

<sup>1</sup> **Supporting Information**

<sup>2</sup> **Upwelling periodically disturbs the ecological assembly of**  
<sup>3</sup> **microbial communities in the Laurentian Great Lakes**

<sup>4</sup> Augustus Pendleton<sup>1\*</sup>, Mathew Wells<sup>2</sup>, & Marian L. Schmidt<sup>1\*</sup>

<sup>5</sup> <sup>1</sup>Department of Microbiology, Cornell University, 123 Wing Dr, Ithaca, NY 14850, USA

<sup>6</sup> <sup>2</sup>Department of Physical and Environmental Sciences, University of Toronto Scarborough,  
<sup>7</sup> 1065 Military Trail, Toronto, Ontario M1C 1A4, Canada

<sup>8</sup> **Corresponding Authors:** Augustus Pendleton: arpt277@cornell.edu; Marian L.  
<sup>9</sup> Schmidt: marschmi@cornell.edu

<sup>10</sup> **Supplemental Methods**

<sup>11</sup> **Sample collection**

<sup>12</sup> Depth profiles were collected as the rosette/CTD descended the water column and water  
<sup>13</sup> samples were collected as the CTD ascended. Samples were collected without regard to  
<sup>14</sup> time of day (though Stn 33 was sampled both night and day in September). As such,  
<sup>15</sup> bottom samples were always filled first, and immediately taken to the lab for filtering.  
<sup>16</sup> For all samples, 5 mL of water was initially sub-sampled for flow cytometry samples and  
<sup>17</sup> kept in the fridge. First replicates at each depth were filtered immediately, while second  
<sup>18</sup> replicates were kept refrigerated until the first replicate was done. Water samples were  
<sup>19</sup> taken in duplicate from separate niskins and pre-filtered consecutively through sterilized  
<sup>20</sup> 200  $\mu$ m and 20  $\mu$ m Nitex mesh (Wildco) into sterile 10L bottles (Nalgene). Bottles  
<sup>21</sup> were rinsed twice with sample water before filling. Microbial samples were filtered using  
<sup>22</sup> MasterFlex Peristaltic Pumps onto a polyethersulfone 0.22  $\mu$ m filter (MilliporeSigma).  
<sup>23</sup> Samples were filtered for a maximum of 25 minutes, with flow through volumes ranging  
<sup>24</sup> from 1500 mL to 8760 mL. Filters were flash frozen in liquid N<sub>2</sub> and stored at -80C.

<sup>25</sup> **Environmental data**

<sup>26</sup> Water chemistry and chlorophyll a data were generated by the U.S. Environmental  
<sup>27</sup> Protection Agency's Great Lakes National Program Office according to their standard pro-  
<sup>28</sup> tocols [1]. Data included NH<sub>4</sub>, NO<sub>x</sub>, Soluble Reactive Phosphorus, Total Nitrogen, Total  
<sup>29</sup> Phosphorus, K, Na, Ca, Mg, Cl, SO<sub>4</sub>, Dissolved Organic Carbon, Si, and Chlorophyll-  
<sup>30</sup> a. Data was downloaded from the Great Lakes Environmental Database system, which  
<sup>31</sup> can be accessed through the Environmental Protection Agency Central Data Exchange  
<sup>32</sup> (<https://cdx.epa.gov/>). Temperature profiles from the Great Lakes Acoustic Telemetry  
<sup>33</sup> Observation System (GLATOS) moorings were collected and provided by the Ontario  
<sup>34</sup> Ministry of Natural Resources, Fisheries and Oceans Canada, the U.S. Fish and Wildlife  
<sup>35</sup> Service, U.S. Geological Survey, and Queen's University, supported in part by the Great  
<sup>36</sup> Lakes Observing System (GLOS). Data are publicly available through the GLOS Seagull  
<sup>37</sup> platform (<https://seagull.glos.org/data-console/groups/379>). Lake Erie water tempera-  
<sup>38</sup> tures at the Niagara River in Buffalo were downloaded from the National Weather Service  
<sup>39</sup> (NOAA, <https://www.weather.gov/buf/LakeErieMay>)

## 40 Microbial Sampling

41 Duplicate microbial samples from separate niskins but the same CTD/rosette cast were  
42 immediately processed on board after water collection. For the first replicate, there was  
43 a maximum time of 96 minutes and a median time of 35.5 minutes from the Niskin bottle  
44 being fired to the sample being frozen. Between replicates of the same station and depth,  
45 1 L of MilliQ was used to rinse the filtration tubing. Between stations, sample bottles  
46 and filtration tubing were sterilized by rinsing and incubating at room temperature for 10  
47 minutes with 10% (v/v) sodium hypochlorite and then rinsed three times consecutively  
48 with MilliQ 18.2 MΩ water. A negative control was prepared on the last day of each  
49 cruise using the same tubing sterilization procedure and running 4 L of MilliQ 18.2 MΩ  
50 water through a 0.22 μm filter. Read depth or richness was not correlated to time from  
51 sampling till freezing, water extracted, or DNA concentrations.

## 52 Quantifying cells with flow cytometry

53 Fixed samples were thawed for 20 minutes at 37C before staining with SYBR I green dye  
54 at a concentration of 1x SYBR I green at 37C for 20 minutes in the dark. Cell abundance  
55 was measured on a BD Accuri C6 Flow Cytometer. Sample volume was set at 50 uL,  
56 but actual volume uptake was estimated using CountBright Absolute Counting Beads  
57 (Thermofisher) to be  $42.5 \pm 0.5 \mu\text{L}$ . Samples were run at a flow rate of 33 uL/min, and  
58 fluorescence was measured using the blue laser FL1 filter. Raw events were filtered with a  
59 FSC-H filter of 100 and FL1-H filter of 400. Flow data was analyzed using the R packages  
60 flowCore and ggcryo [2, 3]. A polygon gate was defined to select for fluorescently-labeled  
61 cells compared to unstained controls, with some allowance of background fluorescence in  
62 unstained samples as many environmental cells are expected to be pigmented (Fig. S1).  
63 This gate is available within the Github repository associated with this manuscript.

## 64 DNA extraction & Illumina sequencing

65 All DNA extractions were carried out using the Qiagen DNeasy PowerWater kit.  
66 Briefly, samples were incubated at 65C in PW1 for 5 minutes, vortexed for 5 minutes,  
67 centrifuged at 4,000g for 1 minute, again incubated for 5 minutes, and vortexed for 5  
68 minutes. Otherwise, extractions were carried out as per manufacturer directions, and  
69 eluted with 40-50μl Elution Buffer. DNA yields were quantified using the dsDNA Broad-  
70 Range kit from Qubit and ranged from 0.302 ng/μL to 151 ng/μL. 149 samples were  
71 sequenced on an Illumina MiSeq using v2 chemistry with 2x250 paired end reads, 10%  
72 underloading and 20% PhiX spike-in. Cluster density was 692, with 15.72 million raw  
73 reads and 13.55 million reads passing filter and an average of 105.5 thousand reads per  
74 sample.

## 75 Microbial bioinformatic processing

76 Raw Illumina sequences were processed into amplicon sequence variants (ASVs) using  
77 the standard workflow in dada2 package in R [4]. Primers were trimmed and sequences  
78 were quality-controlled using the `dada2::filterAndTrim` command with `truncLen =`  
79 `c(240,22)` and `trimLeft = c(20,20)` with `maxEE = c(1,1)` and `maxN = 0`, based on  
80 quality plots. Merged ASVs were length filtered to 252 or 253 bp. A median of 64.1% of  
81 reads were retained.

82 Taxonomy was assigned using both the TaxAss freshwater database and the Silva 16S  
83 rRNA databases [5, 6]. If ASVs were classified using the curated freshwater taxonomy,  
84 that classification was retained; otherwise, the Silva classification was used. For the  
85 freshwater taxonomy, ASVs were classified using the `RunSteps_quickie.sh` script using  
86 the `FreshTrain15Jun2020silva138` database with default parameters [5]. For the Silva  
87 database, ASVs were classified using `dada2::assignTaxonomy` function, against the Silva  
88 non-redundant (99%) v138 database and `dada2::addSpecies` function using the Silva  
89 species v138.1 database [6].

## 90 Measuring community assembly processes with iCAMP

91 iCAMP quantitatively estimates the importance of various assembly mechanisms using  
92 measures of phylogenetic and taxonomic dissimilarity. As assembly processes likely vary  
93 within sub-populations of a given community, iCAMP first divides the community into  
94 smaller, monophyletic lineages, in which the relative importance of assembly processes  
95 are evaluated. Using our scaled environmental parameters (Supporting Information Fig.  
96 S3B), we tested the average phylogenetic signal present in bins across bin.size.limits (12,  
97 24, 36, 48, and 60) and ds values (0.1, 0.2, 0.3, 0.4, and 0.5) using Mantel Tests as per-  
98 formed with the `ps.bin` function. The average correlation between phylogenetic distance  
99 and environmental niche (`MeanR`) was highest with smallest bin sizes, but the highest  
100 percentage of bins with significant correlations was with bin size limits of 48 (3.9%). We  
101 opted to use a bin.size.limit of 24, as the percentage of significant bins was still compara-  
102 ble (3.4%) and more bins allowed us to explore relationships between a bin's taxonomy,  
103 abundance, and assembly processes at a higher resolution. Over-all findings (that homog-  
104 enizing selection and drift dominated community assembly) remained consistent at both  
105 bin size limits.

106 Selective processes were quantified first using  $\beta$  Net Relatedness Index ( $\beta$ NRI) and  
107 then dispersal processes were quantified using a modified Raup-Crick (RC) method based  
108 on the Bray-Curtis dissimilarity of relative abundances [7, 8]. However, after testing for  
109 the normality of null distributions (`null.norm`), most bins had non-normal null distri-  
110 butions, so a one-tail non-parametric confidence test was instead used, by using the  
111 `change.sigindex` function with argument `Confidence`. Significance of inferred assem-  
112 bly processes between each month/depth group was tested using the `icamp.boot` function  
113 with 1000 iterations; all comparisons shown in Fig. 3A were considered significant.

114 The relative importance of an assembly processes in explaining the turnover between  
115 two samples is then estimated as the sum of bins dominated by that process, weighted  
116 by those bins' abundance overall. To divide the relative contribution of each microbial  
117 Class to a given process, we accessed the `Class.maxNamed` column in the `Bin.TopClass`  
118 slot.

119 To estimate a bin's abundance and coefficient of variation, the absolute abundance of all  
120 ASVs in each bin was summed for each sample. The coefficient of variation was calculated  
121 by dividing the standard deviation of absolute abundances across samples by the mean  
122 absolute abundance of that bin across all samples. To facet bin-level processes between  
123 Phyla, bins were divided based on `Top.TaxonPhylum`, and labeled by `Top.TaxonGenus` if  
124 available; otherwise, the lowest taxonomic rank available was used.

125 **Supplemental Results and Discussion**

126 **Microbial Community Composition across the thermocline**

127 Microbial communities in Lake Ontario shared a core set of cosmopolitan taxa, but cer-  
128 tain Classes were closely tied to specific depth-month groups. Alphaproteobacteria (pre-  
129 dominantly LD12, a close relative of the marine SAR11 lineage), Actinobacteria (includ-  
130 ing acI-B1, acI-C2, acI-A7, acI-B1, acI-A6, acSTL-A1, acI-A5, and acI-A3), Bacteroidia  
131 (such as unknown species and bacI-A1, bac-IIA, and bacIII-A), Gammaproteobacteria  
132 (LD28 and PnecB, formerly Betaproteobacteria), and Verrucomicrobiae (mostly LD19,  
133 SH3-11, and unknown species) were ubiquitous across all samples (Figs. 2D & S9A-B).

134 Differential abundance analysis revealed distinct habitat preferences among taxa. Chlo-  
135 roflexi Anaerolineae, including an ASV sharing 99.6% identity with CL500-11 from Crater  
136 Lake [9], were enriched in Deep and shallow May samples (Fig. S12). Deep samples had  
137 elevated levels of nitrifiers, including the ammonia-oxidizing Archaea Nitrososphaeria and  
138 the nitrite-oxidizing Bacteria Nitrospiria. In contrast, Bacteroidia were more prevalent in  
139 shallow waters, with the largest populations in May. Acidimicrobia and Cyanobacteria  
140 (mostly *Cyanobium PCC-6307*) were prominent in shallow September, along with two  
141 potentially harmful cyanobacterial genera (Fig. S13A-B). *Dolichospermum NIES41* was  
142 almost 1% of the community (>60,000 cells/mL) at station 17 near the Welland Canal  
143 (Fig. S13C), while *Microcystis PCC-7914* was abundant across multiple stations, peak-  
144 ing at ~100,000 cells/mL at station 43 near Cobourg, Ontario (Fig. S13D). As these  
145 samples were pre-filtered through a 20 $\mu$ m mesh before sequencing, these values could  
146 underestimate the true abundance of these colonial cyanobacteria.

147 **The Upwelling Niche**

148 It is notable that Great Lakes upwelling events in a given location tend be ephemeral,  
149 lasting a few to several days and create concomitant downwelling events across the lake  
150 (Fig. S6, [10]). These upwelling events create pulses of nutrients and bottom-dwelling  
151 taxa which disturb the dispersal barrier created by the thermocline. However, upwelling  
152 events can propagate laterally as slow-moving Kelvin waves, and occur regularly (e.g. a  
153 few per month) in stratified Lake Ontario (Fig. S6, [10], [11]). In September, upwelling  
154 stations disproportionately featured unique taxa not observed elsewhere (Fig. S9) and  
155 therefore generating unique co-occurring ASV pairs, creating the opportunity for novel  
156 microbial interactions (Fig. S9). Accordingly, upwelling zones represent a consistent, if  
157 ephemeral, niche in Lake Ontario with unique microbial diversity,

158 **Selection in the Hypolimnion**

159 While selection was less influential in shaping the microbial hypolimnion community,  
160 selection in deep waters is likely intense for microbes who actively grow in this environ-  
161 ment, especially over long time scales. For example, biogeographic patterns of Chloroflexi  
162 C500-11, a ubiquitous and (presumably) active member of the hypolimnion, suggest  
163 strong purifying selection under hypolimnion conditions [12–14]. Similar results have  
164 suggested globally distributed freshwater lineages of nitrifying taxa, whose occurrence is  
165 almost exclusively hypolimnetic due to photoinhibition of nitrification [15]. Multi-annual  
166 time series or sequencing approaches which differentiate active from dormant taxa in the  
167 hypolimnion may make homogenizing selection more influential.

168 **Rarity in Lake Ontario**

169 While the distribution and assembly of rare bacteria is consistently more stochastic  
170 (see [16]), different assembly mechanisms can lead to unique “types” of rarity [17, 18].  
171 Four types of rarity have been proposed: permanently rare, permanently rare with pe-  
172 riodic distributions, transiently rare, and conditionally rare [18]. In Lake Ontario, few  
173 conditionally rare taxa were observed (Fig. S12A). Drift and dispersal limitation emerged  
174 as the dominant assembly mechanisms for taxa with low abundance. This suggests that  
175 most rare taxa in Lake Ontario are either consistently rare with periodic distributions or  
176 transiently rare [18]. This finding aligns with the absence of a distance-decay relation-  
177 ship in rare taxa, as analyzed using unweighted UniFrac dissimilarity, which gives equal  
178 weight to both abundant and rare taxa (Fig. S14C & S14E). These results indicate that  
179 dispersal limitation signals in rare bacterioplankton likely reflect drift-driven differenti-  
180 ation across temporarily separated parcels of water, rather than linear range dispersal  
181 limitation [19].

182 **An expanded picture of microbial diversity in Lake Ontario**

183 Previous studies of Lake Ontario microbial diversity focused on a few offshore stations  
184 [20–23] or highly impacted embayments [24]. Our findings confirm dominant taxa like  
185 the *acI* and *LD12* lineages throughout the lake, the oligotrophic heterotroph *CL500-11*  
186 in the hypolimnion, and non-toxic *Synechococcus* in summer surface waters. Notably, we  
187 observed an abundance of Bacteroidia (*Flavo-A*, *bacI*, and *bacII*) in spring and a previ-  
188 ously unreported diversity of Acidomicrobiia (*acIV* lineages like *Iluma-A1*), expanding  
189 known microbial diversity of the lake. Though, these differences may reflect primer bias  
190 or the effects of a different pre-filtration strategy (see *Supplemental Discussion: Biases in*  
191 *Sampling and Sequencing*). Filamentous Bacteroidia may dominate in spring by averting  
192 grazing [25, 26]. Additionally, we detected the *acI-C2* lineage, characteristic of the lower  
193 Great Lakes [20], which was significantly more abundant in shallow September (Fig. S15),  
194 suggesting niche selection to the relatively warmer lower lakes [27].

195 **Biases in iCAMP**

196 These data, while novel in spatial breadth, rely on discerning patterns rather than  
197 directly measuring processes and depend on statistical assumptions, such as niche par-  
198 titioning being phylogenetically conserved and dispersal being random [7, 8]. However,  
199 dispersal, though modeled as a stochastic process, can be determined selectively for a  
200 given microbial clade [28, 29]. As selection is evaluated first by iCAMP, this approach  
201 may bias results toward detecting selective processes, potentially underestimating disper-  
202 sal’s role in community assembly [8].

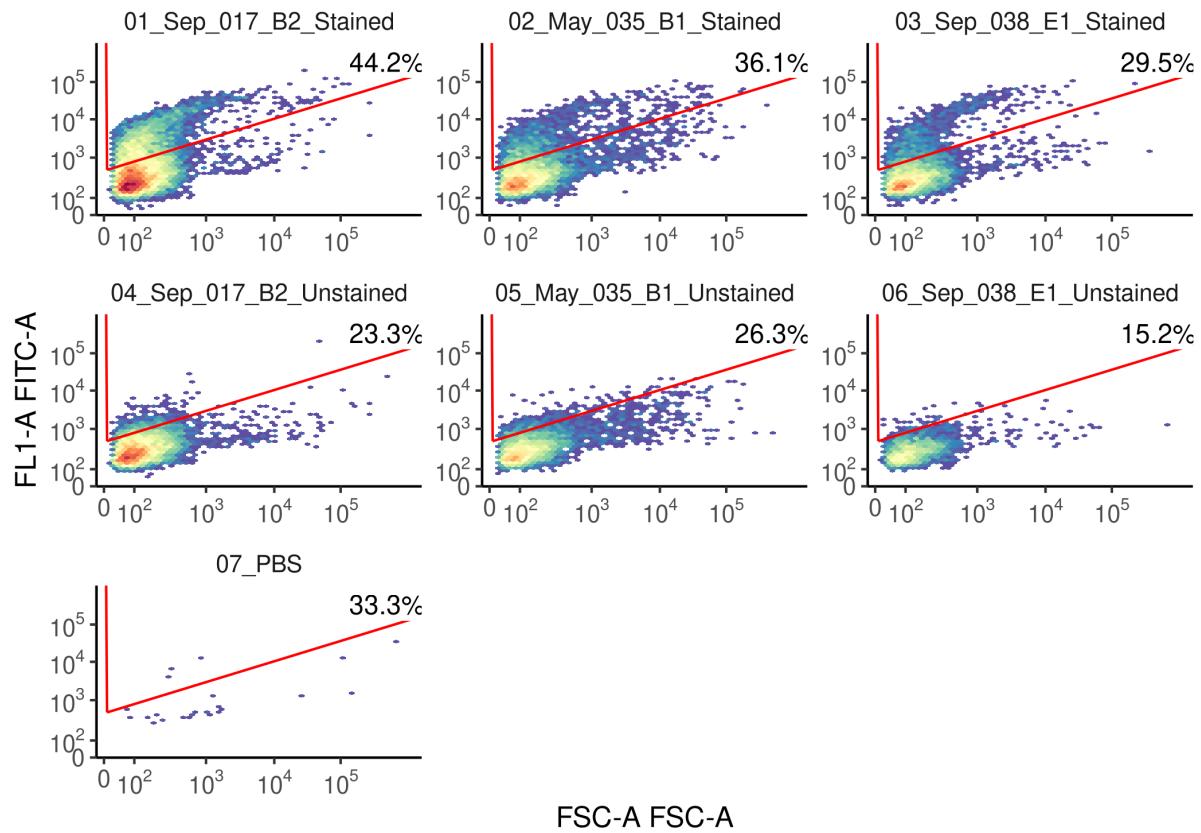
203 **Biases in Sampling and Sequencing**

204 Choices of pre-filtration and sequencing strategy are consequential to the measure-  
205 of microbial diversity. Even at 20 µm pre-filtration, many particle-associated, colonial,  
206 and filamentous bacteria are likely removed from our dataset. In addition, all primers  
207 used in metabarcoding amplification will show some bias [30]. These two factors could  
208 explain some of the differences in our dataset compared to [20]. Their more stringent pre-  
209 filtration at 1.6 µm may have reduced the abundance of large or filamentous taxa, like the  
210 Bacteroidetes and Cyanobacteria (outside of *Cyanobium*). In addition, they tested both

211 V4 and V4-V5 16S rRNA region primer sets, and opted to use the V4-V5 as it provided  
212 greater genomic resolution to differentiate taxa. As such, we may have underestimated  
213 the abundance of Alphaproteobacteria *LD12* and Actinobacteria *aci-B*, while the V4-V5  
214 primer set seemed biased against Acidimicrobiia *aciV* lineages, which were prominent in  
215 our dataset.

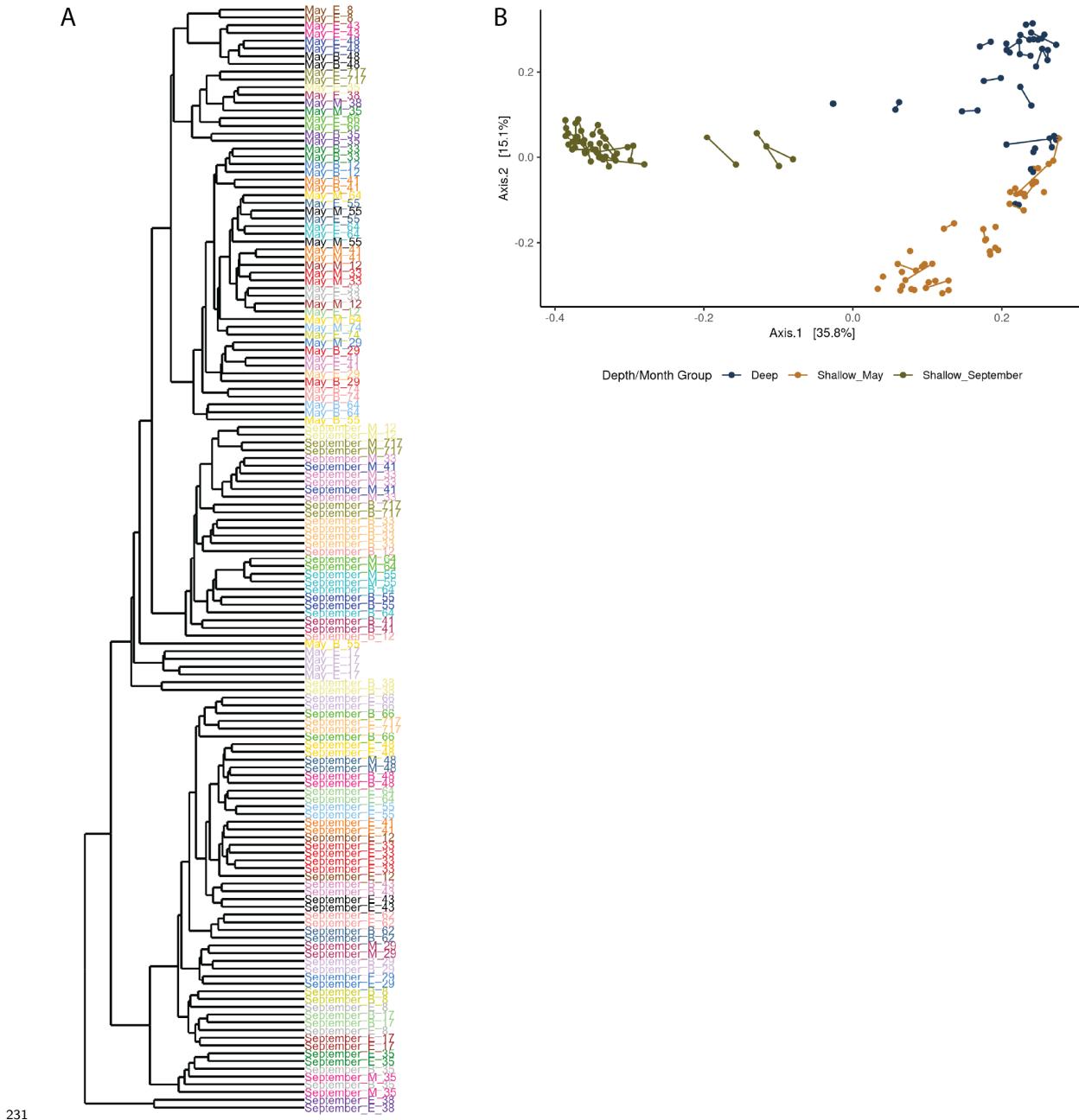
216 In addition, while our dataset is a valuable tool to probe “bottom-up” controls on  
217 microbial dynamics, but we did not consider important “top-down” controls through  
218 zooplankton grazing, which may be particularly relevant in oligotrophic lakes [31, 32].  
219 While simultaneously collected zooplankton samples were collected during the cruises, the  
220 methods used to collect them likely removed the smaller taxa who are the predominant  
221 microbial grazers like ciliates. DNA approaches like 18S rRNA metabarcoding on non-  
222 prefiltered samples would provide valuable insights into the relationships between micro-  
223 eukaryotic and prokaryotic plankton.

<sup>224</sup> Supplemental Figures

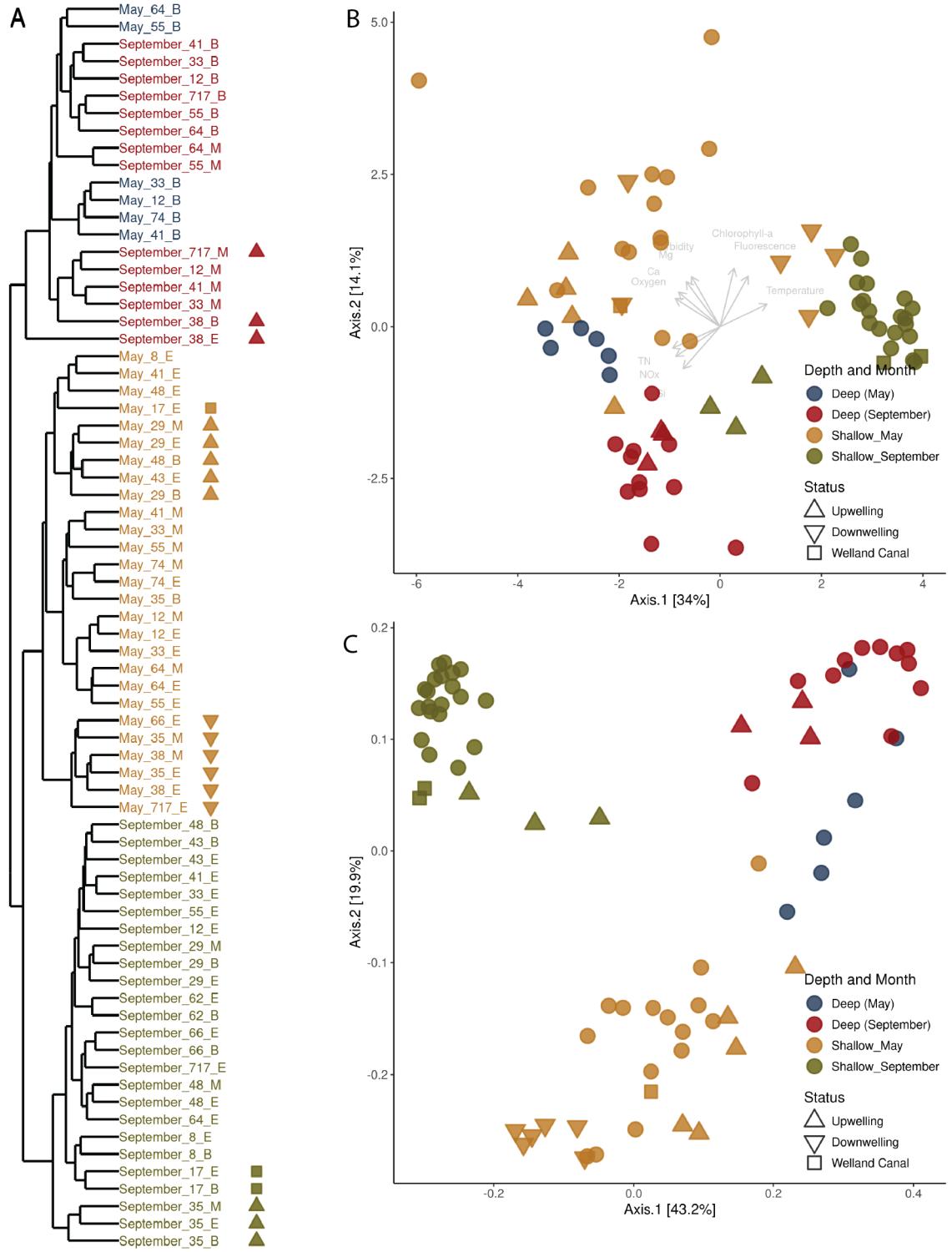


225

226 *Figure S1. Example Gating for SYBR-Positive Events.* Density plots generated showing green fluorescence  
 227 (with the blue laser FL1 filter on y-axis) versus forward scatter (FSC on x-axis) for three samples  
 228 stained with SYBR green (top row) versus unstained (middle row) and a PBS control (bottom row). Raw  
 229 events were filtered with a FSC-H filter of 100 and FL1-H filter of 400. The polygon gate which defines pos-  
 230 itive events (cells) is available within the Github repository @ “data/04\_cytometry\_exports/fcs\_files/”.

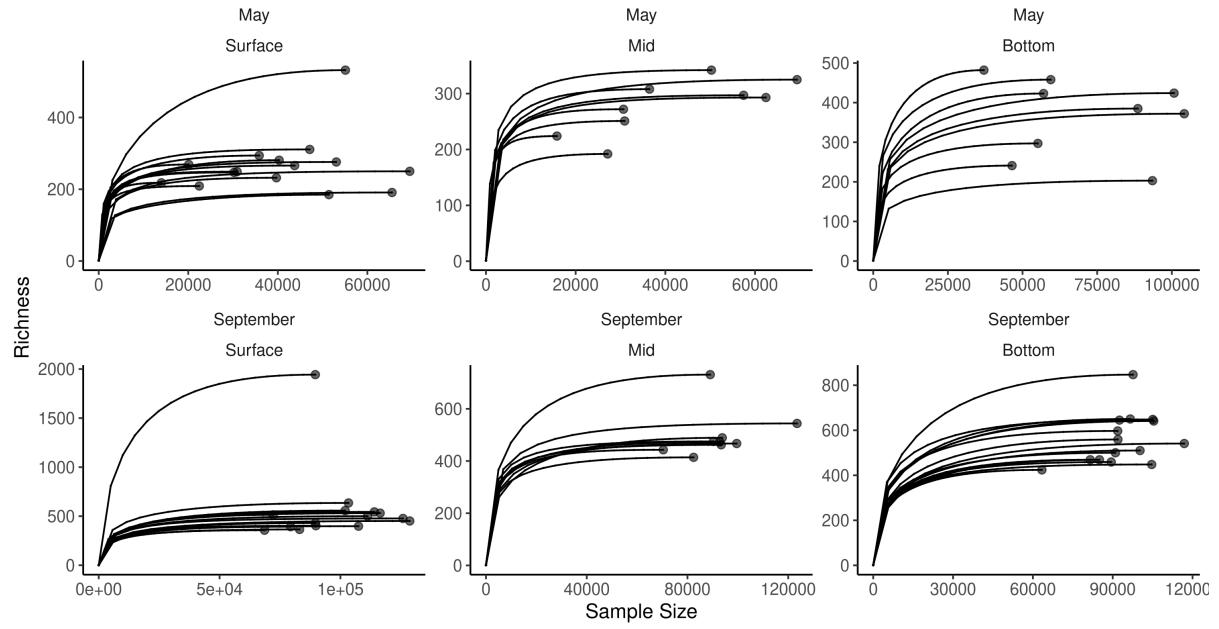


232 *Figure S2. Replicate samples had highly similar composition and were merged for following analyses.* (A)  
 233 Hierarchical clustering of all replicate samples using the abundance-unweighted version of the Bray-Curtis  
 234 dissimilarity, also known as the Sørensen Dissimilarity. Replicates of individual samples share colors; most  
 235 replicates are highly similar to each other. (B) Principal Coordinates Analysis of all replicate samples  
 236 using the Sørensen distance. Replicates are colored by their depth-month groups and connected via a  
 237 line.



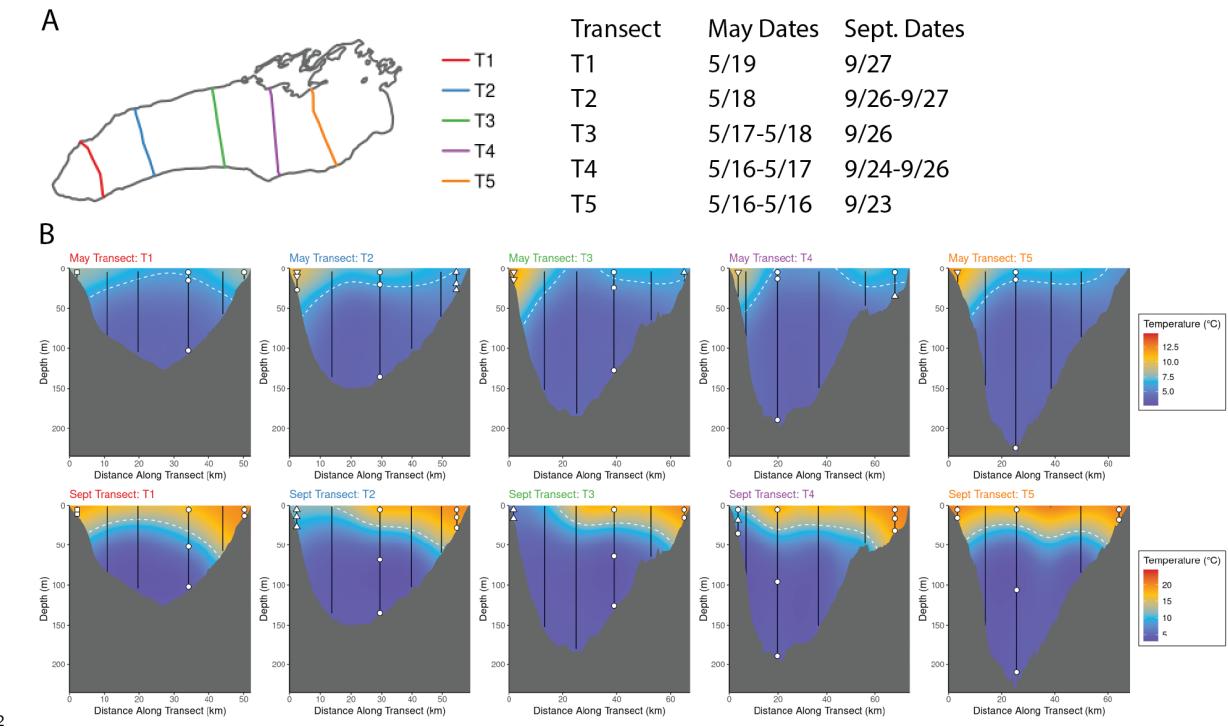
238

239 *Figure S3. Depth-Month Group Splits* (A) Month and depth groups were defined with the abundance-  
 240 weighted UniFrac dissimilarity using a UPGMA hierarchical clustering method, cutting the tree at 3  
 241 groups, though for illustration deep samples from May and September are differentiated by color. Sym-  
 242 bols denoting upwelling and downwelling stations and the Welland Canal are consistent across all three  
 243 panels. (B) PCA of physical and chemical parameters collected via CTD sensors and analyzed by the  
 244 EPA. All variables were scaled and centered before ordination. Only variables assessed as significant via  
 245 `vegan::envfit` ( $R^2 > 0.5$ ,  $p < 0.05$ ) are included. (C) PCoA of microbial community data, using the  
 246 abundance-weighted UniFrac distance as input, calculated with non-normalized absolute abundances of  
 247 each ASV. Shallow May had higher dispersion than shallow September (PERMDISP,  $p = 0.002$ ).



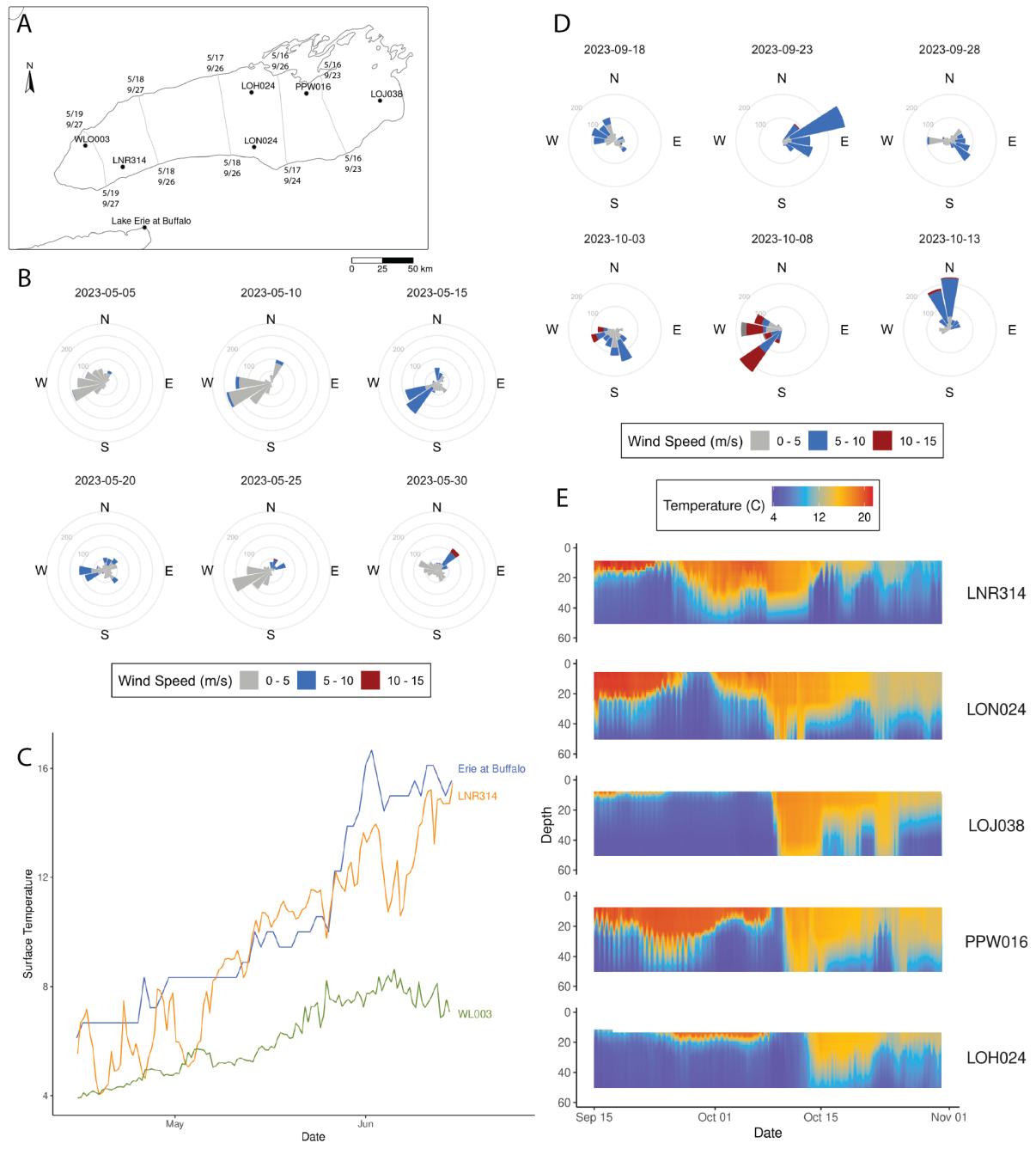
248

249 *Figure S4. Rarefaction curves of merged samples.* Rarefaction was performed using iNEXT on merged  
 250 replicate samples sub-sampling at forty “knots” within each sample. Curves terminate at the observed  
 251 richness.



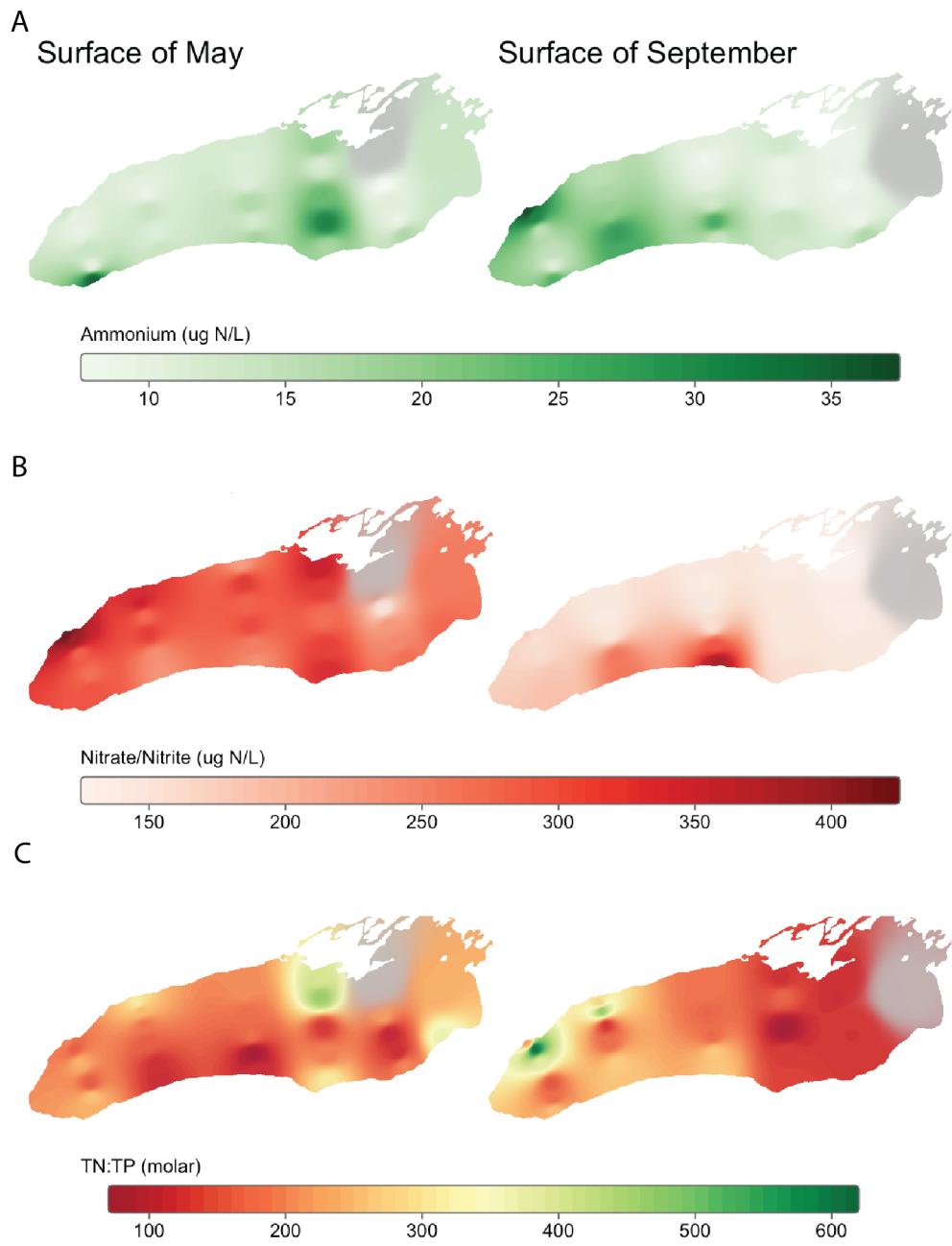
252

253 *Figure S5. Temperature Cross Sections of Lake Ontario in May and September* (A) Transects are labeled  
 254 1 through 5, moving eastward. (B) Temperature data from CTD casts was interpolated between stations  
 255 along a transect using multilevel B-splines in R across a 300x300 step grid for each transect with 5  
 256 hierarchical levels. The grey bottom reflects bathymetry of the lake, moving along the transect from  
 257 the southern shore to the northern shore. A 6°C isotherm is included in May, and a 12°C isotherm  
 258 in September. Sampling locations are noted via white symbols; upwelling stations are upward-facing  
 259 triangles, downwelling stations are downward-facing triangles, and the station nearest the Welland Canal  
 260 is a square. Note that temperature scales are different between May (top row) and September (bottom  
 261 row).



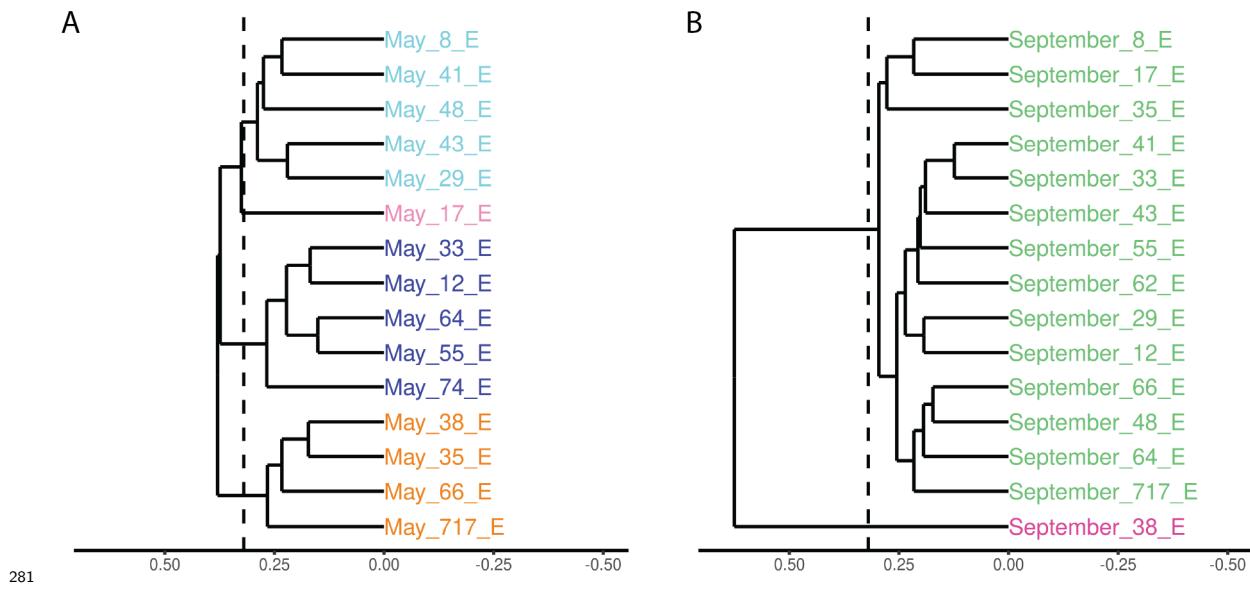
262

263 *Figure S6. Meteorological and Hydrodynamic Time Series Structuring Surface Temperature during Sam-*  
 264 *pling. (A) Map of GLATOS (in Lake Ontario) and National Weather Service temperature loggers. (B)*  
 265 *Binned wind roses in the days leading up to and following sampling in May. (C) Time series of tem-*  
 266 *perature moorings in the days leading up to and following sampling in May. The temperature sensor closest*  
 267 *to 12 meters was selected for WL003 and LNR314; the Lake Erie at Buffalo temperature sensor is 30m*  
 268 *deep. (D) Binned wind roses in the days leading up to and following sampling in September. (E) Time*  
 269 *series of temperature moorings in the days leading up to and following sampling in September. Mooring*  
 270 *depth was capped at 50m. Temperatures were averaged in 2-hour bins. Vertical temperature profiles*  
 271 *were linearly interpolated at each timepoint. For wind roses: All observations were drawn from NOAA*  
 272 *buoy 45012. Bars represent the number of 10-minute observations of wind at a given intensity and*  
 273 *direction. Dates above each rose correspond to the start day of a five-day period over which observa-*  
 274 *tions are summarized.*

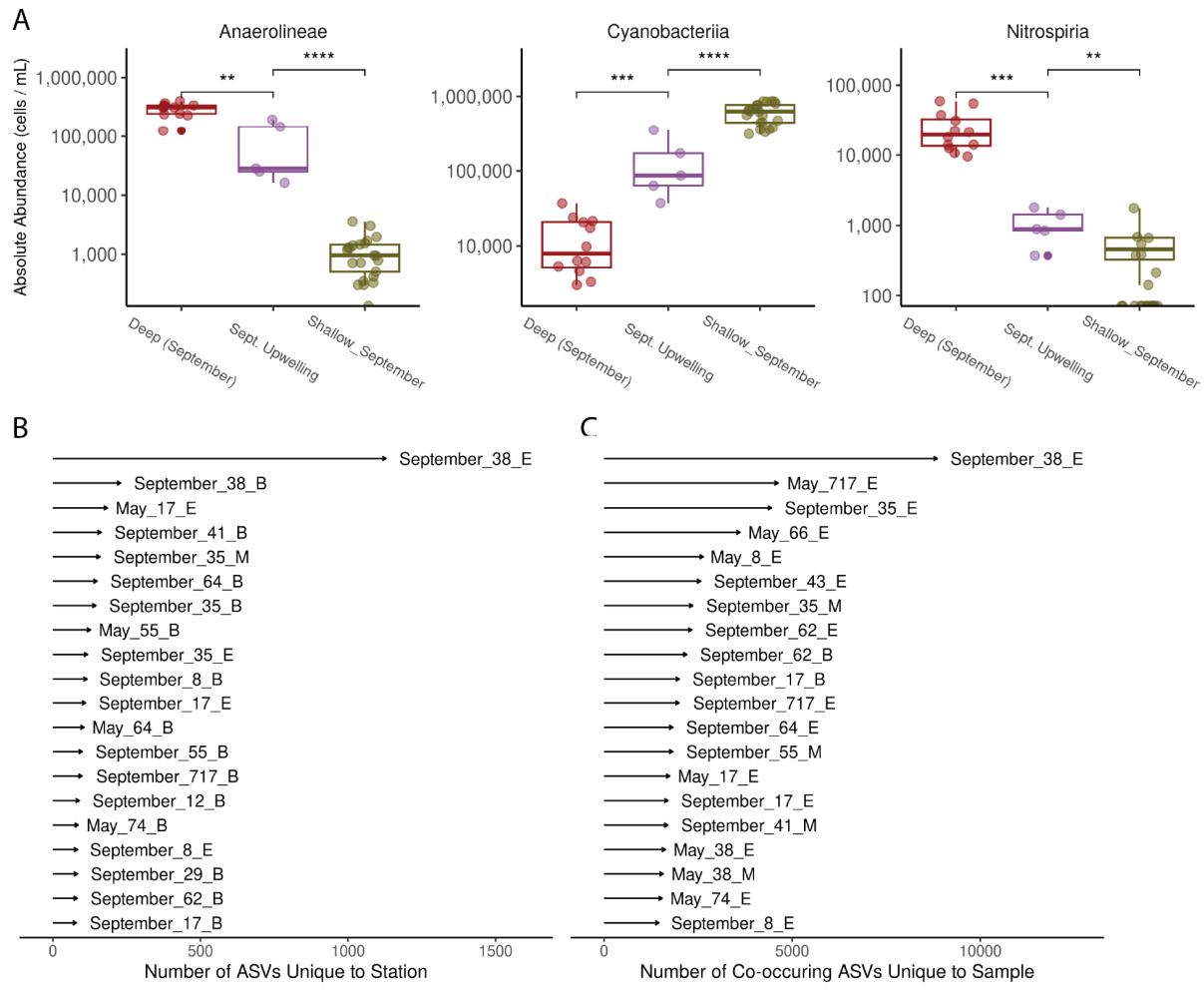


275

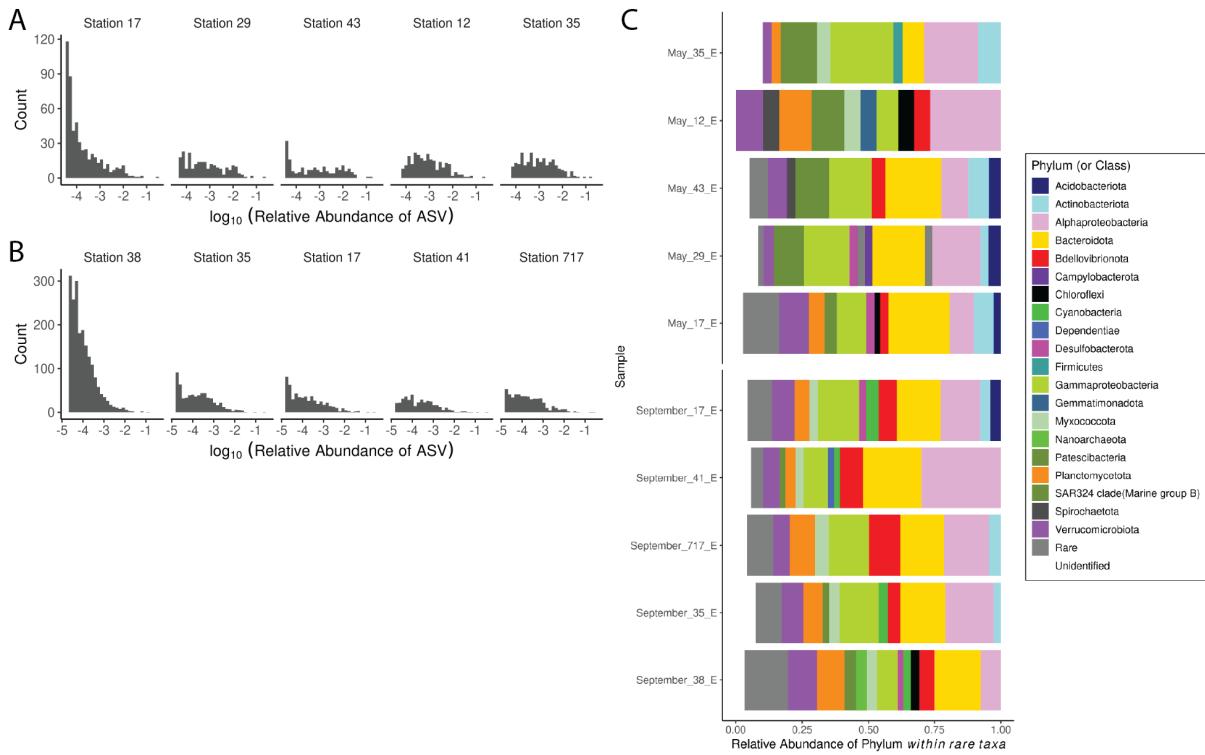
276 *Figure S7. Seasonal and Spatial variations of surface nitrogen species across Lake Ontario.* Multilevel  
 277 B-spline interpolation of measured concentrations of (A) Ammonium ( $\text{NH}_4$ ) (B) Nitrate and Nitrite  
 278 ( $\text{NO}_x$ ) and (C) the molar ratio of Total Nitrogen to Total Phosphorus (TN:TP) in surface samples in  
 279 May and September. Interpolations were performed using inverse-distance weighted interpolation in R,  
 280 with a power of 5.



282 *Figure S8. Hierarchical clustering of surface communities* UPGMA clustering of surface communities in  
 283 (A) May and (B) September using the community dissimilarity. Trees were cut to form clusters (displayed  
 284 in Fig. 1D) at a height of 0.32 (dashed line). Axes are consistent between plots.

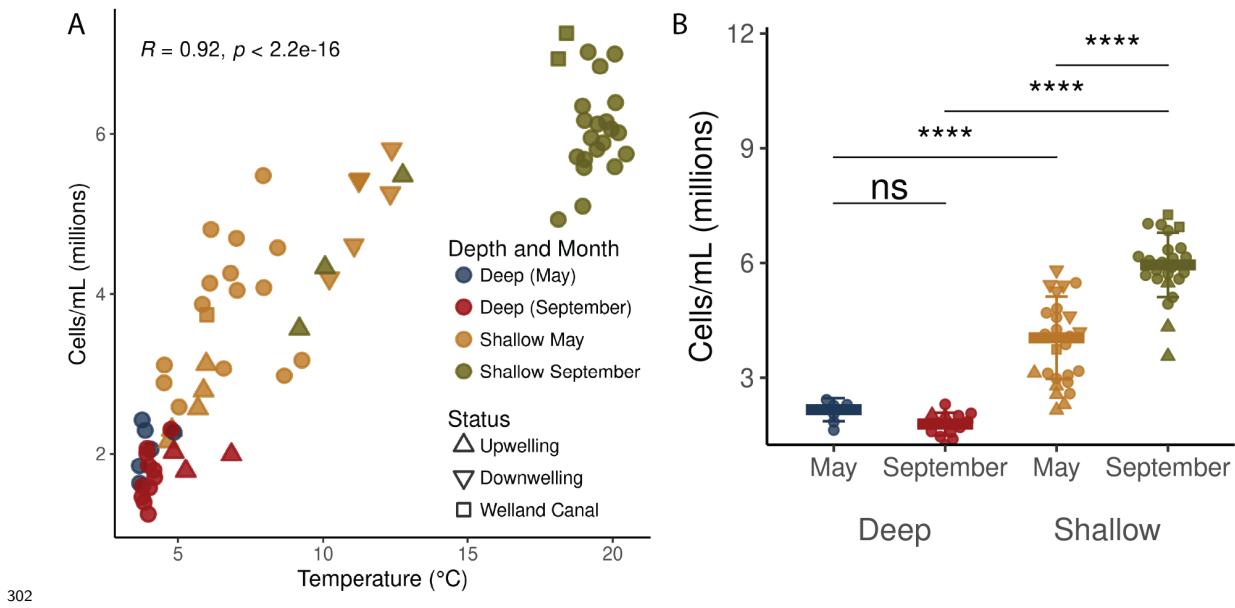


286 *Figure S9. Unique Taxa in the September Southern Upwelling Stations.* (A) Absolute abundances of  
 287 three Classes of Bacteria in September deep, shallow, and upwelling stations. Here, stations 35 and 38  
 288 are used as September upwelling stations. Statistical tests represent results of Two-Sample Wilcoxon  
 289 tests with Bonferroni correction (\* =  $p < 0.05$ , \*\* =  $p < 0.01$ , \*\*\* =  $p < 0.001$ , \*\*\*\* =  $p < 0.0001$ ). (B)  
 290 The number of ASVs that are unique to a given sample. Any ASV which was only observed in a single  
 291 sample was counted. The top twenty samples with the most unique ASVs are shown. (C) Potential  
 292 number of interactions between unique ASVs to a given sample. Only ASVs with an absolute abundance  
 293 greater than 1,000 cells/mL were kept. Unique ASV pairs were counted for each sample. Top twenty  
 294 samples with the most unique ASV pairs are shown.

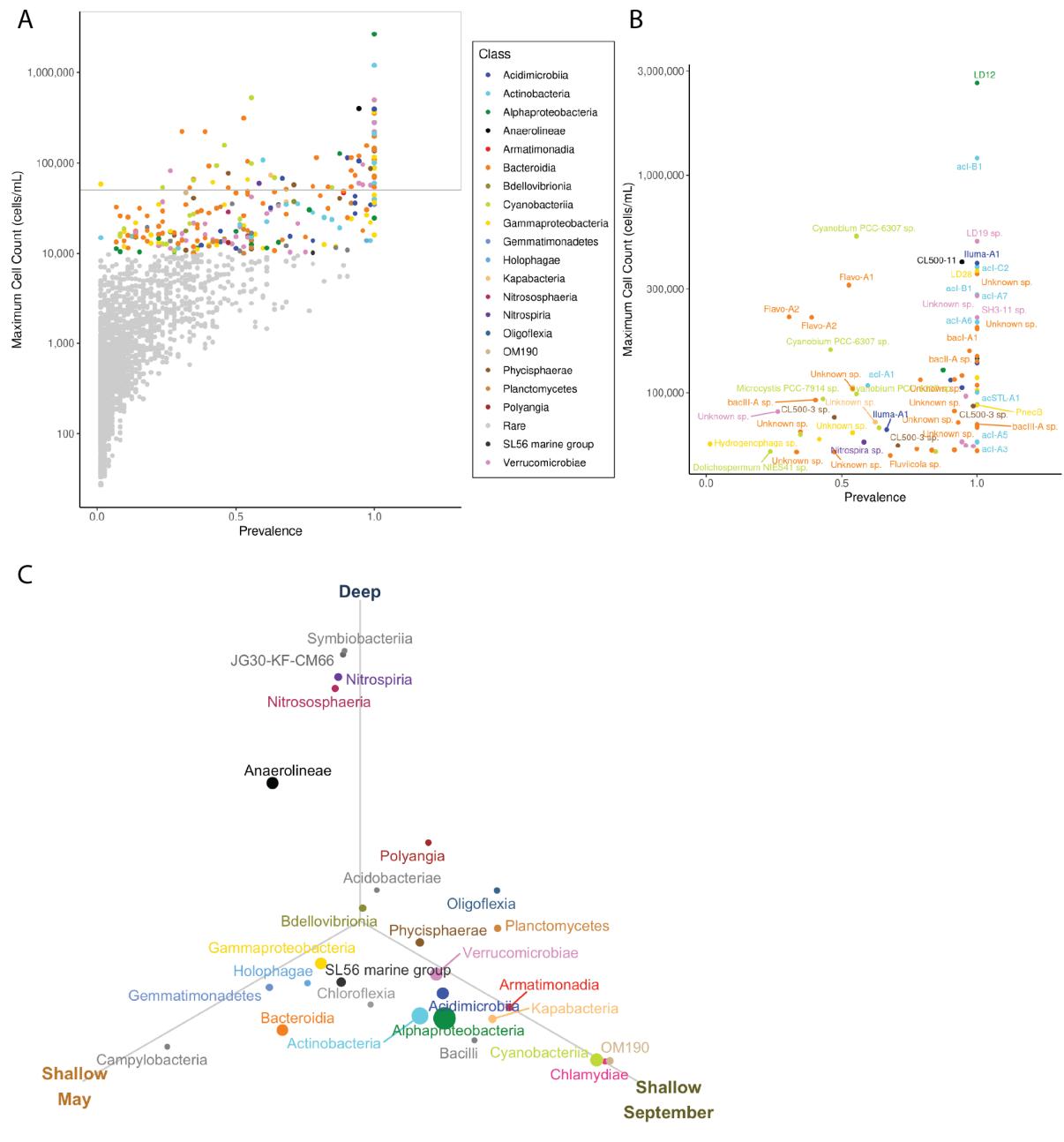


295

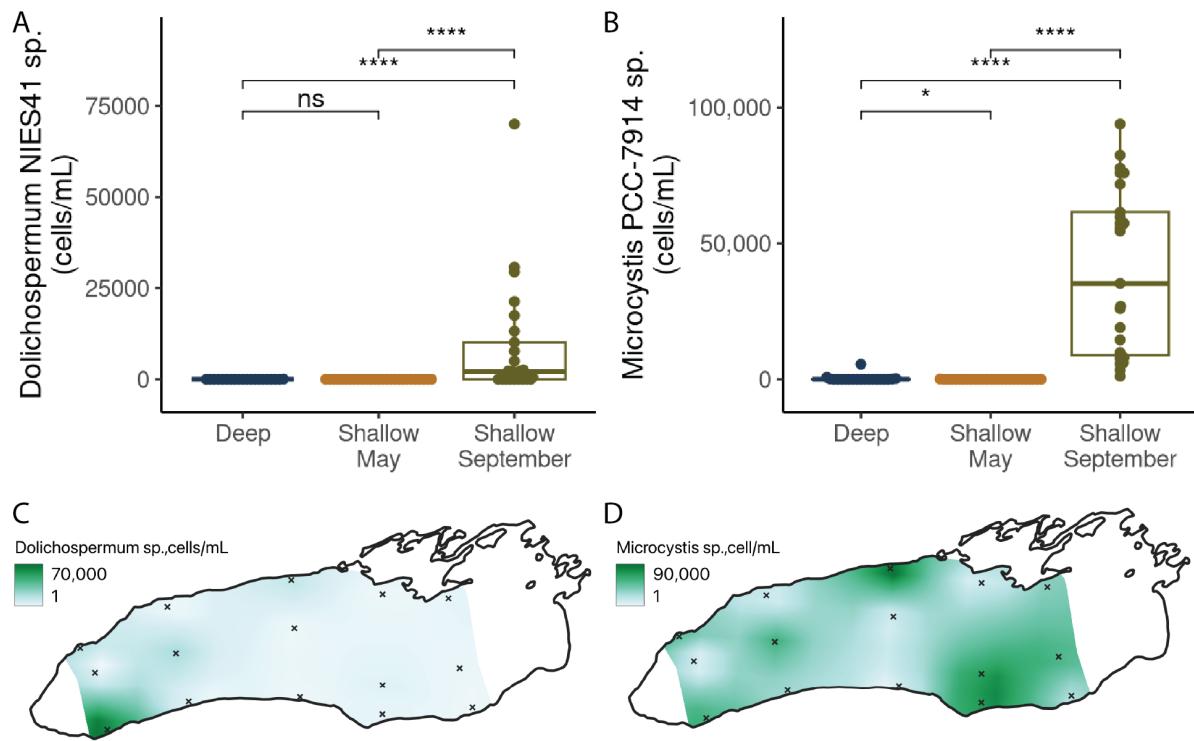
296 *Figure S10. Rare taxa differentiate the microbial communities of the Welland Canal and areas of*  
 297 *Upwelling in Lake Ontario.* (A-B) Distribution of ASV abundances in anomalous stations (Welland  
 298 Canal and upwelling areas) and nearby stations in (A) May and (B) September. (C) Relative abundance  
 299 of rare ASVs in anomalous and nearby stations. ASVs with greater than 0.01% abundance were removed,  
 300 and the relative abundance of remaining ASVs was calculated within each sample and summarized at  
 301 the Phylum level.



302 *Figure S11.* Cell counts are strongly correlated with temperature. (A) Cell abundances in millions  
 303 of cells/mL (y-axis) compared to temperature in Celsius (x-axis).  $R$  corresponds to the Spearman  
 304 correlation. (B) Cell counts split between depth-month groups. Comparisons represent Two-Sample  
 305 Wilcoxon Tests (\* =  $p < 0.05$ , \*\* =  $p < 0.01$ , \*\*\* =  $p < 0.001$ , \*\*\*\* =  $p < 0.0001$ ).  
 306

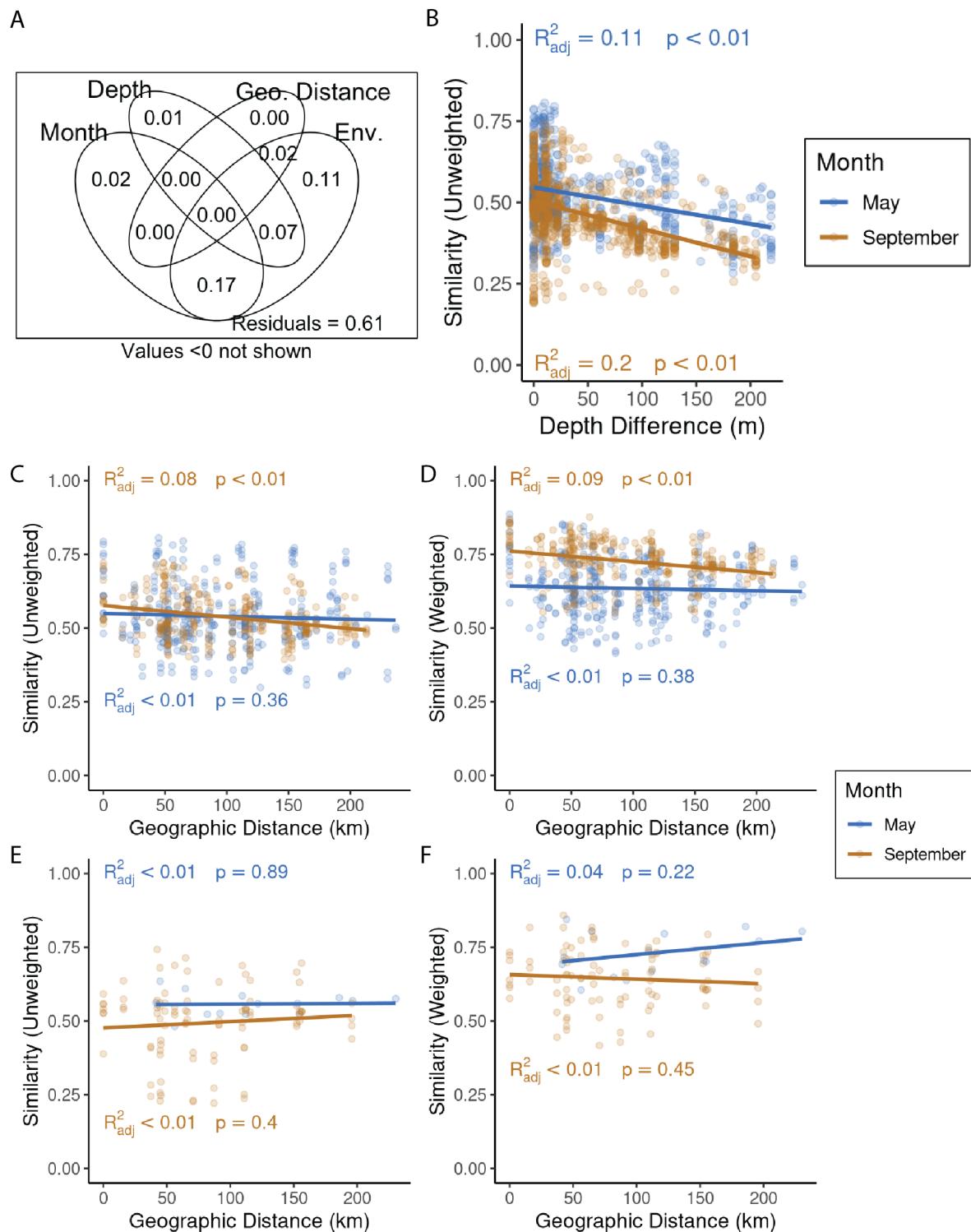


308 *Figure S12. Distribution and Differential Abundance of Microbial Taxa in Lake Ontario* (A) Maximum  
 309 observed cell count and prevalence (percentage of samples for which that ASV was observed) for individual ASVs are plotted and colored by Class. (B) ASVs whose maximum abundance exceeds 50,000  
 310 cells/mL. Labeling reflects Genus and Species as assigned by TaxAss. (C) Differential abundance was  
 311 calculated using ANCOM-BC2 using pairwise comparisons between our three month/depth groups at  
 312 the Class level. For taxa which were differentially abundant ( $p < 0.05$  and a passing sensitivity test) in  
 313 at least one pairwise comparison were retained. A Class's location in the plot represents the centroid of a  
 314 triangle whose vertices are the average absolute abundance of that class in each month/depth group. Class  
 315 colors are consistent across all three plots.



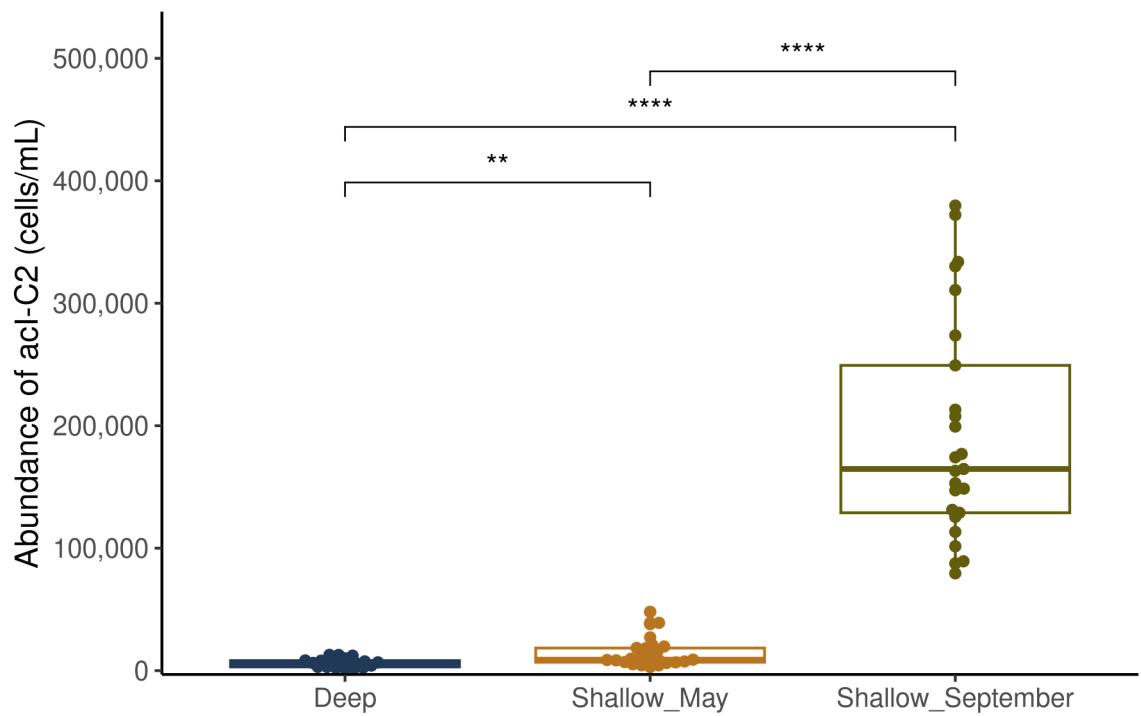
317

318 *Figure S13. Abundance of Potentially Harmful Cyanobacteria in Lake Ontario.* Absolute abundance of  
 319 two genera of potentially harmful bacteria across month-depth groups, including (A) *Dolichospermum*  
 320 *NIES41* of unknown species and (B) *Microcystis PCC-7914* of unknown species. Comparisons represent  
 321 Two-Sample Wilcoxon Tests (\* =  $p < 0.05$ , \*\* =  $p < 0.01$ , \*\*\* =  $p < 0.001$ , \*\*\*\* =  $p < 0.0001$ ). (C-D)  
 322 Spatial distribution of the (C) *Dolichospermum* sp. and (D) *Microcystis* sp. in September surface samples.  
 323 Note that scales are different between both maps. Interpolation was performed using multilevel B-splines  
 324 in QGIS/GRASS with 20km east-west steps, 10km north-south steps, and a Tykhonov regularization of  
 325 0.05.



326

327 *Figure S14. Fine-grain differences in decay relationships.* (A) Variance partitioning results using  
 328 the abundance-unweighted UniFrac dissimilarity as a response. Environmental variables (Env.) corre-  
 329 sponded to scaled physical and chemical parameters from each sample. (B) Depth decay using abundance-  
 330 unweighted UniFrac dissimilarity. Similarity is defined as 1 - unweighted UniFrac. Linear models were  
 331 calculated separately for samples from each month. (C-F) Distance decay relationships in shallow sam-  
 332 ples (C and D) versus deep samples (E and F), and abundance-unweighted UniFrac (C and E) and  
 333 abundance-weighted UniFrac (D and F).



334

335 *Figure S15. Abundance of Actinobacteria acI-C2.*\* Absolute abundance of Actinobacteria species *acI-C2*  
 336 across three depth-month groups. Significance represents Two-sample Wilcoxon Tests with Bonferroni  
 337 correction (\* =  $p < 0.05$ , \*\* =  $p < 0.01$ , \*\*\* =  $p < 0.001$ , \*\*\*\* =  $p < 0.0001$ ).

<sup>338</sup> **Supplemental Tables**

<sup>339</sup> **Table S1. List of Stations for Microbial Sampling**

Station	Date	Time	Latitude	Longitude	Depths	Notes					
					Sampled (m)						
<hr/>											
May											
Cruise											
008	20230519	1:20:00 PM	43.6232	-79.4518	5						
012	20230519	10:32:00 AM	43.5035	-79.3537	5, 15, 102.8						
017	20230519	7:10:00 AM	43.2250	-79.2714	5						
029	20230518	3:51:00 PM	43.8184	-78.8695	5, 19, 27						
033	20230518	11:35:00 AM	43.5969	-78.8133	5, 20, 135						
035	20230518	7:54:00 AM	43.3618	-78.7294	5, 10, 27						
038	20230518	2:24:00 AM	43.3834	-77.9896	5, 14						
041	20230517	8:02:00 PM	43.7166	-78.0273	5, 24, 127						
043	20230517	5:07:00 PM	43.9501	-78.0497	5						
048	20230516	5:58:00 PM	43.8797	-77.4397	5, 35						

Station	Date	Time	Latitude	Longitude	Depths	
					(m)	Notes
055	20230517	10:22:00	43.4442	-77.4389	5, 13, 189	
		AM				
064	20230516	2:15:00	43.5197	-76.91836	5, 14, 224	
		AM				
066	20230516	7:25:00	43.3336	-76.8343	5	
		AM				
074	20230515	4:37:00	43.7473	-76.5115	5, 14, 66	
		PM				
717	20230517	8:00:00	43.3005	-77.4408	5	
		AM				
September						
Cruise						
008	20230927	7:51:00	43.6222	-79.4509	5, 13.4	
		PM				
012	20230927	11:21:00	43.5034	-79.3531	5, 52, 102	
		PM				
017	20230927	2:54:00	43.2253	-79.2711	5, 10.98	
		PM				
029	20230927	4:12:00	43.8194	-78.8711	5, 15, 27.9	
		AM				
033	20230927	9:16:00	43.5965	-78.8143	5, 68.5,	Day
		AM			134.9	Sample
033	20230926	11:42:00	43.5973	-78.8149	5, 68, 134	Night
		PM				Sample
035	20230926	1:49:00	43.3614	-78.7294	5, 14.4,	
		PM			26.7	

Station	Date	Time	Latitude	Longitude	Depths	
					(m)	Sampled
						Notes
038	20230926	10:25:00	43.3831	-77.9887	5, 16.6	
		AM				
041	20230926	6:39:00	43.7174	-78.0263	5, 63.9,	
		AM			126	
043	20230926	3:38:00	43.9489	-78.0468	5, 15.6	
		AM				
048	20230926	12:31:00	43.8797	-77.4392	5, 16.5,	
		AM			31.5	
055	20230925	7:10:00	43.4428	-77.4398	5, 96,	
		PM			189.6	
062	20230923	3:11:00	43.5251	-76.9265	5, 18.3	
		PM				
064	20230923	11:50:00	43.3332	-76.8413	5, 106,	
		AM			210	
066	20230923	9:19:00	43.8600	-77.0012	5, 15.8	
		PM				
717	20230924	9:55:00	43.3004	-77.4407	5, 18.5,	
		AM			34.8	

<sup>340</sup> **Table S2. Software Versions and Citations**

Software	Version	Citation
<b>Languages and Environments</b>		
R	4.3.2	[33]
RStudio	2022.07.0	[34]
QGIS	3.40.5	[35]
<b>CLI Tools</b>		
MAFFT	7.520	[36]
TaxAss	2.1.1	[5]
FAPROTAX	1.2.10	[37]
<b>R Packages</b>		
ape	5.7-1	[38]
ANCOMBC	1.6.4	[39]
biomformat	1.30.0	[40]
Biostrings	2.64.1	[41]
broom	1.0.5	[42]
combinat	0.0-8	[43]
dada2	1.24.0	[44]
flowCore	2.14.2	[2]
ggcyto	1.30.2	[3]
ggdendro	0.2.0	[45]
ggfortify	0.4.16	[46]
ggh4x	0.2.8	[47]
ggpubr	0.6.0	[48]
ggrepel	0.9.4	[49]
ggsicle	0.3.1	[50]
ggtree	3.4.4	[51]
grid	4.3.2	[33]

Software	Version	Citation
gridExtra	2.3	[52]
gstat	2.1-1	[53]
iCAMP	1.5.12	[54]
iNEXT	3.0.0	[55]
lubridate	1.9.3	[56]
microViz	0.12.0	[57]
MBA	0.1-0	[58]
oce	1.8-2	[59]
pacman	0.5.1	[60]
patchwork	1.2.0	[61]
phyloseq	1.40.0	[62]
phytools	2.1-1	[63]
purrr	1.0.2	[64]
readxl	1.4.3	[65]
rstatix	0.7.2	[66]
scales	1.3.0	[67]
sf	1.0-14	[68]
speedyseq	0.5.3.9018	[69]
terra	1.7-71	[70]
tidytree	0.4.6	[71]
tidyverse	2.0.0	[72]
tmap	3.99.9	[73]
tmaptools	3.1-1	[74]
treemapify	2.5.6	[75]
units	0.8-5	[76]
vegan	2.6-4	[77]

<sup>341</sup> **Supplemental References**

- <sup>342</sup> 1. Agency UEP. Sampling and analytical procedures for GLNPO's open lake water quality survey of the Great Lakes. 2003. Great Lakes National Program Office Illinois, USA, 2003.
- <sup>343</sup> 2. Ellis B et al. flowCore: flowCore: Basic structures for flow cytometry data. 2024.
- <sup>344</sup> 3. Van P et al. ggcryo: Next-generation open-source visualization software for cytometry. *Bioinformatics* 2018.
- <sup>345</sup> 4. Callahan BJ et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 2016;13:581–583. <https://doi.org/10.1038/nmeth.3869>
- <sup>346</sup> 5. Rohwer RR et al. TaxAss: Leveraging a Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere* 2018;3:10.1128/msphere.00327-18. <https://doi.org/10.1128/msphere.00327-18>
- <sup>347</sup> 6. Quast C et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 2013;41:D590–D596. <https://doi.org/10.1093/nar/gks1219>
- <sup>348</sup> 7. Stegen JC et al. Quantifying community assembly processes and identifying features that impose them. *ISME J* 2013;7:2069–2079. <https://doi.org/10.1038/ismej.2013.93>
- <sup>349</sup> 8. Ning D et al. A quantitative framework reveals ecological drivers of grassland microbial community assembly in response to warming. *Nat Commun* 2020;11:4717. <https://doi.org/10.1038/s41467-020-18560-z>
- <sup>350</sup> 9. Urbach E et al. Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnology and Oceanography* 2001;46:557–572. <https://doi.org/10.4319/lo.2001.46.3.00557>
- <sup>351</sup> 10. Jabbari A et al. Nearshore-offshore exchanges by enhanced turbulent mixing along the north shore of Lake Ontario. *Journal of Great Lakes Research* 2023;49:596–607. <https://doi.org/10.1016/j.jglr.2023.03.010>

- 352 11. Hlevca B et al. Fine-Scale Hydrodynamic Modeling of Lake Ontario: Has Climate  
Change Affected Circulation Patterns? *Environ Model Assess* 2024. <https://doi.org/10.1007/s10666-024-10003-z>
- 353 12. Mehrshad M et al. Hidden in plain sight—highly abundant and diverse planktonic  
freshwater Chloroflexi. *Microbiome* 2018;6:176. <https://doi.org/10.1186/s40168-018-0563-8>
- 354 13. Okazaki Y et al. The Broad Habitat Spectrum of the CL500-11 Lineage (Phylum  
Chloroflexi), a Dominant Bacterioplankton in Oxygenated Hypolimnia of Deep  
Freshwater Lakes. *Frontiers in Microbiology* 2018;9.
- 355 14. Okazaki Y et al. Ubiquity and quantitative significance of bacterioplankton lin-  
eages inhabiting the oxygenated hypolimnion of deep freshwater lakes. *The ISME  
Journal* 2017;11:2279–2293. <https://doi.org/10.1038/ismej.2017.89>
- 356 15. Podowski JC et al. Genome Streamlining, Proteorhodopsin, and Organic Nitrogen  
Metabolism in Freshwater Nitrifiers. *mBio* 2022;13:e02379–21. <https://doi.org/10.1128/mbio.02379-21>
- 357 16. Langenheder S, Lindström ES. Factors influencing aquatic and terrestrial bacte-  
rial community assembly. *Environmental Microbiology Reports* 2019;11:306–315.  
<https://doi.org/10.1111/1758-2229.12731>
- 358 17. Newton RJ, Shade A. Lifestyles of rarity: understanding heterotrophic strategies  
to inform the ecology of the microbial rare biosphere. *Aquatic Microbial Ecology*  
2016;78:51–63. <https://doi.org/10.3354/ame01801>
- 359 18. Jia X, Dini-Andreote F, Salles JF. Community Assembly Processes of the Micro-  
bial Rare Biosphere. *Trends in Microbiology* 2018;26:738–747. <https://doi.org/10.1016/j.tim.2018.02.011>
- 360 19. Jones SE et al. Spatial and temporal scales of aquatic bacterial beta diversity.  
*Front Microbiol* 2012;3. <https://doi.org/10.3389/fmicb.2012.00318>
- 361 20. Paver SF, Newton RJ, Coleman ML. Microbial communities of the Laurentian  
Great Lakes reflect connectivity and local biogeochemistry. *Environmental Micro-  
biology* 2020;22:433–446. <https://doi.org/10.1111/1462-2920.14862>

- 362 21. Winters AD et al. Molecular characterization of bacterial communities associated  
with sediments in the Laurentian Great Lakes. *Journal of Great Lakes Research*  
2014;40:640–645. <https://doi.org/10.1016/j.jglr.2014.04.008>
- 363 22. Mohiuddin MM et al. Temporal and spatial changes in bacterial diversity in  
mixed use watersheds of the Great Lakes region. *Journal of Great Lakes Research*  
2019;45:109–118. <https://doi.org/10.1016/j.jglr.2018.10.007>
- 364 23. Gale J et al. Diverse and variable community structure of picophytoplankton  
across the Laurentian Great Lakes. *Limnology and Oceanography* 2023;68:2327–  
2345. <https://doi.org/10.1002/lno.12422>
- 365 24. Palermo CN et al. Analysis of bacterioplankton genes in an impaired Great Lakes  
harbour reveals seasonal metabolic shifts and a previously undetected cyanobac-  
terium. *Can J Microbiol* 2023;69:281–295. [https://doi.org/10.1139/cjm-2022-0252](https://doi.org/10.1139/cjm-2022-<br/>0252)
- 366 25. Newton RJ et al. A Guide to the Natural History of Freshwater Lake Bacteria.  
*Microbiology and Molecular Biology Reviews* 2011;75:14–49. [https://doi.org/10.1128/mmbr.00028-10](https://doi.org/10.<br/>1128/mmbr.00028-10)
- 367 26. J P et al. Top-down effects on the size-biomass distribution of a freshwa-  
ter bacterioplankton community. *Aquatic Microbial Ecology* 1996;10:255–263.  
<https://doi.org/10.3354/ame010255>
- 368 27. Kim S et al. Cultivation of Dominant Freshwater Bacterioplankton Lineages Us-  
ing a High-Throughput Dilution-to-Extinction Culturing Approach Over a 1-Year  
Period. *Front Microbiol* 2021;12. <https://doi.org/10.3389/fmicb.2021.700637>
- 369 28. Lowe WH, McPeek MA. Is dispersal neutral? *Trends in Ecology & Evolution*  
2014;29:444–450. <https://doi.org/10.1016/j.tree.2014.05.009>
- 370 29. Evans S, Martiny JBH, Allison SD. Effects of dispersal and selection on stochastic  
assembly in microbial communities. *The ISME Journal* 2017;11:176–185. [https://doi.org/10.1038/ismej.2016.96](https://<br/>doi.org/10.1038/ismej.2016.96)

- 371 30. McNichol J et al. Evaluating and Improving Small Subunit rRNA PCR Primer  
Coverage for Bacteria, Archaea, and Eukaryotes Using Metagenomes from Global  
Ocean Surveys. *mSystems* 2021;6:10.1128/msystems.00565-21. <https://doi.org/10.1128/msystems.00565-21>
- 372 31. Muylaert K et al. Relationship between Bacterial Community Composition and  
Bottom-Up versus Top-Down Variables in Four Eutrophic Shallow Lakes. *Ap-  
plied and Environmental Microbiology* 2002;68:4740–4750. <https://doi.org/10.1128/AEM.68.10.4740-4750.2002>
- 373 32. Pradeep Ram AS, Keshri J, Sime-Ngando T. Differential impact of top-down and  
bottom-up forces in structuring freshwater bacterial communities. *FEMS Micro-  
biology Ecology* 2020;96:fiaa005. <https://doi.org/10.1093/femsec/fiaa005>
- 374 33. R Core Team. R: A language and environment for statistical computing. Vienna,  
Austria: R Foundation for Statistical Computing, 2022.
- 375 34. RStudio Team. RStudio: Integrated Development Environment for R. Boston,  
MA: RStudio, PBC., 2020.
- 376 35. QGIS Development Team. QGIS Geographic Information System. 2021.
- 377 36. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version  
7: Improvements in Performance and Usability. *Molecular Biology and Evolution*  
2013;30:772–780. <https://doi.org/10.1093/molbev/mst010>
- 378 37. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global  
ocean microbiome. *Science* 2016;353:1272–1277. <https://doi.org/10.1126/science.aaf4507>
- 379 38. Paradis E et al. ape: Analyses of phylogenetics and evolution. 2023.
- 380 39. Lin H. ANCOMBC: Microbiome differential abundance and correlation analyses  
with bias correction. 2022.
- 381 40. McMurdie PJ, Paulson JN. biomformat: An interface package for the BIOM file  
format. 2023.
- 382 41. Pagès H et al. Biostrings: Efficient manipulation of biological strings. 2022.

- 383 42. Robinson D, Hayes A, Couch S. *broom*: Convert statistical objects into tidy tibbles.  
2023.
- 384 43. Chasalow S. *combinat*: combinatorics utilities. 2012.
- 385 44. Callahan B, McMurdie P, Holmes S. *dada2*: Accurate, high-resolution sample  
inference from amplicon sequencing data. 2022.
- 386 45. de Vries A, Ripley BD. *ggdendro*: Create dendograms and tree diagrams using  
ggplot2. 2024.
- 387 46. Horikoshi M, Tang Y. *ggfortify*: Data visualization tools for statistical analysis  
results. 2023.
- 388 47. van den Brand T. *ggh4x*: Hacks for ‘ggplot2’. 2024.
- 389 48. Kassambara A. *ggpubr*: ggplot2 based publication ready plots. 2023.
- 390 49. Slowikowski K. *ggrepel*: Automatically position non-overlapping text labels with  
ggplot2. 2023.
- 391 50. Landis J. *ggside*: Side grammar graphics. 2024.
- 392 51. Yu G, Lam TT-Y, Xu S. *ggtree*: an R package for visualization of tree and anno-  
tation data. 2022.
- 393 52. Auguie B. *gridExtra*: Miscellaneous functions for ‘Grid’ graphics. 2017.
- 394 53. Pebesma E, Graeler B. *gstat*: Spatial and spatio-temporal geostatistical modelling,  
prediction and simulation. 2023.
- 395 54. Ning D. *iCAMP*: Infer community assembly mechanisms by phylogenetic-bin-  
based null model analysis. 2022.
- 396 55. Hsieh TC, Ma KH, Chao A. *iNEXT*: Interpolation and extrapolation for species  
diversity. 2022.
- 397 56. Spinu V, Grolemond G, Wickham H. *lubridate*: Make dealing with dates a little  
easier. 2023.
- 398 57. Barnett D. *microViz*: Microbiome data analysis and visualization. 2024.
- 399 58. Finley A, Banerjee S, Hjelle Ø. *MBA*: Multilevel b-spline approximation. 2022.
- 400 59. Kelley D, Richards C. *oce*: Analysis of oceanographic data. 2023.

- 401 60. Rinker TW, Kurkiewicz D. pacman: Package management for R. Buffalo, New  
York, 2018.
- 402 61. Pedersen TL. patchwork: The composer of plots. 2024.
- 403 62. McMurdie PJ et al. phyloseq: Handling and analysis of high-throughput micro-  
biome census data. 2022.
- 404 63. Revell LJ. phytools: Phylogenetic tools for comparative biology (and other things).  
2024.
- 405 64. Wickham H, Henry L. purrr: Functional programming tools. 2023.
- 406 65. Wickham H, Bryan J. readxl: Read excel files. 2023.
- 407 66. Kassambara A. rstatix: Pipe-friendly framework for basic statistical tests. 2023.
- 408 67. Wickham H, Pedersen TL, Seidel D. scales: Scale functions for visualization. 2023.
- 409 68. Pebesma E. sf: Simple features for R. 2023.
- 410 69. McLaren M. speedyseq: Faster implementations of phyloseq functions. 2024.
- 411 70. Hijmans RJ. terra: Spatial data analysis. 2024.
- 412 71. Yu G. tidytree: A tidy tool for phylogenetic tree data manipulation. 2023.
- 413 72. Wickham H. tidyverse: Easily install and load the tidyverse. 2023.
- 414 73. Tennekes M. tmap: Thematic maps. 2024.
- 415 74. Tennekes M. tmaptols: Thematic map tools. 2021.
- 416 75. Wilkins D. treemapify: Draw treemaps in ‘ggplot2’. 2023.
- 417 76. Pebesma E, Mailund T, Hiebert J. Measurement units in R. *R Journal* 2016;8:486–  
494. <https://doi.org/10.32614/RJ-2016-061>
- 418 77. Oksanen J et al. vegan: Community ecology package. 2022.