated inputs or resuspensions of nutrients in eutrophic to hypereutrophic systems may be a key
241 driver of metabolic diversity within the lake microbiome.
242
243 The metabolism of complex carbohydrates of aquatic or terrestrial origin depends on initial
244 degradation through carbohydrate-active enzymes (CAZymes). MAGs contained >13,000 genes assigned
245 to 129 glycoside hydrolase (GH) gene families, implicated in cleaving the glycosidic bonds of diverse
246 polysaccharides. Like carbohydrate metabolism (KEGG category), GH repertoires within bacterial
247 assemblages were linked to lake physicochemistry (**Fig. 4A**). 41% of GH genes were identified in
248 *Bacteroidota* MAGs (**Fig. 4E**) and were often encoded in polysaccharide utilization loci (PULs) (**Fig. S9**),
249 demonstrating that freshwater *Bacteroidota* are central to complex organic matter degradation
250 comparable to their role in marine and gut ecosystems[58]. *Verrucomicrobiota* MAGs were also replete with
251 GH genes (18%), expanding on previous observations that freshwater *Verrucomicrobiota* are central
252 players in organic matter degradation[59]. PCA of GH distributions across MAGs demonstrated a strong
253 taxonomic structure for carbohydrate-processing potential in lake bacteria (**Fig. 4F**). *Bacteroidota*
254 displayed the most distinct GH family repertoires, characterized by an enrichment in GH74 implicated in
255 cellulose and xyloglucan degradation and GH20 implicated in anhydromuropeptides recycling and chitin
256 degradation. *Proteobacteria* and *Bdellovibrionota* displayed overlapping GH repertoires distinct from other
257 taxonomic groups and driven by GH23, GH102, and GH103, which are all lysozymes implicated in
258 peptidoglycan degradation, aligning with potential predatory activity against other bacteria[60]. The
259 thousands of GH genes encoded in the LakePulse MAG catalogue represent a rich resource to further

260 examine organic matter degradation in freshwaters and to expand on previous global explorations of
261 CAZyme diversity across ecosystems[61,62].
262
263 *Connecting lake bacterial diversity to watershed characteristics and human impact*
264
265 While the influence of lake physicochemistry on bacterial ecology has received considerable attention and
266 is elaborated upon in this study, terrestrial influences from the surrounding watershed are less
267 understood. The LakePulse survey was complemented with geomatics descriptions of lake watersheds,
268 including soils and land use[31]. Here, we coupled the geomatics data with the MAG catalogue to provide
269 continental-scale insights into watershed impacts on lake bacterial assemblages across the land-water
270 interface. Soil properties explained significant GDM deviance in both taxonomic (MAGs 12.1%; ASVs
271 23.2%) and functional (MAGs 16.3%; metagenomes 9.2%) turnover (**Fig. 5A**). Notably, across all GDMs
272 developed in this study, the importance of soil was second only to lake physicochemistry, providing
273 evidence for significant terrestrial influence on lake bacterial assemblage structure and function. Turnover
274 across MAG and community profiles was most strongly associated with soil pH and the volumetric fraction
275 of coarse fragments, a variable describing soil texture (**Fig. 5B**; **Fig. S8A**). Soil texture governs water
276 retention, while pH determines the mobility of ions and nutrients[63]. Soil attributes thereby direct the
277 composition and release of catchment runoff, which in turn modifies lake physicochemistry. Hence,
278 terrestrial inputs from soils appear to influence lake bacterial diversity in a moderately predictable manner
279 at continental scale.
280
281 Next, we investigated if bacterial assemblages could be predicted by human impact on the
282 watershed. There is accumulating evidence that the taxonomic diversity of lake bacteria is influenced by
283 land use at large spatial scales[32,64,65], yet effects on function are unclear. We were particularly interested
284 in how agriculture (**Fig. 5C**) and human population density (**Fig. 5D**) within watersheds may shape
285 bacterial diversity, as each contribute to the eutrophication and chemical contamination of lakes. Land
286 use GDMs explained a lesser, but still significant, amount of taxonomic (MAGs 6.8%; ASVs 9.6%) and
287 functional (MAGs 5.4%) turnover compared to soils. There was no statistically significant response of
288 metagenome function to land use gradients, and given that MAGs are an incomplete representation of the
289 broader community, signals in MAG functional turnover should be interpreted cautiously. Based on
290 evidence from GDMs, MAG assemblages are an informative subset through which to investigate – and
291 potentially indicate – anthropogenic pressures on watersheds altering lake bacterial assemblages across
292 the land-water interface. The functional effects of land use on MAG assemblages may reflect the
293 responses of taxonomic composition, yet the coupling of taxonomy and function in MAGs present a dual-
294 faceted resource to study human impact.
295
296 Of all human impact variables, agriculture exhibited a significant effect on both MAG taxonomic

297 and functional turnover, while human population density exhibited a significant effect on functional
298 turnover (**Fig. 5E**). Intriguingly, when analyzing specific metabolic categories, xenobiotic metabolism was
299 the one most strongly explained by land use (**Fig. 5E**). Consistent with general functional turnover (**Fig.**
300 **5F**; **Fig. S8A**), human population density followed by agriculture were the strongest predictors of turnover
301 within xenobiotic metabolism (**Fig. 5G**; **Fig. S8B**). In both cases, the most rapid turnover occurred where
302 human populations were densest (1,000 – 3,000 individuals/km^2), corresponding to the few lakes
303 sampled within the Vancouver and Toronto metropolitan areas (**Fig. 5D**). The underlying effect of
304 population density appeared complex: bacterial assemblages within highly populated landscapes were
305 characterized by a lower diversity of KOs linked to xenobiotic metabolism. Moreover, PCoA showed that
306 although xenobiotic metabolism in lakes within high-population watersheds were distinct from those within
307 low-population watersheds, there was also large variation in xenobiotic metabolism among lakes near
308 population centres (**Fig. S10**). As many of the detected KOs within xenobiotic metabolism were involved
309 in the degradation of aromatic compounds, lake bacteria may be capable of metabolizing aromatic
310 environmental pollutants, such as agricultural pesticides or industrial wastes[66]. The metabolism of land-
311 derived pollution by lake bacteria merits further investigation, considering watersheds are increasingly
312 converted to agriculture and urban development[67], and altered precipitation patterns are increasing
313 runoff[68]. Overall, these results suggest that the size and activities of human populations have a detectable
314 and significant influence on the functional diversity encoded in the freshwater microbiome, as represented
315 in the LakePulse MAG catalogue.
316
317 **Discussion**
318
319 The LakePulse MAG catalogue, compiled from a continental-scale sampling of hundreds of lakes,
320 addresses an urgent need to probe the microbiome underpinning freshwater ecosystems under rapid and
321 accelerating anthropogenic change. This study has substantially expanded the number and diversity of
322 genomes from freshwater ecosystems and demonstrated that lake bacteria are equipped to metabolize
323 diverse substrates originating from both within lakes and their catchments. Trophic state emerged as a
324 leading driver of taxonomic and functional variation among lake bacterial assemblages at continental
325 scale, with implications for understanding bacterial responses to ongoing anthropogenic eutrophication of
326 lakes[9]. The functional responses of MAG assemblages to land use pressures present a useful system for
327 investigating human impact on the lake microbiome across the land-water interface. The links between
328 lake bacterial diversity and watershed variables, more stable over short timescales than lake
329 physicochemical conditions, could be leveraged to predict freshwater assemblages using remote sensing
330 data to decrease the dependence on *in situ* surveys and monitoring[37,40]. Our findings make clear that lake
331 ecosystems cannot be studied in isolation of their watersheds in an era of intensifying land use[67].
332
333 Through its construction and analysis, the LakePulse MAG catalogue makes strides toward

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 (e.g., xenobiotics; antibiotic resistance[39]) in human-altered watersheds. More lake bacterial diversity and ecosystem function remain to be captured in stratified water columns[20,29], sediments[69,70], across seasons[71], and potentially through deeper metagenome binning[21]. Finally, microdiversity[72] 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.

Methods

Lake sampling

Lake sampling was conducted by the Natural Sciences and Engineering Research Council (NSERC) Canadian Lake Pulse Network (LakePulse) between 2017 – 2019[31]. 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 km^2, medium: 0.5 – 5 km^2, and large: 5 – 100 km^2) 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[73]. 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. 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 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,

371 discharge, residence time, and slope were accessed from HydroLakes v. 1.0[30]). Shoreline length and
372 average depth were strongly colinear with other morphometry variables ($r \geq 0.7$) and therefore excluded
373 from the analysis. Surface water physicochemical variables included temperature, pH, colour, and
374 concentrations of Chl-*a*, DIC, DOC, TN, TP, calcium, chloride, magnesium, potassium, sodium, and
375 sulfate. Missing physicochemical data were replaced with ecozone medians. Watershed soil variables
376 estimated for the top 0 – 5 cm soil depth interval were accessed from SoilGrids250m[74] and included bulk
377 density of the fine earth fraction, cation exchange capacity, volumetric fraction of coarse fragments,
378 proportions of clay, sand, and silt particles in the fine earth fraction, total nitrogen, pH, soil organic carbon
379 content in the fine earth fraction, and organic carbon density. Land use variables were calculated as
380 fractions of watershed area not covered by water and included crop agriculture, pasture, forestry, built
381 development, human population density, livestock density, and poultry density. Climate variables
382 measured over the seven days prior to lake sampling were accessed from ERA5-Land hourly reanalysis[75]
383 and included mean air temperature, total precipitations, mean net solar radiation, mean wind speed, and
384 ice disappearance day for the year of sampling. Lake trophic state was categorized by TP concentrations
385 according to the Canadian Water Quality guidelines: ultraoligotrophic (<4 µg/L), oligotrophic (4 – 10 µg/L),
386 mesotrophic (10 – 20 µg/L), mesoeutrophic (20 – 35 µg/L), eutrophic (35 – 100 µg/L), and hypereutrophic
387 (>100 µg/L)[76].
388
389 *DNA extraction and metagenome sequencing*
390
391 DNA was extracted from filters using the DNeasy PowerWater kit (QIAGEN) according to the
392 manufacturer's instructions supplemented by the optional addition of 1 µL ribonuclease A and 30 min
393 incubation at 37 °C. DNA was submitted to Genome Quebec for library preparation using the NEBNext
394 Ultra II DNA Library Prep Kit (New England Biolabs) and 150 bp paired-end shotgun sequencing on an
395 Illumina NovaSeq 6000 platform.
396
397 *Metagenome assembly and binning*
398
399 Adapter-clipping and quality-trimming of raw reads were performed in Trimmomatic v. 0.36 using a Phred
400 quality score of 33 and default settings[77]. We used a co-assembly approach to leverage the differential
401 depth of coverage across contigs and metagenomes and thus maximize the recovery of genomes
402 common to the same ecozone. Metagenomes were co-assembled in MEGAHIT v. 1.2.7 using k-mer
403 lengths 27, 37, 47, 57, 67, 77, 87 and a minimum k-mer multiplicity filter of two[78]. Because only three
404 lakes were sampled in the Taiga Cordillera and these bordered the Boreal Taiga, an ecozone with similar
405 climate, landform, and vegetation, metagenomes from these two ecozones were combined into one co-
406 assembly.
407

408 Local alignment of metagenome reads to co-assemblies was performed with the Burrows-
409 Wheeler Aligner (BWA) v. 0.7.17 using the BWT-Smith-Waterman (BWTSW) index algorithm and BWA-
410 Maximal Exact Matches[79]. Binning was performed in MetaBAT v. 2.12.1 using the default minimum
411 scaffold length of 2,500 bp and based on the average differential coverage of each contig determined by
412 the jgi_summarize_bam_contig_depths script implementing a 97% sequence identity threshold[80]. Bin
413 completeness, contamination, and strain heterogeneity were estimated in CheckM v. 1.0.7[81]. Bins scoring
414 ≥50% completeness, <10% contamination, and <10% strain heterogeneity were retained as MAGs.
415 MAGs were dereplicated at average nucleotide identity (ANI) ≥95% and alignment coverage ≥10% in
416 dRep v. 3.2.0 implementing gANI as the second clustering algorithm[82]. In cases where more than one
417 MAG belonged to the same genomospecies, the highest-quality MAG was selected as the
418 genomospecies representative to be included in the final MAG catalogue and subsequent analysis. MAGs
419 were categorized according to MIMAG standards by which high-quality drafts display completeness
420 >90%, contamination <5%, and ≥18/20 tRNA genes (16S, 23S, and 5S rRNA gene presence criteria were
421 omitted) and medium-quality drafts display completeness ≥50% and contamination <10%[43].
422
423 *Functional analysis*
424
425 Protein-coding genes were annotated in Prokka v. 1.12[83] which implemented Prodigal v. 2.6.3 for gene
426 prediction[84]. Ribosomal RNA genes were predicted in Infernal v. 1.1.2[85] against Rfam v. 14.2[86]. Gene
427 functions were annotated in KofamScan using default settings and a bitscore-to-threshold ratio of 0.7[87].
428 Transporters were predicted in BLASTp against the Transporter Classification Database downloaded on
429 January 27, 2021[88] using a maximum e-value of 10^{-20}. CAZymes were predicted in hmmsearch from
430 HMMER v. 3.1b2[89] against dbCAN2 HMMdb v. 9[90]. GH family 109 was omitted from CAZyme analysis
431 due to nonspecificity of the hidden Markov model.
432
433 Functions were assigned based on the following set of rules. Oxygenic photosynthesis required
434 evidence of at least one marker gene from both photosystems I (psaA-F: K02689, K02690, K02691,
435 K02692, K02693, or K02694) and II (psbA-F: K02703, K02704, K02705, K02706, K02707, or K02708).
436 For anoxygenic photosynthesis, evidence of photosystem II required the presence of both pufL (K08928)
437 and pufM (K08929); evidence of photosystem I required the presence of all four proteins (pscA-D:
438 K08940, K08941, K08942, K08943). Bacteriorhodopsin assignment required evidence of bop (K04641).
439 Autotrophic carbon fixation via three independent pathways was evidenced by rbcL (K01601) and prkB
440 (K00855) for the CBB cycle, mct (K14470) for the glyoxylate assimilation pathway, and abfD (K14534) for
441 the 3HP/4HB cycle. Evidence of glycan import mediated by the susCD protein complex was determined
442 by the presence of both susC (K21573) and susD (K21572). Aromatic compound degradation was
443 determined by the presence of either a funneling or ring fission pathway: evidence of catabolic funneling
444 was identified by ferulate to vanillin conversion (evidenced by fcs (K12508)), vanillin to protocatechuate

445 conversion (evidenced by vanA (K03862)), or protocatechuate ring opening (evidenced by pcaG
446 (K00448) or ligA (K04100)); ring fission was identified by gallate ring opening (evidenced by galA
447 (K04099)), catechol ring opening (evidenced by K00446 or K03381), salicylate to gentisate conversion
448 (evidenced by K00480 or K18242 or K18243), or gentisate ring opening (evidenced by K00450).
449 Methanol oxidation was evidenced by xoxF (K23995). Nitrogen fixation was evidenced by nifD (K02586)
450 or nifH (K02588) or nifK (K02591). Dissimilatory nitrate reduction was evidenced by narG (K00370) and
451 narH (K00371) or napA (K02567) and napB (K02568). Evidence of nitrite reduction was determined by
452 the presence of 1) nirB (K00362) and nirD (K00363), 2) nrfA (K03385) and nrfH (K15876), 3) nirK
453 (K00368), or 4) nirS (K15864). Nitric oxide reduction was evidenced by norB (K04561) and norC
454 (K02305). Nitrous oxide reduction was evidenced by nosZ (K00376). Evidence of sulfur oxidation via SOX
455 was determined by the presence of 1) soxA (K17222) and soxB (K17224) and soxX (K17223) and soxY
456 (K17226) and soxZ (K17227), or 2) soxC (K17225) and soxD (K22622). Dissimilatory sulfate reduction
457 was evidenced by 1) aprA (K00394) and aprB (K00395), or 2) dsrA (K11180) and dsrB (K11181).
458
459 PULs were delineated following the algorithm described in Terrapon *et al*.[91]. First, adjacent SusC
460 and SusD genes located on the same DNA strand (termed tandem susCD pairs) were identified. Tandem
461 susCD pairs within five loci of each other were considered a single susCD-containing locus. Next,
462 operons structured around tandem susCD pairs were delineated by including genes within intergenic
463 distances of ≤200 bp. PUL boundaries were iteratively extended to neighbouring operons when a
464 CAZyme was detected within the first five genes. The selected intergenic distance threshold exceeded
465 the empirically-derived 102 bp cut-off proposed by Terrapon *et al*.[91] so as to include functionally-
466 associated genes other than CAZymes that would otherwise be excluded. Only CAZymes implicated in
467 glycan degradation (glycoside hydrolases, polysaccharide lyases, carbohydrate-binding modules, and
468 carbohydrate esterases) were included in conditional PUL extension. Finally, only PULs containing
469 specified CAZyme types were retained.
470
471 *MAG read recruitment and coverage normalization*
472
473 To avoid recruitment to conserved regions, rRNA and tRNA gene sequences were masked in BEDTools
474 v. 2.26.0[92] prior to mapping. To determine the coverage of MAGs across all lake metagenomes, trimmed
475 metagenome reads were initially recruited to MAGs at a sequence identity threshold of 95% in BBMap v.
476 37.76[93], then filtered with a hard cut-off of 96% identity using a custom Python script. The selection of a
477 95% alignment similarity cut-off reflected a prokaryotic genome ANI species delineation[94]. Mapping files
478 were converted from SAM to sorted BAM format in Samtools[95]. Horizontal coverage was calculated
479 following the method of Rodriguez-R *et al*.[21] as TAD_{80}, the average sequencing depth truncated to the
480 central 80% of mapped positions normalized by the number of genome equivalents. Genome equivalents
481 in each metagenome were estimated in MicrobeCensus v. 1.1.0[96]. A positive TAD_{80} value signified that

482 ≥10% of MAG positions were mapped by metagenome reads; the ≥10% sequencing breadth threshold
483 has been shown to detect species presence at a confidence of >95%[97].
484
485 *Taxonomic classification and phylogenetic analysis*
486
487 MAG taxonomic classification was performed in GTDB-Tk v. 1.3.0 by phylogenetic placement into
488 bacterial reference tree v. r95[98]. A *de novo* phylogenetic tree was constructed in GTDB-Tk using the
489 alignment of 120 marker genes based on the JTT model of evolution and subsequently plotted with the R
490 package ggtree[99].
491
492 *Metagenome analysis*
493
494 Single metagenome assemblies were generated in MEGAHIT[78] using k-mer lengths 23, 43, 63, 83, 103,
495 123 or metaSPAdes[100] v. 3.13.0 using k-mer lengths 21, 33, 55, 77, 99, 127 and mapped in BWA v.
496 0.7.17 using the BWTSW index algorithm and BWA-Maximal Exact Matches[79]. Assemblies and coverage
497 files were submitted to the DOE-JGI Metagenome Annotation Pipeline v. 4.16.6, 5.0.20, or 5.0.23[101].
498 Metagenome functional diversity was quantified by relative gene counts, calculated as KO counts
499 normalized by total contig average depth of coverage.
500
501 *16S rRNA gene amplicon sequencing and analysis*
502
503 A ~300 bp fragment of the V4 region of the 16S rRNA gene was amplified using the primer set 515F (5'-
504 GTGCCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3')[102]. Each PCR
505 mixture consisted of 5 µL Phusion High Fidelity Buffer, 0.5 µL dNTPs (10 mM), 1.8 µL of each primer (5
506 µM), 0.25 µL Phusion polymerase, 13.65 µL H_20, and 10 ng in a 2 µL volume of template DNA. PCR
507 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
508 a final elongation at 72 °C for 60 s. Sequencing of 250 bp paired-end reads was performed on an Illumina
509 MiSeq platform.
510
511 Primer sequences were trimmed in Cutadapt v. 3.1[103]. ASVs were inferred in DADA2 v. 1.16[104],
512 which performed trimming of low-quality 3'-end positions from forward and reverse reads, quality retention
513 of reads 160 – 200 bp in length, ASV inference within individual samples informed by machine-learned
514 sequencing error rates, merging of paired reads, and removal of consensus chimeric ASVs. ASVs <250
515 or >256 bp in length were removed along with ASVs not assigned at the kingdom level and ASVs
516 assigned to archaea, eukaryotes, or chloroplasts. ASV data from two lakes were discarded because they
517 did not meet the threshold 10,000 sequences. ASVs were classified against GTDB r95 16S rRNA gene
518 sequences formatted for DADA2[105].

Ecological analysis

PCA of lake features was performed in the R package vegan v. 2.5.6[106] 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 mS/cm and total major ions <4,000 mg/L. 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[107] 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[108]. PCA of MAG GH composition was performed on Hellinger-transformed data. Data analysis and visualization were performed in R v. 4.0.1[109] using the tidyverse v. 1.3.0 suite of packages[110].

Code availability

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

Data availability

Raw metagenome reads were archived in the European Nucleotide Archive under study accession PRJEB29238 (https://www.ebi.ac.uk/ena/browser/view/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 annotations were deposited in Dryad (https://doi.org/10.5061/dryad.zkh1893fs).

Acknowledgements

This study was funded by the NSERC Canadian LakePulse Network (Strategic Network Grant NETGP-479720) and Canada Research Chairs held by DAW and YH. REG and VEO received scholarships from the NSERC CREATE ÉcoLac Training Program in Lake and Fluvial Ecology. REG 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 Jelena Juric, Paul MacKeigan, Cindy Paquette, Marieke Beaulieu, Katherine Griffiths, Geneviève Potvin, Bruno Cremella, Gabriel Diab, Vincent Fugère, Jihyeon Kim, Anaïs Oliva, and Katherine Velghe. We thank Wentworth Brookes, Shawn Simpson, and Compute Canada for bioinformatics technical support.

References

1. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 14–49 (2011).
2. Pernthaler, J. Competition and niche separation of pelagic bacteria in freshwater habitats. *Environ. Microbiol*. **19**, 2133–2150 (2017).
3. Bertilsson, S. & Mehrshad, M. Diversity and Dynamics of Bacterial Communities in Freshwater Lakes. in *Encyclopedia of Inland Waters* 601–615 (Elsevier, 2022).
4. Kratz, T. K., MacIntyre, S. & Webster, K. E. Causes and Consequences of Spatial Heterogeneity in Lakes. in *Ecosystem Function in Heterogeneous Landscapes* (eds. Lovett, G. M., Jones, C., Turner, M. G. & Weathers, K. C.) 329–347 (Springer New York, 2005).
5. Tranvik, L. J. *et al*. Lakes and reservoirs as regulators of carbon cycling and climate. *Limnol. Oceanogr*. **54**, 2298–2314 (2009).
6. Williamson, C. E., Dodds, W., Kratz, T. K. & Palmer, M. A. Lakes and streams as sentinels of environmental change in terrestrial and atmospheric processes. *Front. Ecol. Environ*. **6**, 247–254 (2008).
7. Grossart, H., Massana, R., McMahon, K. D. & Walsh, D. A. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnol. Oceanogr.* **65**, 1–19 (2020).
8. O'Reilly, C. M. *et al.* Rapid and highly variable warming of lake surface waters around the globe. *Geophys. Res. Lett.* **42**, 10773–10781 (2015).
9. Smith, V. H. & Schindler, D. W. Eutrophication science: where do we go from here? *Trends Ecol. Evol.* **24**, 201–207 (2009).
10. Jane, S. F. *et al.* Widespread deoxygenation of temperate lakes. *Nature* **594**, 66–70 (2021).
11. Dugan, H. A. *et al.* Lakes at Risk of Chloride Contamination. *Environ. Sci. Technol.* **54**, 6639–6650 (2020).
12. Reid, A. J. *et al.* Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biol. Rev.* **94**, 849–873 (2019).
13. Sheridan, E. A. *et al*. Plastic pollution fosters more microbial growth in lakes than natural organic matter. *Nat. Commun*. **13**, 4175 (2022).
14. Tyson, G. W. *et al*. Community structure and metabolism through reconstruction of microbial

593 genomes from the environment. *Nature* **428**, 37–43 (2004).
594 15. Hug, L. A. *et al*. A new view of the tree of life. *Nat. Microbiol*. **1**, 16048 (2016).
595 16. Cabello-Yeves, P. J. *et al.* Genomes of Novel Microbial Lineages Assembled from the Sub-Ice
596 Waters of Lake Baikal. *Appl. Environ. Microbiol.* **84**, 1–21 (2018).
597 17. Okazaki, Y., Nakano, S., Toyoda, A. & Tamaki, H. Long-Read-Resolved, Ecosystem-Wide
598 Exploration of Nucleotide and Structural Microdiversity of Lake Bacterioplankton Genomes.
599 *mSystems* **7**, e00433-22. (2022).
600 18. Mehrshad, M. *et al.* Hidden in plain sight—highly abundant and diverse planktonic freshwater
601 Chloroflexi. *Microbiome* **6**, 176 (2018).
602 19. Arora-Williams, K. *et al*. Dynamics of microbial populations mediating biogeochemical cycling in a
603 freshwater lake. *Microbiome* **6**, 165 (2018).
604 20. Tran, P. Q. *et al.* Depth-discrete metagenomics reveals the roles of microbes in biogeochemical
605 cycling in the tropical freshwater Lake Tanganyika. *ISME J.* **15**, 1971–1986 (2021).
606 21. Rodriguez-R, L. M., Tsementzi, D., Luo, C. & Konstantinidis, K. T. Iterative subtractive binning of
607 freshwater chronoseries metagenomes identifies over 400 novel species and their ecologic
608 preferences. *Environ. Microbiol.* **22**, 3394–3412 (2020).
609 22. Kavagutti, V. S. *et al*. High-resolution metagenomic reconstruction of the freshwater spring bloom.
610 *Microbiome* **11**, 15 (2023).
611 23. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-
612 assembled genomes from the global oceans. *Sci. Data* **5**, 170203 (2018).
613 24. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected
614 biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
615 25. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504
616 (2019).
617 26. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from
618 uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
619 27. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000
620 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20
621 (2019).
622 28. Nayfach, S. *et al.* A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
623 29. Buck, M. *et al.* Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater
624 lakes and ponds. *Sci. Data* **8**, 131 (2021).
625 30. Messager, M. L., Lehner, B., Grill, G., Nedeva, I. & Schmitt, O. Estimating the volume and age of
626 water stored in global lakes using a geo-statistical approach. *Nat. Commun.* **7**, 13603 (2016).
627 31. Huot, Y. *et al.* The NSERC Canadian Lake Pulse Network: A national assessment of lake health
628 providing science for water management in a changing climate. *Sci. Total Environ.* **695**, 133668
629 (2019).

630 32. Kraemer, S. A. *et al.* A large-scale assessment of lakes reveals a pervasive signal of land use on
631 bacterial communities. *ISME J.* **14**, 3011–3023 (2020).
632 33. Wight, J., Varin, M.-P., Robertson, G. J., Huot, Y. & Lang, A. S. Microbiology in the Field:
633 Construction and Validation of a Portable Incubator for Real-Time Quantification of Coliforms and
634 Other Bacteria. *Front. Public Heal*. **8**, 1–11 (2020).
635 34. Garner, R. E., Gregory-Eaves, I. & Walsh, D. A. Sediment Metagenomes as Time Capsules of
636 Lake Microbiomes. *mSphere* **5**, e00512-20 (2020).
637 35. Garner, R. E. *et al.* Protist Diversity and Metabolic Strategy in Freshwater Lakes Are Shaped by
638 Trophic State and Watershed Land Use on a Continental Scale. *mSystems* **7**, (2022).
639 36. MacKeigan, P. W. *et al.* Comparing microscopy and DNA metabarcoding techniques for identifying
640 cyanobacteria assemblages across hundreds of lakes. *Harmful Algae* **113**, 102187 (2022).
641 37. Oliva, A., Garner, R. E., Walsh, D. & Huot, Y. The occurrence of potentially pathogenic fungi and
642 protists in Canadian lakes predicted using geomatics, in situ and satellite-derived variables:
643 Towards a tele-epidemiological approach. *Water Res.* **209**, 117935 (2022).
644 38. Griffiths, K. *et al.* Pervasive changes in algal indicators since pre-industrial times: A
645 paleolimnological study of changes in primary production and diatom assemblages from ~200
646 Canadian lakes. *Sci. Total Environ.* **838**, 155938 (2022).
647 39. Kraemer, S. A., Barbosa da Costa, N., Oliva, A., Huot, Y. & Walsh, D. A. A resistome survey
648 across hundreds of freshwater bacterial communities reveals the impacts of veterinary and human
649 antibiotics use. *Front. Microbiol*. **13**, 1–13 (2022).
650 40. Oliva, A. *et al*. Geospatial analysis reveals a hotspot of fecal bacteria in Canadian prairie lakes
651 linked to agricultural non-point sources. *Water Res*. **231**, 119596 (2023).
652 41. Sánchez Schacht, J. R., MacKeigan, P. W., Taranu, Z. E., Huot, Y. & Gregory-Eaves, I.
653 Agricultural and urban land use explain substantial variation in lake water quality across national-
654 scale physiographic and climatic gradients. Preprint at https://doi.org/10.1101/2022.08.29.505280
655 (2022).
656 42. Albertsen, M. *et al*. Genome sequences of rare, uncultured bacteria obtained by differential
657 coverage binning of multiple metagenomes. *Nat. Biotechnol*. **31**, 533–538 (2013).
658 43. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a
659 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731
660 (2017).
661 44. Fernández-Gómez, B. *et al.* Ecology of marine Bacteroidetes: a comparative genomics approach.
662 *ISME J.* **7**, 1026–1037 (2013).
663 45. De Corte, D. *et al.* Metagenomic insights into zooplankton-associated bacterial communities.
664 *Environ. Microbiol.* **20**, 492–505 (2018).
665 46. Ezzedine, J. A., Desdevises, Y. & Jacquet, S. Bdellovibrio and like organisms: current
666 understanding and knowledge gaps of the smallest cellular hunters of the microbial world. *Crit.*

667 *Rev. Microbiol.* **48**, 428–449 (2022).
668 47. Brown, C. T. *et al*. Unusual biology across a group comprising more than 15% of domain Bacteria.
669 *Nature* **523**, 208–211 (2015).
670 48. Berg, I. A. Ecological Aspects of the Distribution of Different Autotrophic CO_2 Fixation Pathways.
671 *Appl. Environ. Microbiol.* **77**, 1925–1936 (2011).
672 49. He, C. *et al.* Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR
673 bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* **6**, 354–365 (2021).
674 50. Tian, R. *et al.* Small and mighty: adaptation of superphylum Patescibacteria to groundwater
675 environment drives their genome simplicity. *Microbiome* **8**, 51 (2020).
676 51. Ferrier, S., Manion, G., Elith, J. & Richardson, K. Using generalized dissimilarity modelling to
677 analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.*
678 **13**, 252–264 (2007).
679 52. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**,
680 1261359–1261359 (2015).
681 53. Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**,
682 320–325 (2018).
683 54. Hahn, M. W., Huemer, A., Pitt, A. & Hoetzinger, M. Opening a next-generation black box:
684 Ecological trends for hundreds of species-like taxa uncovered within a single bacterial >99% 16S
685 rRNA operational taxonomic unit. *Mol. Ecol. Resour.* **21**, 2471–2485 (2021).
686 55. Azan, S. S. E. & Arnott, S. E. The impact of calcium decline on population growth rates of
687 crustacean zooplankton in Canadian Shield lakes. *Limnol. Oceanogr.* **63**, 602–616 (2018).
688 56. Weyhenmeyer, G. A. *et al.* Widespread diminishing anthropogenic effects on calcium in
689 freshwaters. *Sci. Rep.* **9**, 10450 (2019).
690 57. Kaushal, S. S. *et al.* Freshwater salinization syndrome on a continental scale. *Proc. Natl. Acad.*
691 *Sci.* **115**, E574–E583 (2018).
692 58. McKee, L. S. *et al.* Polysaccharide degradation by the Bacteroidetes: mechanisms and
693 nomenclature. *Environ. Microbiol. Rep.* **13**, 559–581 (2021).
694 59. He, S. *et al.* Ecophysiology of Freshwater Verrucomicrobia Inferred from Metagenome-Assembled
695 Genomes. *mSphere* **2**, e00277-17 (2017).
696 60. Harding, C. J. *et al.* A lysozyme with altered substrate specificity facilitates prey cell exit by the
697 periplasmic predator Bdellovibrio bacteriovorus. *Nat. Commun.* **11**, 4817 (2020).
698 61. Berlemont, R. & Martiny, A. C. Glycoside Hydrolases across Environmental Microbial
699 Communities. *PLOS Comput. Biol.* **12**, e1005300 (2016).
700 62. Garron, M.-L. & Henrissat, B. The continuing expansion of CAZymes and their families. *Curr.*
701 *Opin. Chem. Biol.* **53**, 82–87 (2019).
702 63. Delgado, A. & Gómez, J. A. The Soil. Physical, Chemical and Biological Properties. in *Principles of*
703 *Agronomy for Sustainable Agriculture* (eds. Villalobos, F. J. & Fereres, E.) 15–26 (Springer

International Publishing, 2016).

64. Marmen, S. *et al*. The Role of Land Use Types and Water Chemical Properties in Structuring the Microbiomes of a Connected Lake System. *Front. Microbiol*. **11**, 89 (2020).

65. Sperlea, T. *et al*. The relationship between land cover and microbial community composition in European lakes. *Sci. Total Environ*. **825**, 153732 (2022).

66. Phale, P. S., Sharma, A. & Gautam, K. Microbial degradation of xenobiotics like aromatic pollutants from the terrestrial environments. in *Pharmaceuticals and Personal Care Products: Waste Management and Treatment Technology* 259–278 (Elsevier, 2019).

67. Foley, J. A. *et al*. Global Consequences of Land Use. *Science* **309**, 570–574 (2005).

68. Kaushal, S. S. *et al.* Land Use and Climate Variability Amplify Carbon, Nutrient, and Contaminant Pulses: A Review with Management Implications. *JAWRA J. Am. Water Resour. Assoc.* **50**, 585–614 (2014).

69. Salcher, M. M., Schaefle, D., Kaspar, M., Neuenschwander, S. M. & Ghai, R. Evolution in action: habitat transition from sediment to the pelagial leads to genome streamlining in Methylophilaceae. *ISME J*. **13**, 2764–2777 (2019).

70. Biessy, L., Pearman, J. K., Waters, S., Vandergoes, M. J. & Wood, S. A. Metagenomic insights to the functional potential of sediment microbial communities in freshwater lakes. *Metabarcoding and Metagenomics* **6**, 59–74 (2022).

71. Linz, A. M. *et al*. Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *PeerJ* **6**, e6075 (2018).

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

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

74. Hengl, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* **12**, e0169748 (2017).

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

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

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

78. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).

79. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.

741 *Bioinformatics* **25**, 1754–1760 (2009).
742 80. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome
743 reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
744 81. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing
745 the quality of microbial genomes recovered from isolates, single cells, and metagenomes.
746 *Genome Res.* **25**, 1043–1055 (2015).
747 82. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic
748 comparisons that enables improved genome recovery from metagenomes through de-replication.
749 *ISME J.* **11**, 2864–2868 (2017).
750 83. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
751 84. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification.
752 *BMC Bioinformatics* **11**, 119 (2010).
753 85. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*
754 **29**, 2933–2935 (2013).
755 86. Kalvari, I. *et al.* Rfam 14: expanded coverage of metagenomic, viral and microRNA families.
756 *Nucleic Acids Res.* **49**, D192–D200 (2021).
757 87. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive
758 score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
759 88. Saier, M. H. et al. The Transporter Classification Database (TCDB): recent advances. *Nucleic*
760 *Acids Res*. **44**, D372–D379 (2016).
761 89. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity
762 searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
763 90. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation.
764 *Nucleic Acids Res.* **46**, W95–W101 (2018).
765 91. Terrapon, N., Lombard, V., Gilbert, H. J. & Henrissat, B. Automatic prediction of polysaccharide
766 utilization loci in Bacteroidetes species. *Bioinformatics* **31**, 647–655 (2015).
767 92. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features.
768 *Bioinformatics* **26**, 841–842 (2010).
769 93. Bushnell, B. BBMap. *https://sourceforge.net/projects/bbmap/* (2015).
770 94. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI
771 analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114
772 (2018).
773 95. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079
774 (2009).
775 96. Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative
776 metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.*
777 **16**, 51 (2015).

97. Castro, J. C. *et al*. imGLAD: accurate detection and quantification of target organisms in metagenomes. *PeerJ* **6**, e5882 (2018).
98. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
99. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
100. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. **27**, 824–834 (2017).
101. Huntemann, M. *et al*. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand. Genomic Sci*. **11**, 17 (2016).
102. Caporaso, J. G. *et al*. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci*. **108**, 4516–4522 (2011).
103. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
104. Callahan, B. J. *et al*. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
105. Alishum, A. DADA2 formatted 16S rRNA gene sequences for both bacteria & archaea. (2020) https://doi.org/10.5281/zenodo.3951383.
106. Oksanen, J. *et al.* vegan: an R package for community ecologists. (2020).
107. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
108. Fitzpatrick, M. C. *et al.* gdm: Generalized dissimilarity modeling. (2021).
109. R Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* (2021).
110. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).

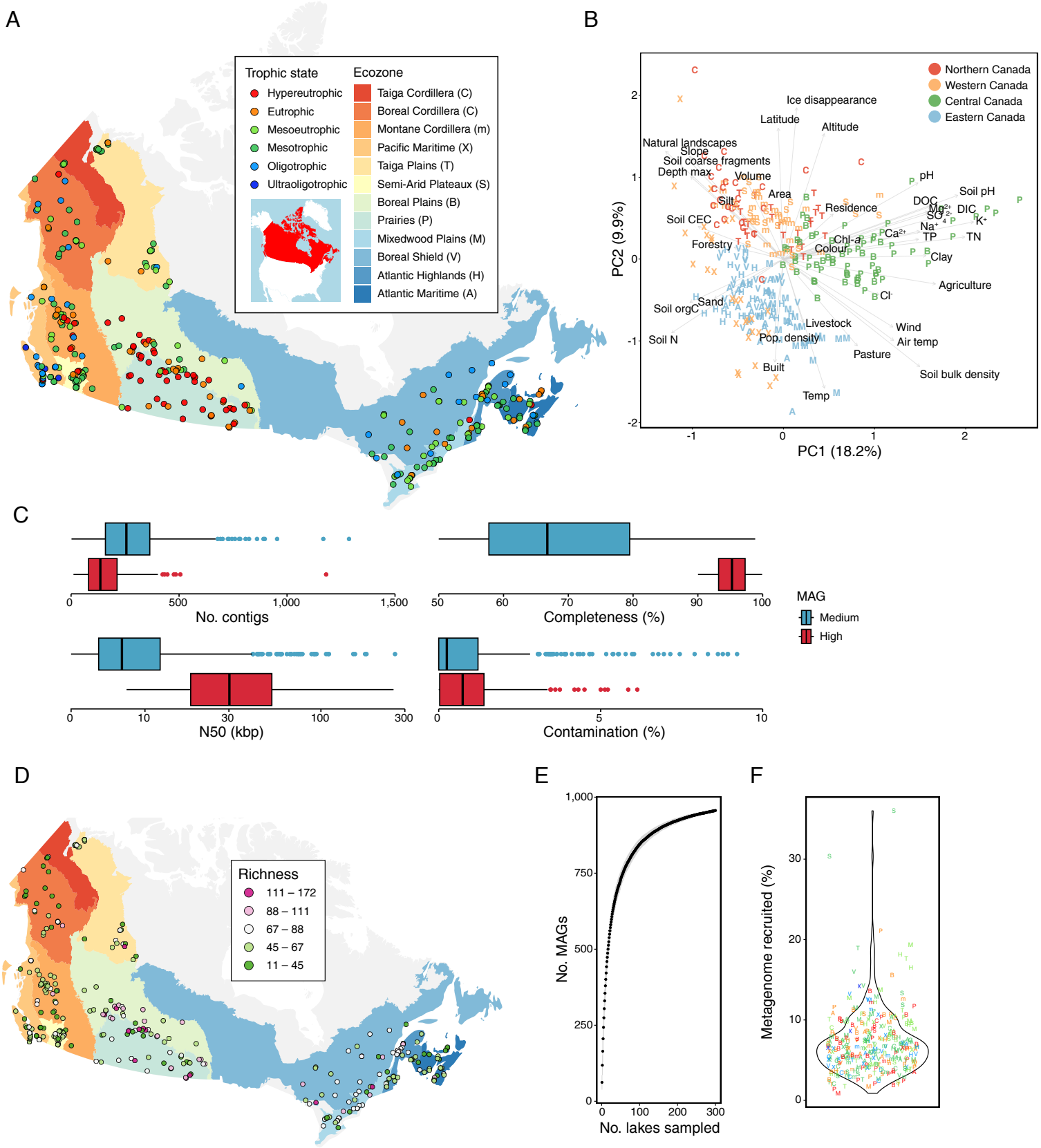
Figure legends

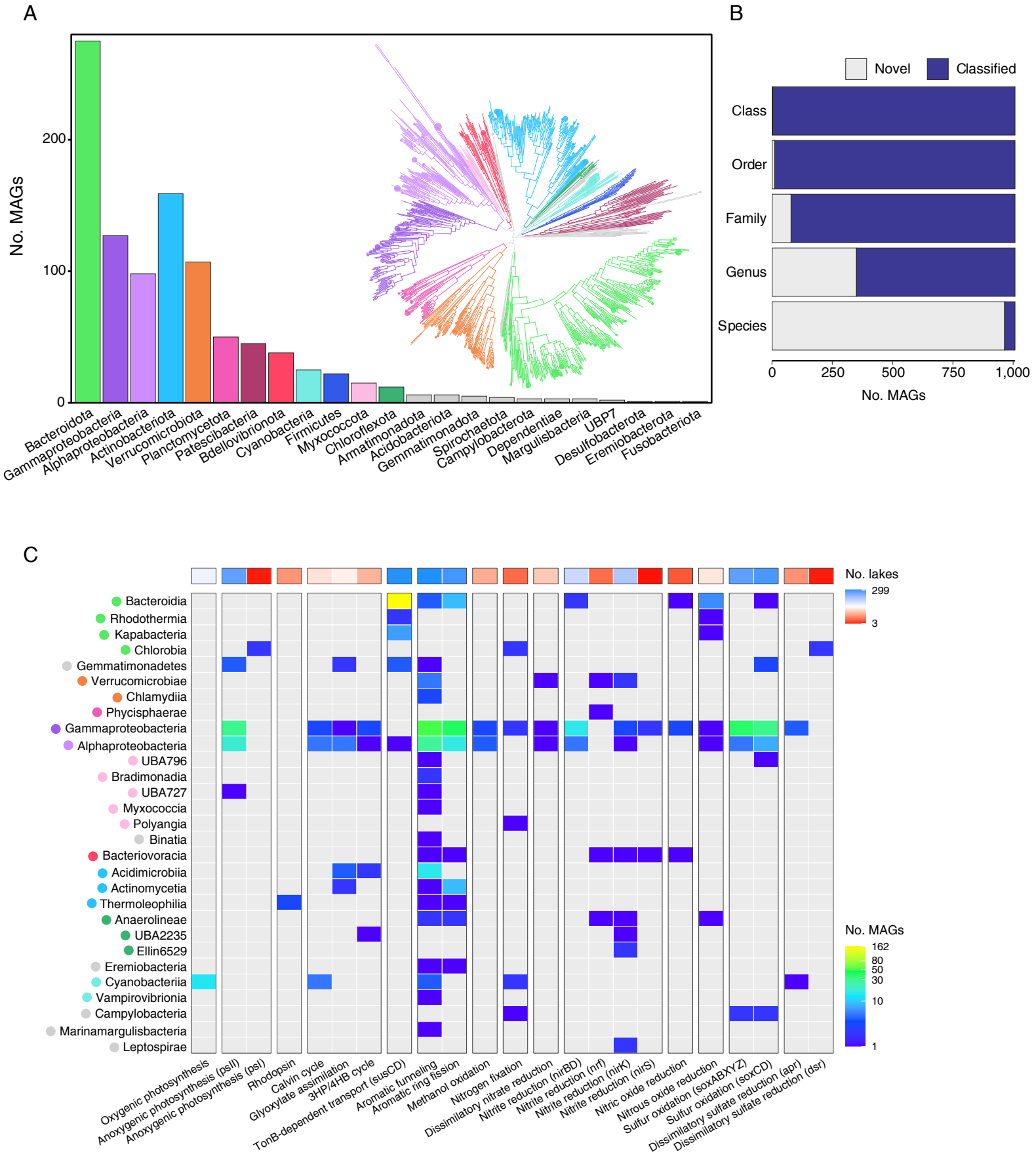
Figure 1. The LakePulse MAG catalogue: geographical and environmental scope, quality, and patterns of local-scale diversity. **(A)** Map of 308 sampled lakes from across a broad trophic state gradient (shown as point colours) in 12 Canadian ecozones (shown as base map colours). **(B)** Principal component analysis illustrating the variation in environmental features among sampled lakes. The letter symbols (see panel 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. **(D)** MAG richness within lakes. Point colours represent richness, which is the number of MAGs detected

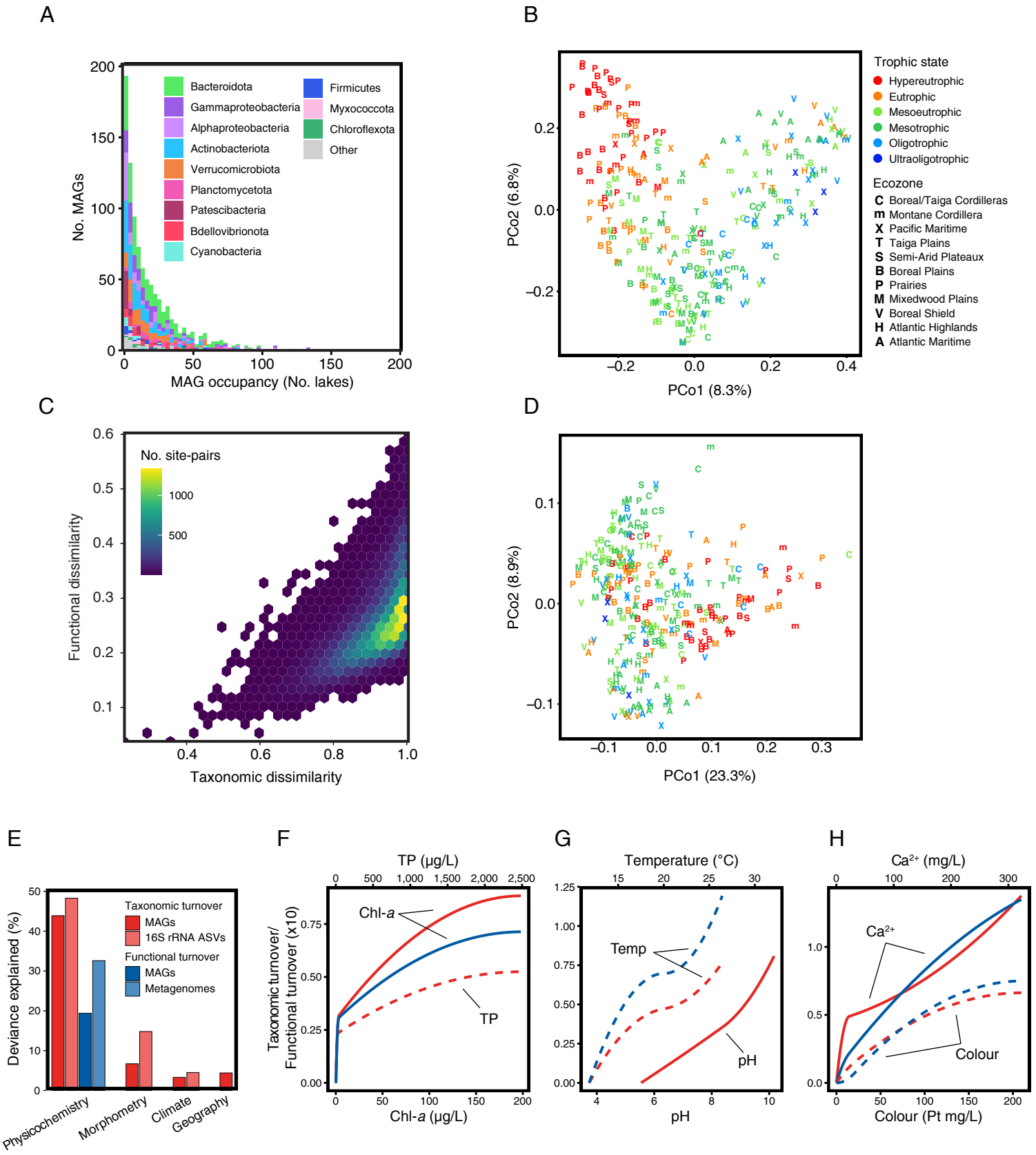
815 at each site based on positive TAD_{80} values (see panel A for ecozone colour legend). **(E)** MAG
816 accumulation curve (i.e. the number of new genomes detected) in a random ordering of lakes. **(F)**
817 Fraction of LakePulse metagenomes captured within MAGs (measured as the percentage of each
818 metagenome recruited to the MAGs). Point colours represent lake trophic states and letters represent
819 ecozones (see legend of panel A).
820
821 **Figure 2**. Taxonomic, phylogenetic, and functional diversity of the LakePulse MAG catalogue. **(A)**
822 Number of MAGs assigned to each phylum and phylogenetic tree of MAGs constructed from 120
823 bacterial marker genes. **(B)** Taxonomic novelty of the LakePulse MAG catalogue. Bar plots are divided by
824 colour to show the number of novel (i.e. unclassified) and classified MAGs from class to species level. **(C)**
825 Functional diversity across taxonomic orders. The lower heat map shows the number of MAGs within
826 each class containing genetic markers for a selection of metabolic functions (grey = no evidence; point
827 colours next to class names indicate the phylum-level affiliation of MAGs shown in panel A). The upper
828 heat map shows the number of lakes containing MAGs with evidence of metabolic functions.
829
830 **Figure 3**. Niche breadth and environmental filtering of MAGs based on taxonomic and functional
831 biogeography. **(A)** Frequencies of MAG incidence (i.e. the number of lakes occupied by each MAG).
832 Colours show MAG phylum-level assignments. **(B)** Principal coordinate analysis (PCoA) showing the
833 variation in MAG taxonomic composition among lakes based on Jaccard dissimilarities. Point colours
834 represent lake trophic state and letter symbols represent ecozones. **(C)** Relationship between functional
835 and taxonomic diversity among MAG assemblages in different lakes. The colour gradient shows the
836 density of pairwise comparisons. **(D)** PCoA showing the variation in functional gene content among MAG
837 assemblages based on Jaccard dissimilarities. Point colours and letter symbols are described in the
838 legend of panel B. **(E)** Percent deviance explained by generalized dissimilarity models (GDMs) fitting the
839 responses of taxonomic or functional MAG assemblages or community profiles to physicochemical,
840 morphometric, geographic, or climatic gradients. (**F-H)** Partial effects of **(F)** total phosphorus (TP) and
841 chlorophyll-*a* (Chl-*a*), **(G)** temperature and pH, and **(H)** calcium and lake colour on the taxonomic and
842 functional turnover of MAG assemblages across lakes. Curve shapes represent the rate of turnover and
843 the maximum height of curves represents the magnitude of turnover across environmental gradients. Line
844 colours represent different MAG assemblage response data (red = taxonomic turnover, blue = functional
845 turnover) and line types (solid or dashed) represent different environmental explanatory variables.
846
847 **Figure 4**. Diversity of specific bacterial functions among lake assemblages and individual MAGs. **(A)**
848 Percent deviance explained by generalized dissimilarity models (GDMs) fitting the responses of specific
849 MAG assemblage functions to physicochemical, morphometric, geographic, or climatic gradients. **(B)**
850 Partial effects of chlorophyll-*a* and lake colour on the carbohydrate and lipid metabolism turnover of MAG
851 assemblages across lakes. **(C)** PCoA showing the variation in carbohydrate metabolism among MAG

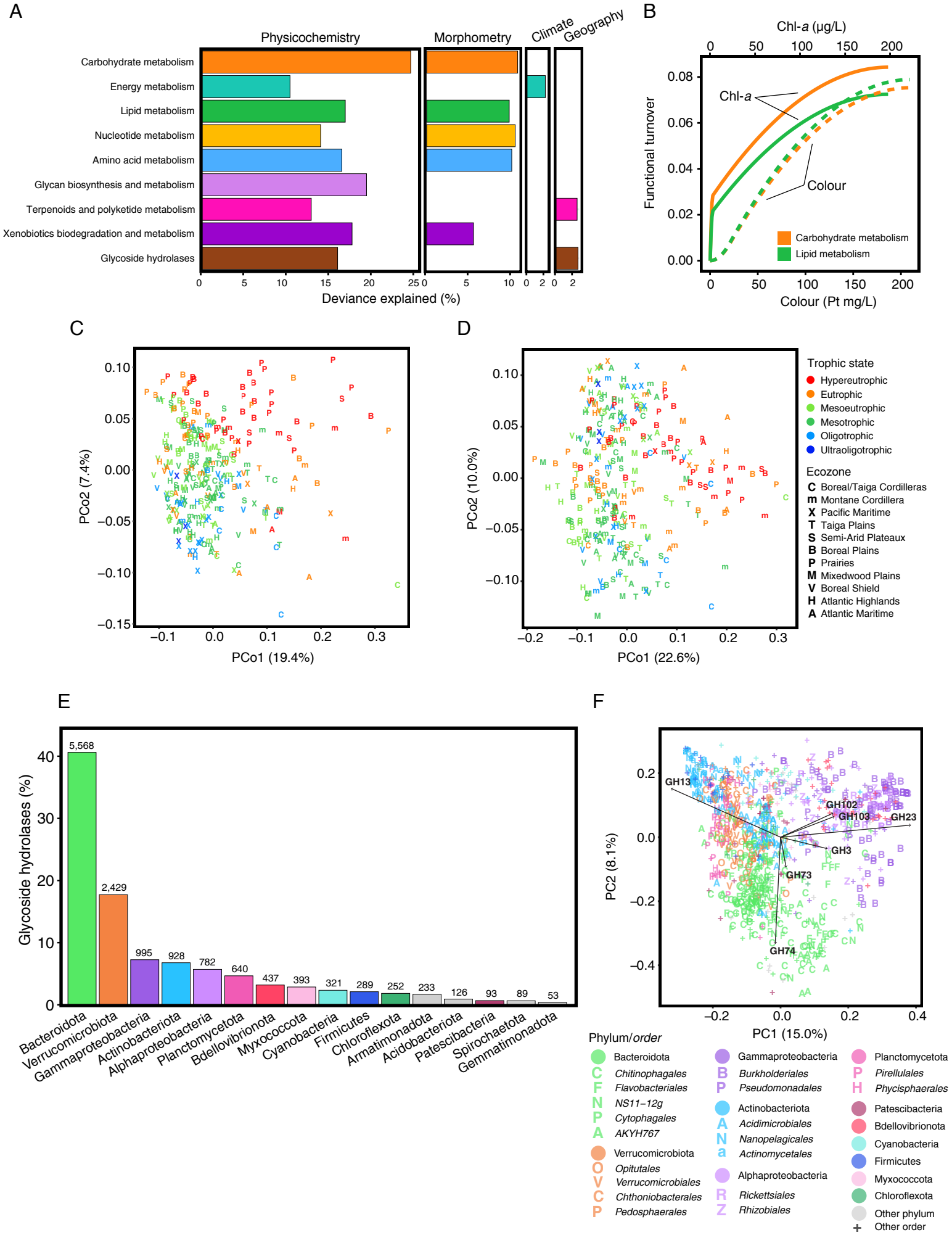
852 assemblages. **(D)** PCoA showing the variation in lipid metabolism among MAG assemblages. Point
853 colours in **C** and **D** represent lake trophic state and letter symbols represent ecozones. **(E)** Percentage
854 and number of glycoside hydrolase (GH) genes distributed across MAG phyla. **(F)** PCA showing the
855 variation in GH gene families among MAGs. Point colours represent MAG phyla and letter symbols
856 represent order-level taxonomy within phyla.
857
858 **Figure 5**. Influences of watershed soil and land use on lake bacterial assemblages. **(A)** Percent deviance
859 explained by generalized dissimilarity models (GDMs) fitting the responses of taxonomic or functional
860 MAG assemblages or community profiles to watershed soil characteristics. **(B)** Partial effects of soil pH
861 and volumetric fraction of coarse fragments on the taxonomic and functional turnover of MAG
862 assemblages across lakes. **(C)** Map of proportion of land used for crop agriculture in lake watersheds. **(D)**
863 Map of human population densities in lake watersheds. **(E)** Percent deviance explained by generalized
864 dissimilarity models (GDMs) fitting the responses of taxonomic or functional MAG assemblages or
865 community profiles to watershed land use characteristics. Partial effects of crop agriculture and human
866 population density **(F)** on the taxonomic and functional turnover of MAG assemblages across lakes and
867 **(G)** on the turnover of xenobiotics biodegradation and metabolism within MAG assemblages across lakes.
868
869 **Supplementary information**
870
871 **Figure S1**. Project workflow illustrating the generation and analysis of the LakePulse MAG catalogue.
872
873 **Figure S2**. Distributions of **(A)** geography, **(B)** lake morphometry, **(C)** watershed soil, **(D)** land use, **(E)**
874 climate, and **(F)** surface water physicochemistry variables. Colours represent ecozones and ecozone
875 medians are indicated by dashed lines.
876
877 **Figure S3**. Number of MAGs generated within each ecozone co-assembly.
878
879 **Figure S4**. Comparison of taxonomic diversity in MAG and 16S rRNA gene ASV datasets: percentages of
880 MAGs and ASVs assigned to **(A)** phyla and **(B)** orders.
881
882 **Figure S5**. Phylum-level biogeographic distributions of MAGs.
883
884 **Figure S6**. Hierarchical clustering (Ward's linkage) of MAG phylum-level taxonomic assemblages. Letter
885 symbols above each MAG assemblage (scaled to relative TAD_{80}) represent ecozones and coloured
886 squares represent lake chlorophyll-*a* concentration.
887
888 **Figure S7**. PCoAs of **(A)** the taxonomic variation among ASV assemblages based on Jaccard

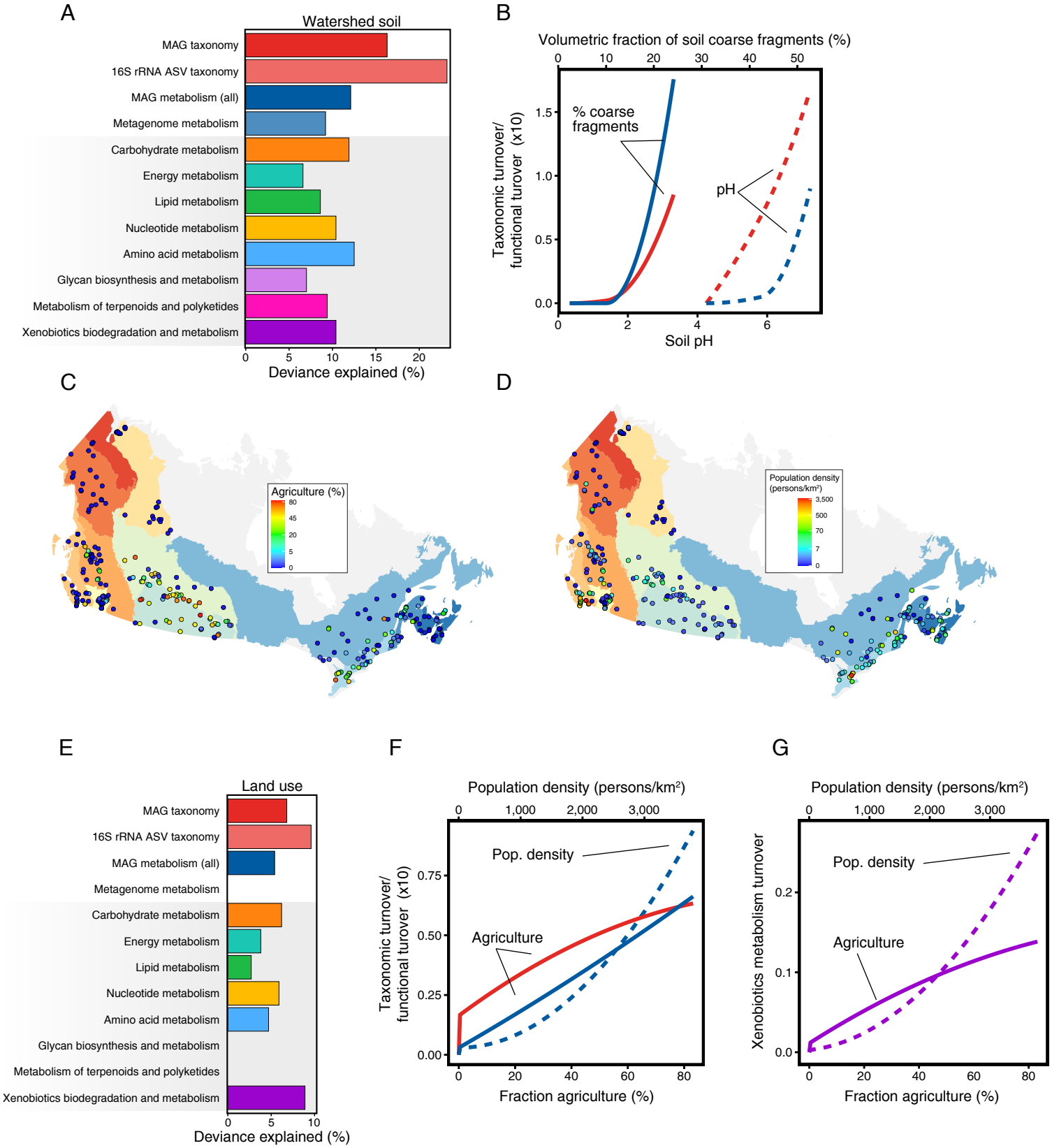
889 dissimilarities and **(B)** the variation in metabolic gene content among metagenomes based on Bray-Curtis
890 dissimilarities.
891
892 **Figure S8**. Significant predictors of **(A)** taxonomic and functional turnover across MAG assemblages and
893 community profiles and **(B)** turnover in specific metabolic functions across MAG assemblages. The height
894 of bars represents the relative importance of predictors within the generalized dissimilarity model. MAG
895 assemblage turnover is shown above zero and community (ASV or metagenome) turnover is shown
896 below zero.
897
898 **Figure S9**. Gene maps of polysaccharide utilization loci (PULs) identified across *Bacteroidota* MAGs.
899 Arrows indicate gene directions. Colours represent gene types (SusCD, CAZymes, tRNA genes, other
900 KOs).
901
902 **Figure S10**. Principal coordinate analysis (PCoA) showing the variation in xenobiotics biodegradation and
903 metabolism among MAG assemblages.
904
905 **Table S1**. Summary of metagenome information: lake names, accession information, sampling
906 coordinates, and assembly characteristics.
907
908 **Table S2**. Summary of MAG accession information, quality, genome characteristics, taxonomy,
909 associated file names, marker gene content, and TAD_{80} across 300 freshwater to oligosaline lakes.
910
911 **Table S3**. Number of novel MAGs in each phylum.
912
913 **Table S4.** Generalized dissimilarity modelling results for MAG assemblages and community (ASV and
914 metagenome) profiles across lakes based on taxonomic and functional composition.


A
Trophic state
Hypereutrophic
Eutrophic
Mesoeutrophic
Mesotrophic
Oligotrophic
Ultraoligotrophic
Ecozone
Taiga Cordillera (C)
Boreal Cordillera (C)
Montane Cordillera (m)
Pacific Maritime (X)
Taiga Plains (T)
Semi-Arid Plateaux (S)
Boreal Plains (B)
Prairies (P)
Mixedwood Plains (M)
Boreal Shield (V)
Atlantic Highlands (H)
Atlantic Maritime (A)
B
Northern Canada
Western Canada
Central Canada
Eastern Canada
PC2 (9.9%)
PC1 (18.2%)
Ice disappearance
Latitude
Altitude
Natural landscapes
Slope
Soil coarse fragments
Depth max
Volume
Area
Silt
Residence
Soil CEC
Forestry
pH
Soil pH
DOC
DIC
TP
TN
Chl-a
Colour
Clay
Agriculture
Soil orgC
Sand
Livestock
Wind
Air temp
Soil N
Pop. density
Pasture
Built
Soil bulk density
Temp
C
No. contigs
Completeness (%)
N50 (kbp)
Contamination (%)
MAG
Medium
High
D
Richness
111 – 172
88 – 111
67 – 88
45 – 67
11 – 45
E
No. MAGs
No. lakes sampled
F
Metagenome recruited (%)


A
No. MAGs
0
100
200
Bacteroidota
Gammaproteobacteria
Alphaproteobacteria
Actinobacteriota
Verrucomicrobiota
Planctomycetota
Patescibacteria
Bdellovibrionota
Cyanobacteria
Firmicutes
Myxococcota
Chloroflexota
Armatimonadota
Acidobacteriota
Gemmatimonadota
Spirochaetota
Campylobacterota
Dependentiae
Margulisbacteria
UBP7
Desulfobacterota
Eremiobacterota
Fusobacteriota
B
Novel
Classified
Class
Order
Family
Genus
Species
0
250
500
750
1,000
No. MAGs
C
No. lakes
299
3
No. MAGs
162
80
50
30
10
1
Bacteroidia
Rhodothermia
Kapabacteria
Chlorobia
Gemmatimonadetes
Verrucomicrobiae
Chlamydiia
Phycisphaerae
Gammaproteobacteria
Alphaproteobacteria
UBA796
Bradimonadia
UBA727
Myxococcia
Polyangia
Binatia
Bacteriovoracia
Acidimicrobiia
Actinomycetia
Thermoleophilia
Anaerolineae
UBA2235
Ellin6529
Eremiobacteria
Cyanobacteriia
Vampirovibrionia
Campylobacteria
Marinamargulisbacteria
Leptospirae
Oxygenic photosynthesis
Anoxygenic photosynthesis (pslI)
Anoxygenic photosynthesis (psI)
Rhodopsin
Calvin cycle
Glyoxylate assimilation
3HP/4HB cycle
TonB-dependent transport (susCD)
Aromatic funneling
Aromatic ring fission
Methanol oxidation
Nitrogen fixation
Dissimilatory nitrate reduction
Nitrite reduction (nirBD)
Nitrite reduction (nrf)
Nitrite reduction (nirK)
Nitrite reduction (nirS)
Nitric oxide reduction
Nitrous oxide reduction
Sulfur oxidation (soxABXYZ)
Sulfur oxidation (soxCD)
Dissimilatory sulfate reduction (apr)
Dissimilatory sulfate reduction (dsr)


A
Bacteroidota
Gammaproteobacteria
Alphaproteobacteria
Actinobacteriota
Verrucomicrobiota
Planctomycetota
Patescibacteria
Bdellovibrionota
Cyanobacteria
Firmicutes
Myxococcota
Chloroflexota
Other
No. MAGs
MAG occupancy (No. lakes)
B
Trophic state
Hypereutrophic
Eutrophic
Mesoeutrophic
Mesotrophic
Oligotrophic
Ultraoligotrophic
Ecozone
C Boreal/Taiga Cordilleras
m Montane Cordillera
X Pacific Maritime
T Taiga Plains
S Semi-Arid Plateaux
B Boreal Plains
P Prairies
M Mixedwood Plains
V Boreal Shield
H Atlantic Highlands
A Atlantic Maritime
PCo2 (6.8%)
PCo1 (8.3%)
C
No. site-pairs
1000
500
Functional dissimilarity
Taxonomic dissimilarity
D
PCo2 (8.9%)
PCo1 (23.3%)
E
Taxonomic turnover
MAGs
16S rRNA ASVs
Functional turnover
MAGs
Metagenomes
Deviance explained (%)
Physicochemistry
Morphometry
Climate
Geography
F
TP (µg/L)
Chl-a
TP
Taxonomic turnover/
Functional turnover (x10)
Chl-a (µg/L)
G
Temperature (°C)
Temp
pH
pH
H
Ca2+ (mg/L)
Ca2+
Colour
Colour (Pt mg/L)


A
Physicochemistry
Morphometry
Climate
Geography
Carbohydrate metabolism
Energy metabolism
Lipid metabolism
Nucleotide metabolism
Amino acid metabolism
Glycan biosynthesis and metabolism
Terpenoids and polyketide metabolism
Xenobiotics biodegradation and metabolism
Glycoside hydrolases
Deviance explained (%)
B
Chl-a (µg/L)
Functional turnover
Chl-a
Colour
Carbohydrate metabolism
Lipid metabolism
Colour (Pt mg/L)
C
PCo2 (7.4%)
PCo1 (19.4%)
D
PCo2 (10.0%)
PCo1 (22.6%)
Trophic state
Hypereutrophic
Eutrophic
Mesoeutrophic
Mesotrophic
Oligotrophic
Ultraoligotrophic
Ecozone
C Boreal/Taiga Cordilleras
m Montane Cordillera
X Pacific Maritime
T Taiga Plains
S Semi-Arid Plateaux
B Boreal Plains
P Prairies
M Mixedwood Plains
V Boreal Shield
H Atlantic Highlands
A Atlantic Maritime
E
Glycoside hydrolases (%)
5,568
2,429
995
928
782
640
437
393
321
289
252
233
126
93
89
53
Bacteroidota
Verrucomicrobiota
Gammaproteobacteria
Actinobacteriota
Alphaproteobacteria
Planctomycetota
Bdellovibrionota
Myxococcota
Cyanobacteria
Firmicutes
Chloroflexota
Armatimonadota
Acidobacteriota
Patescibacteria
Spirochaetota
Gemmatimonadota
F
PC2 (8.1%)
PC1 (15.0%)
GH13
GH102
GH103
GH23
GH3
GH73
GH74
Phylum/order
Bacteroidota
C Chitinophagales
F Flavobacteriales
N NS11–12g
P Cytophagales
A AKYH767
Verrucomicrobiota
O Opitutales
V Verrucomicrobiales
C Chthoniobacterales
P Pedosphaerales
Gammaproteobacteria
B Burkholderiales
P Pseudomonadales
Actinobacteriota
A Acidimicrobiales
N Nanopelagicales
a Actinomycetales
Alphaproteobacteria
R Rickettsiales
Z Rhizobiales
Planctomycetota
P Pirellulales
H Phycisphaerales
Patescibacteria
Bdellovibrionota
Cyanobacteria
Firmicutes
Myxococcota
Chloroflexota
Other phylum
+ Other order


A
Watershed soil
MAG taxonomy
16S rRNA ASV taxonomy
MAG metabolism (all)
Metagenome metabolism
Carbohydrate metabolism
Energy metabolism
Lipid metabolism
Nucleotide metabolism
Amino acid metabolism
Glycan biosynthesis and metabolism
Metabolism of terpenoids and polyketides
Xenobiotics biodegradation and metabolism
0 5 10 15 20
Deviance explained (%)
B
Volumetric fraction of soil coarse fragments (%)
0 10 20 30 40 50
% coarse fragments
pH
Taxonomic turnover/ functional turover (x10)
0.0 0.5 1.0 1.5
0 2 4 6
Soil pH
C
Agriculture (%)
80 45 20 5 0
D
Population density (persons/km²)
3,500 500 70 7 0
E
Land use
MAG taxonomy
16S rRNA ASV taxonomy
MAG metabolism (all)
Metagenome metabolism
Carbohydrate metabolism
Energy metabolism
Lipid metabolism
Nucleotide metabolism
Amino acid metabolism
Glycan biosynthesis and metabolism
Metabolism of terpenoids and polyketides
Xenobiotics biodegradation and metabolism
0 5 10
Deviance explained (%)
F
Population density (persons/km²)
0 1,000 2,000 3,000
Pop. density
Agriculture
Taxonomic turnover/ functional turover (x10)
0.00 0.25 0.50 0.75
0 20 40 60 80
Fraction agriculture (%)
G
Population density (persons/km²)
0 1,000 2,000 3,000
Pop. density
Agriculture
Xenobiotics metabolism turnover
0.0 0.1 0.2
0 20 40 60 80
Fraction agriculture (%)