Diversity, distribution, and expression of opsin genes in freshwater lakes

Shaomei He[1,2], Alexandra M. Linz[1], Sarah L.R. Stevens[1], Patricia Q. Tran[1,3], Francisco Moya-Flores[4], Ben O. Oyserman[4], Jeffrey R. Dwulit-Smith[1,5], Katrina T. Forest[1,5], Katherine D. McMahon[1,4]

[1]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA

[2]Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA

[3]Department of Integrative Biology, University of Wisconsin-Madison, Madison, WI, USA

[4]Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, WI, USA

[5]Program in Biophysics, University of Wisconsin-Madison, Madison, WI, USA

Corresponding Author

Katherine D. McMahon (trina.mcmahon@wisc.edu)

Running Title: Opsin genes in freshwater

Key Words

Type I rhodopsin; Opsin genes; Clearwater lake; Humic lake; Dissolved organic carbon (DOC); Diversity; Diel gene expression

ABSTRACT

Microbial rhodopsins are widely distributed in aquatic environments and may significantly contribute to phototrophy and energy budgets in global oceans. However, the study of freshwater rhodopsins has been largely limited. Here, we explored the diversity, ecological distribution, and expression of opsin genes that encode the apoproteins of Type I rhodopsins in humic and clearwater lakes with contrasting physicochemical and optical characteristics. Using metagenomes and metagenome-assembled genomes, we recovered opsin genes from a wide range of taxa, mostly predicted to encode green light-absorbing proton pumps. Viral opsin and novel bacterial opsin clades were recovered. Opsin genes occurred more frequently in taxa from clearwater than from humic water, and opsins in some taxa have non-typical ion-pumping motifs that might be associated with physicochemical conditions of these two freshwater types. Analyses of the surface layer of 33 freshwater systems revealed an inverse correlation between opsin gene abundance and lake dissolved organic carbon (DOC). In humic water with high terrestrial DOC and light-absorbing humic substances, opsin gene abundance was low and dramatically declined within the first few meters, whereas the abundance remained relatively high along the bulk water column in clearwater lakes with low DOC, suggesting opsin gene distribution is influenced by lake optical properties and DOC. Gene expression analysis confirmed the significance of rhodopsin-based phototrophy in clearwater lakes and revealed different diel expressional patterns among major phyla. Overall, our analyses revealed freshwater opsin diversity, distribution and expression patterns, and suggested the significance of rhodopsin-based phototrophy in freshwater energy budgets, especially in clearwater lakes.

INTRODUCTION

Microbial rhodopsin is a retinal-activated photosystem. In marine environments, microbial rhodopsins are found in many diverse prokaryotes and exhibit a rich genetic and functional diversity (McCarren & DeLong 2007; Olson *et al.* 2018; Rozenberg *et al.* 2021; Rusch *et al.* 2007). As microbial rhodopsins often occur in more than half of the cells in a wide range of marine samples (Finkel *et al.* 2013; Maresca *et al.* 2018; Olson *et al.* 2018; Sieradzki *et al.* 2018), rhodopsin-based phototrophy has been suggested to be a prevalent form of phototrophy and may significantly influence energy budgets in the global oceans (Finkel *et al.* 2013; Gomez-Consarnau *et al.* 2019), although whether the harvested sunlight is used for growth, survival or other physiological and ecological benefits may differ among bacterial species (Pinhassi *et al.* 2016). Contrasting to the rich knowledge regarding marine microbial rhodopsins, the diversity, ecological distribution, and significance of rhodopsins in freshwater have been much less studied. Particularly, rhodopsins in humic aquatic environments (with high terrestrially-derived allochthonous organic carbon, dimly lit water columns, and often acidic pH) are largely unexplored.

Being a simple photosystem, a rhodopsin consists of an opsin, which is an integral membrane protein with seven transmembrane helices, and a covalently bound retinal as a light-absorbing chromophore. Microbial (also termed Type I, as compared to Type II in animals) rhodopsins play important roles in microbial physiology, with known functions including light-driven ion transport (e.g. proton, H^+, chloride, Cl^-, and sodium, Na^+ pumps), cation channeling, and sensing that directs phototaxis (Rozenberg *et al.* 2021). More recently, a novel family of microbial rhodopsins distinct from Type I rhodopsins was discovered through functional metagenomics and named

68 heliorhodopsin (HeR) (Pushkarev *et al.* 2018). As HeRs are only distantly related to Type I
69 rhodopsins and their functions are largely unknown, in our study, we solely focused on Type I
70 rhodopsins (hereinafter referred to as "rhodopsins" and genes encoding their apoproteins, opsins,
71 were referred to as "opsin genes" for simplicity), although we note that HeRs have been found in
72 diverse monoderm bacteria in freshwater environments (Chazan *et al.* 2022; Flores-Uribe *et al.*
73 2019).
74
75 The first microbial opsin genes in freshwater were discovered through metagenomic studies, where
76 the predominant opsin genes were recovered from Actinobacteria and therefore named
77 "actinorhodopsin" (ActR) (Sharma *et al.* 2009; Sharma *et al.* 2008). ActR is phylogenetically
78 closer to xanthorhodopsins than to other rhodopsin groups (Vollmers *et al.* 2013) and functions as
79 an outward H^+ pump (Dwulit-Smith *et al.* 2018; Keffer *et al.* 2015). In addition to ActR, opsin
80 genes were also recovered from some other phyla in freshwater lakes (Atamna-Ismaeel *et al.* 2008;
81 Cabello-Yeves *et al.* 2017; Galachyants *et al.* 2021; Jaffe *et al.* 2022; Martinez-Garcia *et al.* 2012;
82 Podowski *et al.* 2021; Sharma *et al.* 2009).
83
84 Rhodopsin has been shown to enhance bacterial growth and survival under carbon-limited
85 conditions when exposed to light (Gómez-Consarnau *et al.* 2010; Gómez-Consarnau *et al.* 2007;
86 Johnson *et al.* 2010; Steindler *et al.* 2011; Yoshizawa *et al.* 2014), suggesting the importance of
87 light and carbon availability in the regulation of this photoreceptive protein. In freshwater, lake
88 optical properties and organic carbon quality/quantity are often correlated, and the former can be
89 greatly impacted by the latter (Maizel *et al.* 2017; Read & Rose 2013). Therefore, an interesting

yet unexplored hypothesis is that the diversity and distribution of opsin genes in freshwater varies with different lake optical properties and organic carbon quality/quantity.

In this study, we surveyed two contrasting freshwaters, a eutrophic clearwater lake and a humic brown-water bog lake with distinct physicochemical characteristics. In particular, the humic bog has much higher concentrations of dissolved organic carbon (DOC), most of which is allochthonous (entering the aquatic system from its catchment). Such terrestrially-derived DOC comprises light-absorbing humic substances which stain the water brown, and thus the bog exhibits very different optical properties from the eutrophic lake, which is "clear" as opposed to "stained" (Maizel *et al.* 2017). To investigate the opsin gene diversity in these environments, we used metagenomic sequences and metagenome-assembled genomes (MAGs) obtained from the co-assembly of a large number of metagenomes collected from a time-series study of these two lakes (Linz *et al.* 2018), as well as the 12665 MAGs from the stratfreshDB which was generated from 41 lakes and ponds (Buck *et al.* 2021). To further study the relationship between opsin gene distribution and freshwater DOC and optical properties, we included metagenomes from a total of 33 freshwater lakes around the globe that span a wide range of DOC levels and clarity. We estimated the relative opsin gene abundance in the surface layer of these lakes, as well as depth-resolved opsin abundance profiles from a subset of these lakes. Lastly, we explored opsin gene expression and diel regulation in the epilimnion of two clearwater lakes and the humic bog. Overall, our analysis recovered a greater than previously appreciated diversity of freshwater opsin genes, revealed the importance of DOC and optical properties in opsin distribution, and suggested the significance of rhodopsin-based phototrophy in clearwater lakes.

113 **MATERIALS AND METHODS**

114 **Lake Mendota and Trout Bog metagenomes.** Two contrasting temperate lakes in Wisconsin,
115 USA were studied in detail. Lake Mendota (43.10ºN, 89.41ºW) is a eutrophic clearwater lake with
116 an average DOC of 5.0 mg/L. Trout Bog (46.01ºN, 89.61ºW) is a small humic lake, with an
117 average DOC of 20.0 mg/L (data from NTL-LTER, averaged over the course of the study). More
118 detailed lake physicochemical and limnological characteristics are listed in **Table S1**. For Lake
119 Mendota, 0 to 12 m depth-integrated water samples (mostly consisting of the epilimnion layer)
120 were collected at 94 time points from 2008 to 2012, and samples are referred to as "ME". For
121 Trout Bog, the integrated epilimnion layer and the integrated hypolimnion layer were collected at
122 45 time points respectively from 2007 to 2009, and samples were referred to as "TE" and "TH",
123 respectively.

124

125 Details of metagenome sequencing, assembly, and binning were previously described (Bendall *et*
126 *al.* 2016; Linz *et al.* 2018). Briefly, shotgun metagenome libraries were constructed from each of
127 the 184 DNA samples. Assembly was performed on individual metagenomes. In addition, three
128 "combined" assemblies were generated by co-assembling reads from all metagenomes within the
129 ME, TE, and TH groups, respectively. Binning was conducted on the three combined assemblies
130 to obtain metagenome-assembled genomes (MAGs), and this resulted in 486 genome bins. By
131 further fine-tuning to remove outlier contigs, and excluding bins with high contamination (>10%)
132 or low completeness (<50%), a total of 202 high-quality MAGs (referred to henceforth as MAGs
133 to distinguish them from the remaining bins which have lower qualities) were generated (Linz *et*
134 *al.* 2018). Taxonomic assignment of MAGs and bins was conducted using PhyloSift based on 37
135 conserved phylogenetic marker genes (Darling *et al.* 2014). The 184 individual metagenome

136 assemblies, the three combined assemblies, and recovered MAGs/bins were submitted to the DOE
137 Joint Genome Institute's Integrated Microbial Genome (IMG) database for gene prediction and
138 function annotation (Markowitz *et al.* 2013).
139
140 **Opsin phylogenetic tree reconstruction and inference of taxonomy.** Opsin sequences were
141 identified by hmmsearch with model-specific trusted cut-offs (--cut_tc) for the hidden Markov
142 model (HMM) of pfam01036 (http://pfam.xfam.org) implemented by the IMG annotation pipeline.
143 Representative opsin reference sequences were chosen by clustering bacterial and archaeal opsin
144 protein sequences from isolate genomes at IMG based on 90% amino acid sequence identity using
145 the UCLUST algorithm provided by the USEARCH tool (v5.2.32) (Edgar 2010). Representative
146 sequences from each cluster, opsins from previously sequenced freshwater single-amplified
147 genomes (SAGs) and MAGs, and viral opsin sequences from the study by Yutin and Koonin
148 (Yutin & Koonin 2012) were included in the phylogenetic analysis, together with the 317 opsin
149 sequences longer than 200 amino acids (which is ~80% of opsin full length) in our three combined
150 assemblies (including sequences in MAGs/bins and the remaining un-binned). Amino acid
151 sequences were aligned using MUSCLE (v3.8.31) (Edgar 2004), and the alignment was trimmed
152 to exclude columns that contained gaps for more than 30% of the included sequences. Poorly
153 aligned positions and divergent regions were further eliminated using Gblocks (v0.91b)
154 (Castresana 2000). A maximum likelihood phylogenetic tree was constructed using PhyML 3.0
155 (Guindon *et al.* 2010), with the LG substitution model (Le & Gascuel 2008) and the gamma
156 distribution parameter estimated by PhyML. Bootstrap values were calculated based on 100
157 replicates. After constructing this phylogenetic tree, sequences shorter than 200 amino acids in the
158 three combined assemblies were added to the existing tree using PPlacer (v1.1.alpha17) (Matsen

159 *et al.* 2010). The resulting phylogenetic tree was used for taxonomic classification of un-binned
160 metagenome opsin sequences at the phylum level.
161
162 **Searching opsin genes in stratfreshDB MAGs.** We first performed gene prediction using
163 Prodigal (v2.6.3) on all stratfreshDB MAGs deposited at European Nucleotide Archive (Buck *et*
164 *al.* 2021). We then performed hmmsearch (HMMER v3.3) with the model-specific trusted cutoff
165 (--cut_tc) option for pfam01036 to identify opsin genes. Taxonomic information of these MAGs
166 was directly from the classification with GTDB-Tk in the original study (Buck *et al.* 2021).
167
168 **Estimation of opsin gene relative abundance in ME, TE, and TH metagenomes.** The
169 abundance of an individual gene in an assembled metagenome was estimated using its average
170 per-base coverage, which indicates how many times a base was sequenced and was determined by
171 mapping metagenome reads to assembled contigs. This estimate is available in the IMG database
172 as "read depth", and we referred to it as "coverage-weighted abundance" in this study to reflect
173 the abundance of the population that contributed to this gene in the metagenome. To compare opsin
174 gene abundance among different metagenomes, the opsin gene relative abundance was estimated
175 by the sum of the coverage-weighted abundance of all opsin genes in the metagenome, and further
176 normalized (i.e. divided) by the average of the coverage-weighted abundance of three single-copy
177 conserved housekeeping genes (represented by pfam00121, pfam00573, pfam00687,
178 http://pfam.xfam.org) in the same metagenome. The three single-copy genes were chosen because
179 they have comparable HMM lengths to opsin genes to minimize the bias in gene abundance
180 estimates associated with gene lengths (He *et al.* 2015).
181

182 **Estimation of opsin gene abundance for additional lakes.** To further study opsin gene
183 distribution, we included metagenomes from 31 additional freshwater lakes for which organic
184 carbon concentrations were available. Among these additional lakes, 18 are from the stratfreshDB
185 deposited at European Nucleotide Archive (Buck *et al.* 2021) and 13 are from publicly available
186 metagenomes at IMG, including two unpublished from our lab and 11 from published studies
187 (Dam *et al.* 2020; Denef *et al.* 2016; Linz *et al.* 2020; Podowski *et al.* 2021; Tran *et al.* 2018; Tran
188 *et al.* 2021) (**Tables S2 and S3**). Together with Mendota and Trout Bog, the 33 lakes span a wide
189 range of DOC levels and exhibit different clarity (examples of clarity data are listed in **Table S1**).
190 Some of these lakes have samples collected from different depths, allowing the study of opsin gene
191 vertical distribution (**Table S4**). For metagenomes available at IMG, the relative abundance of
192 opsin genes was estimated as described above. For metagenomes from the stratfreshDB, we
193 performed gene prediction using Prodigal (v2.6.3) on the assemblies, and then hmmsearch
194 (HMMER v3.3) with the model-specific trusted cutoff (--cut_tc) option for the HMM models of
195 pfam01036, pfam00121, pfam00573, and pfam00687 to identify opsin and the three single-copy
196 conserved genes respectively. To obtain gene coverage information, we first processed the raw
197 reads to remove adapter sequences and low-quality bases using fastp (v0.23.1) with default settings.
198 We then mapped the processed reads to corresponding metagenome assemblies using bowtie2
199 (v2.2.2) with default settings for gene coverage. Coverage-weighted abundance from these genes
200 was used to estimate opsin gene abundance as described earlier. The in-house bioinformatics
201 workflow that we used on the stratfreshDB metagenomes, when used on a subset of the included
202 IMG metagenomes, yielded results very comparable to results obtained from the IMG pipeline.
203 Therefore, no systemic bias was associated with different bioinformatic workflows.
204

205 **Analysis of opsin gene expression and diel trends.** We used metatranscriptomes from the
206 GEODES (Gene Expression in Oligotrophic, Dystrophic, and Eutrophic Systems) dataset (Linz *et*
207 *al.* 2020) to study the expression of opsin genes. This dataset includes triplicate samples collected
208 every four hours for two days from Lake Mendota and Sparking Lake, an oligotrophic clearwater
209 lake in Northern Wisconsin (**Table S1**). Triplicate samples were similarly collected from Trout
210 Bog but samples from the second day did not yield sufficient RNA for sequencing. Notably, an
211 internal standard, an *in vitro* transcribed pFN18A of a known concentration was added to each
212 sample after the bead beating step during the RNA extraction (Linz *et al.* 2020). This type of
213 internal standard enables absolute quantification of transcript abundance as copies of transcripts
214 per liter of aquatic sample and thus allows comparing transcript abundance across different
215 samples (Gifford *et al.* 2011). As two metagenomes were collected along with the
216 metatranscriptomes for each lake, we first clustered opsin sequences from the two replicate
217 metagenomes with 97% nucleotide identity using the UCLUST algorithm provided by the
218 USEARCH tool (v3.8.31) (Edgar 2010) to remove redundancy, followed by removing sequences
219 shorter than 600 bp. The resulting opsin sequences from Sparkling Lake were used as references
220 for mapping Sparkling metatranscriptome reads. For Mendota and Trout Bog, the resulting
221 GEODES opsin sequences from the above steps were compared to opsin sequences longer than
222 600 bp from the Mendota and Trout Bog combined assembly respectively by BLASTN to identify
223 GEODES sequences not represented in the corresponding combined assembly as defined by <97%
224 nucleotide identity. The identified GEODES sequences were added to the >600 bp opsin sequences
225 from the corresponding combined assembly to generate the mapping reference for Mendota and
226 Trout Bog respectively. Metatranscriptome reads used for mapping were previously processed to
227 remove adapter sequences, low-quality or short reads, and rRNA sequences by Linz *et al*. (2020),

228 and were mapped to corresponding opsin references with the internal standard sequence by using
229 bowtie2 (v2.2.2) with default settings. Metatranscriptomes with low read counts from the internal
230 standard were discarded according to Linz *et al.* (2020). Opsin transcript read counts were
231 normalized using the read count from the internal standard and converted to transcript
232 concentrations. RAIN test (Thaben & Westermark 2014) was conducted to detect genes with diel
233 cyclic trends with *p*-values of less than 0.05 for cyclic genes.
234

235 **RESULTS**

236 **1. Diversity of freshwater opsin genes**

237 ***Overview***

238 Here, we studied opsin gene diversity using three combined metagenome assemblies, generated
239 by co-assembling reads from all metagenomes within the ME (Mendota Epilimnion, 94
240 metagenomes), TE (Trout Bog Epilimnion, 45 metagenomes), and TH (Trout Bog Hypolimnion,
241 45 metagenomes) groups (Linz *et al.* 2018). Altogether, from the three combined assemblies, opsin
242 genes were recovered in 42 of 184 MAGs (**Table 1**) plus 50 bins that did not pass our criteria for
243 quality MAGs. Genome completeness/contamination assessments and taxonomic classification of
244 these MAGs and bins, and their opsin gene IMG IDs are provided in **Table S5**. At the phylum
245 level, the classification of these opsin genes through classifying their resident MAG/bins was
246 largely consistent with the IMG classification of opsin gene-containing contigs based on the
247 consensus BLASTP hits for all genes on the contig (**Table S5**). Therefore, we assumed that opsin
248 genes from these MAGs/bins were suitable phylogenetic anchors to classify metagenome opsin
249 sequences in un-binned contigs to at least the phylum level.
250

251 We constructed a phylogenetic tree with previously published references and opsin sequences
252 longer than 200 amino acids in the three combined assemblies, which include sequences from
253 MAGs/bins and un-binned contigs (**Figure 1**, with a rectangular view of this tree with more
254 detailed gene information in **Figure S1**). Metagenome opsin sequences shorter than 200 amino
255 acids were aligned and added to the phylogenetic tree (**Figure S2**). Based on the coverage-
256 weighted opsin gene abundance in the assemblies, we estimated the percent distribution of opsin
257 genes among different taxa and rhodopsin types in each lake/layer (**Figure S3**).
258
259 The majority of recovered opsin sequences belong to known H^+-pumping branches, such as
260 actinorhodopsin (ActR), proteorhodopsin (PR), xanthorhodopsin (XR) and *Exiguobacterium*
261 *sibiricum* rhodopsin (ESR) (**Figure 1**). Most sequences have the conserved DTE (aspartic acid,
262 threonine, glutamic acid) motif corresponding to Positions 85, 89, and 96 (BR numbering) for H^+
263 accepting and donation (Beja & Lanyi 2014). The spectral tuning residue on the recovered opsin
264 sequences is a hydrophobic residue methionine (M) or leucine (L) for absorbing green light, with
265 a few exceptions that have the hydrophilic glutamine (Q) for absorbing blue light (Gómez-
266 Consarnau *et al.* 2007; Man *et al.* 2003). Therefore the majority of recovered opsin genes were
267 predicted to encode green light-absorbing H^+ pumps. In Mendota, opsin genes were recovered
268 from all major freshwater phyla except Cyanobacteria. These included Actinobacteria,
269 Bacteroidetes, Proteobacteria, Planctomycetes, Verrucomicrobia, and Chloroflexi, and the first
270 three phyla were the major contributors to opsin gene abundance (**Figure S3**). Whereas in Trout
271 Bog, opsin genes were restricted to fewer taxa, mainly from Actinobacteria (acI-B2 in particular)
272 and some from Proteobacteria and *Candidatus* Saccharibacteria (TM7) (**Figure S3**).
273

274 As we only recovered a small number of opsin sequences in Trout Bog MAGs (**Table S6**), we also
275 searched the 12665 high-quality MAGs from the stratfreshDB which was generated from 41
276 freshwater lakes and ponds, including many "clearwater" (not stained brown) and brown-water
277 humic environments. We found a total of 2011 opsin genes, most of which were also predicted to
278 encode green light-absorbing H^+-pumps based on the DTE motif (**Figure 2a**). Similarly,
279 Actinobacteria, Bacteroidetes and Proteobacteria were the major opsin-containing phyla (**Table**
280 **S7**). In addition to the phyla found in Mendota and Trout Bog, a very small number of opsin genes
281 were also recovered from phyla such as Acidobacteriota and Bdellovibrionota from clearwater
282 environments. Like the contrast between Mendota and Trout Bog (**Figure 3a**), opsin genes
283 occurred more frequently in clearwater MAGs than in humic water MAGs from the stratfreshDB
284 (**Table S7**).
285
286 ***Opsin gene prevalence and diversity in freshwater Actinobacteria***
287 Actinobacteria was the phylum possessing the largest number of opsin genes, accounting for 46-
288 61% of all opsin gene coverage in each of the three combined assemblies (**Figure S3**). In ME,
289 opsin genes were widely distributed in nearly all lineages/clades of Actinobacteria that were
290 previously known to harbor opsin genes (**Figure S4**). The acI lineage, particularly Ca. Planktophila
291 (acI-A) and Nanopelagicus (acI-B), was the most prominent contributor to opsin genes in ME
292 (**Figure S3**). Although acI was also the most abundant contributor to opsin genes in TE and TH,
293 their ActR genes were confined within a branch of acI-B, likely from acI-B2. This branch
294 exclusively contained opsin sequences from humic lakes, consistent with the finding that acI-B2
295 prefers humic lakes as compared to acI-B1 (Garcia *et al.* 2014; Newton *et al.* 2006).
296

297 Most Actinobacterial opsin genes belong to ActR, an outward H^+-pumping group phylogenetically
298 close to XR. Subgroup I of XR also has a secondary chromophore, a noncovalently bound
299 carotenoid functioning as a light-harvesting antenna, in addition to the retinal (Balashov *et al.* 2005;
300 Béjà *et al.* 2000; Imasheva *et al.* 2009; Vollmers *et al.* 2013). Notably, nearly all ActR sequences
301 recovered in our study contain the signature glycine that replaces a bulky residue on helix 5, which
302 permits binding of the carotenoid antenna salinxanthin in the well-characterized XR from
303 *Salinibacter ruber* (Balashov *et al.* 2005). Whether any individual ActR-encoding microbe
304 synthesizes a carotenoid antenna and whether particular ActR exemplars bind a secondary
305 carotenoid may need to be addressed on a case-by-case basis but the presence of the signature
306 glycine residue and the genetic machinery for a secondary carotenoid biosynthesis have been noted
307 (Dwulit-Smith *et al.* 2018), and in at least one case the binding of a native actinobacterial
308 carotenoid to an ActR has been demonstrated (Chuon *et al*., 2021).
309
310 In addition to ActR, a rhodopsin group phylogenetically distinct from the typical ActR was
311 previously discovered in both freshwater and marine actinobacterial MAGs belonging to the order
312 of Acidimicrobiales and was named acidirhodopsin (short as "AcidiR" in our study in contrast to
313 ActR) (Ghai *et al.* 2014; Mizuno *et al.* 2015). According to the freshwater taxonomy, the AcidiR-
314 containing freshwater MAG (Ghai *et al.* 2014) belongs to the Iluma-B clade (alternative name
315 acIV-B). Here in ME, we also found sequences belonging to this Iluma-B AcidiR clade (**Figures**
316 **1** and **S4)**, accounting for ~5% of all opsin gene coverage in the combined ME assembly (**Figure**
317 **S3**). This AcidiR clade did not form a monophyletic group with ActR, but instead was a sister
318 clade to the branch formed by ActR and XR (**Figures 1** and **S1**), similar to its placement in
319 previously reconstructed opsin phylogenetic trees (Cabello-Yeves *et al.* 2017; Durán-Viseras *et al.*

320 2019; Mizuno *et al.* 2015). In addition, we identified a different branch of opsin sequences
321 belonging to Iluma (likely Iluma-A), accounting for ~6% of all opsin gene coverage in the
322 combined ME assembly. This branch did not form a monophyletic clade with the previously
323 identified AcidiR branch (**Figures 1 and S1**). Therefore, we use "AcidiR (Group I)" and "AcidiR
324 (Group II)" to distinguish them. A difference between these two groups is that the conserved
325 proline in the middle of helix 4 of ActR is also conserved in AcidiR Group I but missing in Group
326 II (as well as XR). Such a proline residue may impact opsin structure and thus render different
327 photoproperties as previously postulated (Dwulit-Smith *et al.* 2018). Notably, the bootstrap values
328 associated with the relative placement among the ActR, XR and AcidiR branches were lower than
329 50%, and this was also the case in other studies (Durán-Viseras *et al.* 2019; Mizuno *et al.* 2015).
330 As such, the phylogeny and evolutional history among these clades needs to be further resolved.
331
332 Besides ActR and AcidiR, an ME MAG classified to the Myco tribe within the acTH2 lineage of
333 Actinobacteria contains an opsin gene that has an unusual H^+-pumping motif, DTG (G for glycine),
334 forming a clade with another ME sequence with the DTG motif. This clade is phylogenetically
335 closer to the sensory inward H^+-pumping xenorhodopsins (XeR) than to other well-studied
336 rhodopsin groups (**Figures 1** and **S1**). We tentatively named it "MycoR" as it was identified in the
337 Myco tribe. Many more DTG opsin sequences belonging to this cluster were found in numerous
338 stratfreshDB MAGs classified to Mycobacteriales and Solirubrobacterales within Actinobacteria,
339 and notably most of these MAGs were recovered from clearwater environments (**Figures 2a** and
340 **2b**). DTG rhodopsins were initially discovered in soil- and plant-associated Proteobacteria (Harris
341 *et al.* 2015; Sudo & Yoshizawa 2016; Suzuki *et al.* 2022) and our findings expand the taxonomic
342 diversity of DTG rhodopsins.

343

344 ***Wider than previously thought opsin gene distribution in Bacteroidetes***

345 Actinobacteria have long been regarded as the major phylum containing opsin genes among
346 freshwater bacteria. For example, among single cells recovered from freshwater lakes, the six
347 opsin-containing Sphingobacteriia SAGs were dwarfed by the large number (~100) of opsin-
348 containing Actinobacteria SAGs (Martinez-Garcia *et al.* 2012). In our study, opsin genes from
349 Bacteroidetes were exclusively within the PR group, and correspond to MAGs placed in the
350 Sphingobacteriia, Flavobacteriia, and Cytophagia classes (**Table 1** and **Figure S1**). They were
351 mostly from low-abundance populations in ME (**Figure 1**), yet when added together, they
352 accounted for ~13% of all opsin gene coverage in the combined ME assembly (**Figure S3**).
353 Notably, in both Mendota and clear waters from the stratfreshDB Actinobacteria and Bacteroidetes
354 had the highest occurrences of opsin-containing MAGs among all MAGs (**Tables S6** and **S7**),
355 suggesting that, similar to Actinobacteria, Bacteroidetes in clear waters also widely possess opsin
356 genes, and therefore, could be a previously overlooked photoheterotrophic lineage in clearwater
357 lakes.

358

359 ***Opsin gene diversity in Proteobacteria***

360 Proteobacteria are abundant and prevalent in freshwater, which is also reflected by the large
361 numbers of Proteobacteria MAGs recovered from Mendota, Trout Bog and the stratfreshDB
362 dataset. Yet compared to Actinobacteria and Bacteroidetes, relatively fewer opsin sequences were
363 recovered from this phylum, especially in humic water (**Tables S6** and **S7**). For example, the
364 ubiquitous *Polynucleobacter* spp., which frequently occur in humic water, mostly lack opsin genes.
365 Freshwater proteobacterial opsin genes in our study belong to the PR and XR groups (**Figures 1**

366 and **S3**). Notably, several un-binned metagenome sequences in the PR branch, including one
367 occurring at an extremely high abundance, fall within the freshwater *Candidatus* Fonsibacter
368 (LD12) (**Figures 1** and **S1**). LD12 is a sister clade of marine SAR11, which is the major contributor
369 to rhodopsins in the ocean (Maresca *et al.* 2018; Olson *et al.* 2018; Sieradzki *et al.* 2018). As LD12
370 are very abundant in Mendota, opsin genes in the LD12 clade alone contributed to nearly 11% of
371 all opsin gene coverage in the combined ME assembly. Among the stratfreshDB MAGs, besides
372 LD12, ubiquitous freshwater genera such as *Limnohabitans*, *Rhodoferax*, and "*Ca*.
373 Methylopumilus" were major bacterial taxa with opsin genes. Most proteobacterial opsins have
374 the DTE motif for pumping H^+, especially in clearwater MAGs. Members of a few taxa such as
375 LD12 and *Rhodoferax* have ion-pumping motifs other than DTE in humic water (**Figure 2a**),
376 although the typical DTE motif also occurred in other members of these taxa. For example, some
377 humic water LD12 MAGs have the DTQ (Q for glutamine) motif (**Figure 2c**). Proteobacterial
378 opsins with such unusual motifs were almost exclusively from humic water, accounting for nearly
379 40% of all proteobacterial opsin sequences from humic water MAGs in the stratfreshDB, and
380 therefore their importance in humic aquatic environments is unlikely to be negligible.
381
382 ***Opsin genes in other bacterial taxa***
383 Opsin genes belonging to Planctomycetes, Verrucomicrobia, and Chloroflexi were distributed in
384 the PR, XR and ESR branches (**Figure 1**). ESR, a branch phylogenetically closest to PR, was
385 named after the firmicute *Exiguobacterium sibiricum* which has an unusual H^+-translocating
386 motif, DTK (K for lysine) (Balashov *et al.* 2013). The ESR branch includes Firmicutes and
387 freshwater Planctomycetes, Verrucomicrobia, and Actinobacteria Luna1 and Luna2 (**Figures 1**
388 and **S1**). Like in *Exiguobacterium* spp., Planctomycetes sequences in the ESR branch also have

389 the DTK motif, and notably these planctomycetes ESR sequences were recovered from clearwater
390 environments (**Figure 2a**), although planctomycetes with the DTE motif in other branches were
391 recovered from both clear and humic waters.
392
393 Opsin genes belonging to *Candidatus* Saccharibacteria (previously candidate division TM7) were
394 recovered from un-binned contigs in Trout Bog. Their sequences are highly similar (~93% amino
395 acid identity) to the opsin gene in a genome bin classified as TM7 (TM7 FNED7-bin-18, IMG
396 genome ID 2648501560) recovered in an enrichment culture originating from the bog-like basin
397 of Grosse Fuchskuhle in Germany (**Figure S1**) (Garcia *et al.* 2014). In fact, Saccharibacteria
398 rhodopsins (SacR) were also recently discovered from stratfreshDB MAGs and function as a light-
399 driven outward H^+ pump (Jaffe *et al.* 2022). We noted that all the 46 SacR sequences from the
400 stratfreshDB also exclusively occurred in humic waters. SacR have the DTS (S for serine) H^+-
401 pumping motif and with MycoR, they form a DTG/DTS cluster (**Figures 1**, **2b, S1**).
402
403 ***Viral opsin genes***
404 Viral opsin genes were first discovered in giant viruses that infect marine unicellular eukaryotes,
405 such as Phycodnaviruses and *Phaeocystis globosa* virus in Organic Lake, an Antarctic lake with
406 extremely salty water (Yutin & Koonin 2012). They were later detected in the Red Sea and were
407 referred to as "VR" for viral rhodopsins (Philosof & Beja 2013). Protein structure characterization
408 and functional analyses revealed that they encode light-driven H^+-pumps (Needham *et al.* 2019)
409 or light-gated ion channels (Bratanov *et al.* 2019; Zabelskii *et al.* 2020). In our study, we recovered
410 a large opsin gene cluster outside of all bacterial and archaeal branches in the phylogenetic tree
411 (**Figures 1** and **S1**). Three viral bins (see **Supplementary Text** and **Tables S8** and **S9** for their

412 classification as viruses) recovered from ME each contain one to two copies of opsin genes
413 belonging to this cluster (**Tables 1** and **S5**). This cluster also harbored opsin sequences from algal
414 viruses previously recovered from marine or salty water (**Figure S1**), supporting the viral origin
415 of this cluster. These freshwater viral opsin genes also form two distinct sub-clusters, Viral Group
416 I and II as previously defined (Yutin & Koonin 2012), each harboring both freshwater and marine
417 sequences (**Figure S1**). The function of these freshwater viral rhodopsins, their roles in virus-host
418 interactions, and their influences on host physiology and ecology are yet to be studied.
419
420 **2. Retinal biosynthesis**
421 Retinal is a photosensitive chromophore cofactor of rhodopsin. One important question is whether
422 the opsin gene-containing genomes we recovered also have the genetic potential for retinal
423 biosynthesis to form functional rhodopsin. For this, we checked key genes responsible for the last
424 few steps in retinal biosynthesis, including phytoene synthesis, phytoene desaturation, lycopene
425 cyclization, and β-carotene cleavage. We found many opsin-containing MAGs also have the
426 genetic potential for retinal biosynthesis (**Table 1**). Due to incompleteness, it is difficult to
427 conclude the lack of genes based on MAGs. However, some patterns are consistent. For example,
428 all viral bins lacked retinal biosynthesis genes, and therefore, they might rely on their hosts or the
429 environment for the chromophore. In addition, among Actinobacteria, retinal biosynthesis genes
430 were mostly found within acI, particularly acI-A and acI-B. This trend is supported by a total of
431 74 freshwater Actinobacteria genomes, including SAGs and MAGs from other studies that span
432 most freshwater Actinobacteria lineages (**Table S10**). Although opsin genes are widely distributed
433 among these lineages (except acV), oxygenase genes responsible for β-carotene cleavage are only
434 present in some tribes within the acI-A and acI-B clades (**Table S10**). Thus, Actinobacteria without

435 the retinal biosynthesis capability need to take up retinal from the environment if their opsins are
436 to be photoactive. For example, the freshwater Actinobacterium *Rhodoluna lacicola* WH-Ta8^{T}
437 conducted light-induced H^{+} pumping only after exogenous retinal was provided (Keffer *et al.*
438 2015).
439
440 **3. Ecological distribution of opsin genes.**
441 We estimated the relative abundance of opsin genes in metagenomes by comparing the coverage-
442 weighted abundance of opsin genes to single-copy essential housekeeping genes that have
443 comparable lengths to opsin genes. As opsin genes in our MAGs are mostly single-copy (**Table**
444 **S5**) and viral contribution to the total opsin genes is relatively small (**Figure S3**), such estimation
445 largely reflects the fraction of bacteria that possess opsin genes in the total community. We
446 performed this estimation using individual metagenomes generated from each time point from ME,
447 TE, and TH. Comparison between the two epilimnia indicated that opsin gene abundance was
448 higher in ME (~70%) than in TE (~14%) (**Figure 3b**). We therefore hypothesized that opsin gene
449 distribution could be influenced by lake DOC and optical properties, in particular, that high DOC
450 and light-absorbing humic substances and thus low light availabilities can lead to the low
451 occurrence of opsin genes, whereas low DOC and good light penetration may enrich for opsin
452 genes.
453
454 **Comparison among the surface layers of 33 lakes**
455 To test this hypothesis, we included a total of 33 freshwater systems for which metagenome data
456 and DOC levels are publicly available (**Table S2**). These lakes range from oligotrophic,
457 mesotrophic, and eutrophic clearwater (not stained brown) lakes, to dystrophic brown-water humic

458 lakes, spanning a wide range of DOC levels and exhibiting different clarities. Examples of lake
459 DOC levels, clarity as reflected in the Secchi depth and light diffuse attenuation coefficient, and
460 other physicochemical and limnological characteristics are listed in **Table S1**. As opsin gene
461 abundance can be influenced by water depth as discussed later, only metagenomes collected from
462 the surface layer were included in this analysis to reduce variables. For lakes from the stratfreshDB,
463 which largely consists of small boreal lakes, metagenomes collected within the first meter were
464 included. An inverse correlation (Pearson correlation coefficient r = -0.77, *p*-value = 1.4e-7)
465 between opsin gene abundance and DOC was observed (**Figure 4**). Opsin gene abundance was >50%
466 in clearwater lakes irrespective of their trophic status. Especially, the abundance was extremely
467 high (~90%) in the Laurentian Great Lakes (Podowski *et al.* 2021) and Lake Lanier (Dam *et al.*
468 2020) where DOC levels are low (~2 mg/L). Opsin genes were much less prominent in humic
469 lakes with high terrestrial DOC. Extremely low abundance (<10%) was observed in the surface
470 water of Lake Mekkojärvi, a highly humic boreal lake with ~30 mg/L of DOC (Buck *et al.* 2021).
471

472 **Comparison at different depths within lakes**

473 Within Trout Bog, comparing individual metagenomes sampled on the same day indicates the
474 epilimnion had a higher opsin gene abundance than the hypolimnion, and the difference between
475 layers is statistically significant (paired student T-test *p*-value < 1e-8, n = 29) (**Figure 3b**). Indeed,
476 analyses of depth-discrete metagenome profiles show a steep decline of opsin gene abundance
477 within the first few meters in humic lakes, leading to extremely low opsin gene levels in the deeper
478 layer (**Figure 5**). By contrast, such a sharp decrease was not observed in clearwater lakes, where
479 a depth-associated change in opsin gene abundance was much more moderate and slower (**Figure**
480 **5**).

481

482 **4. Opsin gene expression and diel trends**

483 We used metatranscriptomes previously generated from the epilimnia of Mendota, Trout Bog, and
484 Sparkling Lake by Linz *et al.* (2020) to study opsin gene expression and diel cyclic rhythms. From
485 metatranscriptome mapping outputs in the original analyses by Linz *et al.* (2020), the total opsin
486 transcripts are about 0.5%, 0.3%, and 0.02% of the total mRNA for Mendota, Sparkling Lake, and
487 Trout Bog respectively when averaged from all time points, and notably, a few opsin genes were
488 among the top 500 most highly expressed genes within the entire metatranscriptomes in Mendota
489 and Sparkling Lake. This indicates that opsin genes were highly expressed in these two clearwater
490 lakes.

491

492 In our current analysis with more tailored and curated references for opsin genes, we found that
493 201 out of the 396 opsin genes in the Mendota opsin reference set had expression levels higher
494 than the cutoff, and these numbers were 105 out of 173 and 18 out of 69 for Sparkling Lake and
495 Trout Bog respectively. The total opsin transcripts per liter of lake water averaged from all time
496 points showed that the total expression in the two clearwater lakes was about an order of magnitude
497 higher than that in the humic bog (**Figure 6**). Bacterial phyla contributing to the most opsin
498 expression were Actinobacteria, followed by Bacteroidetes and Proteobacteria in the two
499 clearwater lakes, whereas opsin expression from Proteobacteria was minimal in the humic bog.

500

501 Ranking individual opsin genes by their average transcript abundances indicated that opsin from
502 Actinobacteria (mostly acI-A and acI-B), Bacteroidetes, and Proteobacteria (LD12 in particular)
503 are among the most highly expressed opsins in the two clearwater lakes, whereas the most highly

504 expressed opsins in Trout Bog were mainly from Actinobacteria acI-B (**Table S11**). Notably, both
505 groups of AcidiR were expressed in the two clearwater lakes.
506
507 Among the expressed opsin genes, some exhibited diel trends. Depending on the
508 photosynthetically active radiation (PAR) measured just below the surface of lake water, samples
509 were categorized into "day" and "night" groups (Linz *et al.* 2020). Some taxa-specific diel patterns
510 emerged. Cyclic genes from Actinobacteria tend to have peak expression at night, and
511 Bacteroidetes opsins tend to peak at daytime in both Mendota and Sparkling Lake (**Figure 7**). In
512 Trout Bog, only eight opsin genes were identified as cyclically transcribed. Seven of them belong
513 to Actinobacteria and showed increased expression during the night and peaked before sunrise
514 (**Figure S5**). However, Trout Bog results need further verification as only one day of data is
515 available for this lake (Linz *et al.* 2020).
516
517 **DISCUSSION**
518 **Diversity of opsin genes in clear and humic waters**
519 Despite numerous studies on freshwater opsin genes, their diversity in brown-color, often acidic
520 humic lakes had been largely unexplored, and the question remained about the difference, if any,
521 between clear and humic waters with contrasting pH, DOC quantity/quality, and optical conditions.
522
523 Here we focused on the two main lakes which we have rich time-series metagenome data (and
524 metatranscriptome data) to investigate the diversity and quantify the contribution of different taxa
525 to the total opsin genes. Owing to the large number of MAGs from the stratfreshDB, we gained a
526 more complete view of opsin diversity in both clear and humic waters. In general, findings from

527 the stratfreshDB are consistent with what we observed with the Mendota and Trout Bog dataset.
528 In both datasets, most opsin genes were predicted to encode green light-absorbing H^+ pumps with
529 the typical DTE variant for H^+ donating and accepting, and Actinobacteria is the most abundant
530 contributor of opsin genes in both humic and clearwater environments. Among the known taxa
531 with opsin genes, we found wider than previously thought distribution within freshwater
532 Bacteroidetes and identified novel clades, such as LD12 with the DTQ motif and MycoR. Our
533 analysis also expanded the taxonomic spectra of freshwater opsins, especially in branches with
534 unusual H^+-pumping motifs.
535
536 Some differences between clear and humic waters emerged. Opsin genes were found more
537 frequently in taxa from clear waters than in taxa from humic waters, which could lead to an
538 observed low taxonomic diversity of opsin genes in a given humic lake, as we observed from Trout
539 Bog. In addition, some opsins have non-typical H^+-pumping motifs that might be associated with
540 the two different lake types as discussed below (**Figure 2a**).
541
542 The DTG rhodopsin was previously recovered in soil- and plant-associated Proteobacteria as an
543 light-driven outward H^+-pump; and its photochemical behavior is pH dependent and H^+-pumping
544 slows down at high extracellular pH, which might prevent intracellular alkalization in alkalized
545 environments (Harris *et al.* 2015; Sudo & Yoshizawa 2016; Suzuki *et al.* 2022). DTG rhodopsins
546 are phylogenetically close to DTS rhodopsins, forming a DTG/DTS cluster (Suzuki *et al.* 2022).
547 The DTS rhodopsin from *Sphingomonas paucimobilis* functions as an outward H^+-pump at pH
548 lower than 7.0 and its photocycle speed is also pH dependent and slows down with increasing
549 extracellular pH (Maliar *et al.* 2020). The pH dependance of DTG/DTS rhodopsins might be due

550 to the lack of a carboxylic H^+-donating residue as in the DTE/DTD variants (Suzuki *et al.* 2022),
551 and therefore it is plausible that rhodopsins in the DTG/DTS cluster are all pH-dependent. In our
552 study, we found DTG opsin genes in Mycobacteriales and Solirubrobacterales within
553 Actinobacteria, and 121 out of the 130 DTG sequences were from four clearwater lakes with an
554 average pH of 7.6 (averaged from these lakes at depths where these sequences were recovered)
555 (**Figure 2b**). Notably, these four lakes have surface water pH values around 8, which is higher
556 than most other lakes included in the stratfreshDB (Buck *et al.* 2021). By contrast, we found DTS
557 opsin sequences all belonging to Saccharibacteria (SacR) exclusively in humic waters and the
558 average pH of these SacR-occurring waters is 5.2 (**Figure 2b**). Notably, these SacR sequences
559 were initially discovered from the same set of MAGs, functioning as an outward H^+-pump (Jaffe
560 *et al.* 2022). As characterized DTG and DTS rhodopsins are both pH-dependent outward H^+ pumps
561 that function more efficiently at low extracellular pH than at high pH, their recovery from non-
562 neutral pH freshwaters in our study is interesting, yet their contrasting occurrence patterns between
563 slightly alkaline and acidic environments seems contradicting. This contradiction may reflect the
564 different aspects of the pH dependance. On one side, stopping pumping out H^+ at high extracellular
565 pH is beneficial to avoid further cytoplasmic alkalization in slightly alkaline lakes where natural
566 fluctuations of pH can sometimes impose an alkaline pH stress on bacterial cells. On the other side,
567 efficient pumping out H^+ at low extracellular pH is advantageous for cells to generate energy
568 and/or prevent cytoplasmic acidification in acidic humic waters. This may probably explain the
569 contrasting occurrence patterns of DTG and DTS variants between slightly alkaline clear waters
570 and acidic humic waters and may suggest their association with habitat pH conditions.
571

572 In our study, the DTQ motif was exclusively from members of LD12 in humic waters, whereas
573 the typical DTE variant from LD12 occurred in both clear and humic waters. The DTE and DTQ
574 variants were also present in SAR11 (the marine sister clade of LD12), with DTE as the dominant
575 variant in the surface water (0-200 m) (Olson *et al.* 2018). In the same study, a deep-water SAR11
576 DTQ opsin, when expressed in *Escherichia coli*, functioned as a light-driven outward H^+-pump,
577 and SAR11 DTQ opsins were expressed especially in deep sea (500–1000 m). This observation,
578 together with the exclusive occurrence of DTQ variants of LD12 in humic waters, may suggest the
579 association of the DTQ variant with poor light conditions.
580
581 Unknown variants (i.e. ATT, ATV, or EGV motifs) were exclusively found in humic waters.
582 Interestingly these sequences have the hydrophilic glutamine (Q) residue for spectral tuning
583 (**Figure 2a**) and may represent novel rhodopsins with different functions and/or photochemical
584 behaviors.
585
586 As a summary, opsin genes are more widely distributed among clearwater taxa than humic taxa.
587 Conversely, the ion-translocating motif seemed more diverse in humic than in clear waters. The
588 different distributions of these motifs might be associated with the pH, light, or other
589 physicochemical contrasts between clear and humic waters. Further studies are needed to confirm
590 this.
591
592 **Significance of rhodopsins in clearwater lakes**
593 Opsin genes are widely distributed in marine environments where they often occur in more than
594 half of the cells in a wide range of samples (Finkel *et al.* 2013; Maresca *et al.* 2018; Olson *et al.*

595 2018). Some have suggested that opsin genes are more abundant in marine than in freshwater
596 systems (Finkel *et al.* 2013; Oh *et al.* 2011). In particular, opsin gene abundance was positively
597 correlated to salinity (Brindefalk *et al.* 2016; Maresca *et al.* 2018). However, a very limited number
598 of freshwater systems were included in these studies.
599
600 We found that in contrast to previous predictions opsin genes are indeed prevalent in clearwater
601 lakes with low DOC and good light penetration. Opsin gene abundance was usually higher than
602 60% in lakes with DOC less than 5 mg/L. In particular opsin gene abundance is remarkably high
603 (~90%) in the Laurentian Great Lakes, where DOC is around 2 mg/L and the terrestrial influence
604 is small due to the large surface area. This suggests that rhodopsin-based phototrophy is an
605 important light-harvesting mechanism and its significance may have been underestimated in such
606 lakes. Contrasting to that, opsin genes appear to be less important overall in humic lakes which are
607 often associated with a large terrestrial influence, high allochthonous DOC, brown water color,
608 and a shallow photic zone as discussed in more detail later.
609
610 **Opsin gene expression and diel regulation**
611 In Mendota and Sparkling Lake, the majority of bacterial opsin genes were expressed. The total
612 opsin transcripts accounted for 0.3-0.5% of all mRNA, and a few opsin genes were among the
613 most highly expressed genes in these two lakes. These observations support the importance of
614 rhodopsins in bacterial ecophysiology in the surface water of clearwater lakes. Although
615 rhodopsins may also be important in populations that express opsin genes in the bog, the higher
616 total expression of opsin genes in the two clearwater lakes compared to in the humic bog suggests
617 the overall greater significance of rhodopsin-based phototrophy in clearwater lakes.

618

619 Opsin gene regulation in freshwater has until now largely unexplored. In marine systems, some
620 studies showed that opsin gene expression exhibited diel rhythms, with maximal expression at the
621 darker periods instead of peak sunlight (e.g. pre-sunrise, post-sunrise, pre-sunset, and post-sunset)
622 (Ottesen *et al.* 2014; Zehnpfennig *et al.* 2022), whereas others suggested opsin genes may be
623 constitutively expressed (Riedel *et al.* 2010; Shi *et al.* 2011). In freshwater, an earlier study of
624 Actinorhodopsin expression indicated their maximal expression at dawn and minimal expression
625 at dusk for samples collected every six hours (Wurzbacher *et al.* 2012). In our study, samples were
626 collected every four hours for two days. From both clearwater lakes, we observed consistent
627 patterns showing diel cyclic behaviors differing among major taxa. Specifically, Actinobacterial
628 opsin genes tend to peak at darker periods of the day, and Bacteroidetes tend to peak during the
629 daytime, especially at peak sunlight.

630

631 Taxa-specific diel patterns might suggest differences in a circadian-like clock among bacteria. As
632 opsin gene expression can be regulated by light (Lami *et al.* 2009), carbon availability and nutrient
633 availability (Kwon *et al.* 2013; Wang *et al.* 2012), the taxa-specific patterns might also suggest
634 dual regulation of opsin genes by both light regime and carbon/nutrient availabilities, as the latter
635 can also experience diel cascade changes following photosynthesis in the lake (Linz *et al.* 2020).
636 We should note that only a portion of the expressed opsin genes exhibited a statically significant
637 diel pattern, contributing to about half of all opsin transcripts in the two clearwater lakes (**Figure**
638 **6**). It is not clear whether those without cyclic trends are constitutively expressed as some studies
639 suggested (Keffer *et al.* 2015; Riedel *et al.* 2010) or if their diel changes are too small to be
640 detected given the noise in the metatranscriptome data.

641

642 **Influence of DOC and optical properties on opsin gene distribution**

643 Lake optical properties are strongly influenced by DOC quality/quantity (Maizel *et al.* 2017; Read
644 & Rose 2013), and in particular, a linear relationship between the light attenuation coefficient for
645 PAR and lake DOC was previously revealed (Read & Rose 2013). Humic lakes often have higher
646 DOC levels than clearwater lakes, as they receive large amounts of terrestrially-derived
647 allochthonous DOC, most of which are in the forms of humic and fulvic acids. These carbon
648 molecules are light-absorbing chromophoric dissolved organic matter (CDOM), mainly absorbing
649 light at wavelengths below 500 nm, including UV-A, UV-B, and the blue light spectrum of
650 photosynthetically active radiation (400-700nm) (Markager & Vincent 2000). In particular, the
651 high concentrations of gilvin and tripton (mostly dissolved and particulate humic substances,
652 respectively) in peat lakes were shown to rapidly absorb blue and green light (Stomp *et al.* 2007).
653 The light spectral distribution can influence phototroph distribution as revealed previously, as
654 increasing humic substance concentration in freshwater increased phototrophs absorbing red light
655 and decreased those absorbing the central part of the PAR spectrum (500–600 nm) (Vila *et al.*
656 1998). Therefore, as expected, we found that the inferred lower availability of green light strongly
657 limited opsin gene occurrence in humic lakes since the encoded rhodopsins in our analysis were
658 predicted to absorb green light.

659

660 Light availability also changes with water depth and such changes differ between clearwater and
661 humic lakes. Opsin gene abundance similarly varied with depth. Depth-associated decreases in
662 opsin gene abundance were relatively moderate and gradual in clearwater lakes. An extreme
663 example is from Lake Michigan, one of the Laurentian Great Lakes with low DOC (~2 mg/L),

664 where opsin genes remained abundant throughout the water column to 110 m deep. In Lake
665 Tanganyika, the opsin gene abundance estimate was around 30% even at 1100 m. These findings
666 are similar to marine studies, where sunlight penetrates deep into the open ocean, and opsin genes
667 and rhodopsin pigments were abundant in the photic zone (the upper 200 m) (Gomez-Consarnau
668 *et al.* 2019; Olson *et al.* 2018) and opsin genes were still present and expressed at deeper depths
669 (down to 1000 meters, albeit at low abundances) (Olson *et al.* 2018). Therefore, in very large and
670 clear lakes like Michigan and Tanganyika, rhodopsin-based phototrophy may contribute
671 significantly to the total phototrophy and energy budgets within the lake, similar to its counterpart
672 in the open ocean. By contrast, opsin gene abundance quickly declined to very low levels within
673 the first few meters in humic lakes. Therefore, humic lakes not only have lower opsin abundance
674 in the surface layer, but also have a much shallower opsin-containing zone, both contributing to a
675 lower significance of rhodopsin-enabled phototrophy in such environments.
676
677 The consistency between light climate and opsin gene profiles with depth suggests the importance
678 of lake optical properties in opsin gene distribution. At the same time, the inverse correlation
679 between opsin gene abundance and DOC levels for samples collected near the surface suggests a
680 strong influence of DOC on opsin gene occurrence. However, as DOC, especially CDOM, and
681 optical properties are generally correlated, it is difficult to tease apart their individual influence on
682 opsin gene distribution from our data. It is plausible that CDOM indirectly influences opsin gene
683 distribution by impacting light availability. In addition, we also postulate a direct influence of DOC
684 on opsin gene distribution, as low DOC may select for opsin-containing bacteria as discussed later.
685

686 We should note that DOC not only influences lake optical properties but also influences lake water
687 pH and dissolved oxygen (DO). Humic lakes often have slightly acidic pH (~5-6, **Table S1**), due
688 to the high levels of humic and fulvic acids. In addition, DO usually decreases sharply along the
689 depth due to lower oxygenic photosynthesis and higher biological oxygen demand as compared to
690 clearwater lakes, leading to a shallow oxygenated zone. Therefore, compared among lakes, acidic
691 pH seems to be associated with low opsin abundance, and compared within a lake, low DO is
692 associated with low opsin abundance. Given that rhodopsin is a light-sensitive protein, the pH and
693 DO influence on opsin gene abundance is likely less important than light availability, although we
694 cannot exclude their potential impacts based on the available data.
695
696 **Ecological roles and opsin gene selection in freshwater**
697 Light availability may be one of the main reasons for the wide distribution of opsin genes in
698 clearwater lakes. At the same time, we speculate that competitive advantages for surviving
699 provided by rhodopsins may favor their selection in clearwater lakes. Nearly all bacterial opsin
700 genes that we analyzed encode H^+-pumps, generating proton motive force (PMF), which
701 contributes to bacterial energy metabolism and renders physiological advantages, especially when
702 bacteria are under stress or carbon-limited conditions as discussed below.
703
704 As many vitamins are vulnerable to UV light, vitamins in surface water are subjected to
705 photodegradation in the photic zone. This may impose pressure on microbes to quickly take up
706 vitamins as soon as they are available, particularly in clearwater lakes. Proteorhodopsin gene
707 expression was linked to the light-enhanced growth of two marine vitamin-B_1 auxotrophic
708 Flavobacteria strains by providing PMF to power the Ton-B dependent acquisition of vitamin-B_1

709 (Gomez-Consarnau *et al.* 2016). Therefore, bacteria with H^+-pumping rhodopsins may compete
710 better for vitamin uptake compared to bacteria without this additional PMF-generating revenue,
711 and are selected for in clearwater lakes where bacteria are subjected to more intense UV radiation
712 than in brown-water humic bogs.
713
714 Rhodopsin-containing bacteria are photoheterotrophs relying on organic carbon for growth.
715 Timely carbon substrate scavenging is a key for heterotrophs to succeed under substrate-limited
716 conditions. Due to the additional PMF/ATP generated from light energy, bacteria with H^+-
717 pumping rhodopsins may be able to enhance the active uptake of organic substrates. More
718 importantly, the energy captured from the sunlight may reduce their reliance on respiration. For
719 example, when organic carbon became limiting, the respiration rate of '*Ca.* Pelagibacter ubique'
720 was lower under light as compared to in the dark while maintaining comparable cell densities
721 (Steindler *et al.* 2011). Therefore, possessing rhodopsins may be more advantageous in carbon-
722 limited than in carbon-rich environments. For example, some rhodopsin-possessing bacteria
723 exhibited light-promoted growth on low organic carbon media or light-enhanced survival under
724 carbon starvation (Gómez-Consarnau *et al.* 2010; Gómez-Consarnau *et al.* 2007; Johnson *et al.*
725 2010; Steindler *et al.* 2011; Yoshizawa *et al.* 2014). As lakes with high DOC, particularly humic
726 bogs, are not carbon-limited, whereas many clearwater lakes are, rhodopsin-containing bacteria
727 may be strongly selected for in clearwater lakes. This may partly and directly contribute to the
728 inverse correlation between opsin gene abundance and lake DOC levels, in addition to the indirect
729 impact of DOC through its influence on the availability of rhodopsin-absorbing spectra and UV
730 light.
731

732 **Rhodopsin-based phototrophy in carbon and energy fluxes in freshwaters**

733 We propose that rhodopsin-based phototrophy influences the carbon and energy fluxes in
734 freshwaters, and the significance of this influence differs between clearwater and humic lakes, as
735 illustrated in **Figure 8**. In clearwater lakes, a large fraction of heterotrophs are actually
736 photoheterotrophs using rhodopsins to harvest sunlight while respiring DOC. As DOC may be
737 limited by photoautotrophic primary producers or present in low levels in clearwater lakes,
738 rhodopsin-phototrophy may be a significant pathway for harvesting energy in such lakes. By
739 contrast, in humic lakes, a large amount of light energy is absorbed by CDOM and particulate
740 humic substances, and such photochemical reactions generate labile DOC, which can be used for
741 heterotrophic respiration (Cory & Kling 2018). As such, although the photodegradation process
742 leaves a small amount of light for rhodopsin-phototrophy, photodegradation provides bacteria with
743 a steady labile organic substrate supply for energy generation. Therefore, compared to clearwater
744 lakes, the contribution of heterotrophy is larger, whereas the contribution of light harvesting by
745 rhodopsins is smaller in the total energy budget in humic lakes. This further influences carbon flux,
746 as the greater reliance on organic carbon based respiration increases the net emission of CO_2 from
747 humic lakes. It was shown that CO_2 emission and net heterotrophy were associated with
748 allochthonous DOC concentrations, and were higher in humic than in clearwater lakes (Jonsson *et*
749 *al.* 2003). From our study, it is possible that rhodopsin-based phototrophy may be a previously
750 overlooked factor contributing to such a difference in CO_2 emission from these two lake types.

751

752 **SUMMARY**

753 In this study, by analyzing opsin genes recovered in metagenomes and MAGs collected from clear
754 and humic waters with contrasting pH, optical properties and DOC quantity/quality, we revealed

755 differences in opsin diversity between two major freshwater types and recovered novel diversity
756 within freshwater opsin genes. Many of these opsin genes were expressed, and some of them
757 exhibited diel cyclic rhythms that differ among taxa. Further, we found opsin gene abundance is
758 associated with lake DOC and depth in the water column, suggesting their ecological distribution
759 can be influenced by organic matter and lake optical properties. In particular, opsin genes are
760 widely present and highly expressed in clearwater lakes, and may be strongly selected for due to
761 the higher light availability, lower DOC, and thus greater competitive advantages provided by
762 rhodopsins in such environments. Therefore, we assert that rhodopsin-based phototrophy is a
763 significant light harvesting mechanism, substantially influencing energy and carbon budgets in
764 clearwater lakes.
765

766 **ACKNOWLEDGEMENTS**

767 We thank the North Temperate Lakes Microbial Observatory 2007-2012 field crews, UW-Trout
768 Lake Station, the UW Center for Limnology, and the Global Lakes Ecological Observatory
769 Network for field and logistical support. We acknowledge efforts by many McMahon Lab
770 undergrads and technicians related to sample collection and DNA extraction. We thank the
771 individual program directors and leadership at the National Science Foundation for their
772 commitment to the continued support of long-term ecological research. We also thank Dr.
773 Andrey Rozenberg for providing valuable and helpful comments on the preprint of this manuscript.
774 The project was supported by funding from the United States National Science Foundation
775 Microbial Observatories program (MCB-0702395), the Long-Term Ecological Research program
776 (NTL-LTER DEB-1440297), an INSPIRE award (DEB-1344254) and U.S. Department of
777 Agriculture (Hatch Project WIS03052). This material is also based upon work supported by the

National Institute of Food and Agriculture, U.S. Department of Agriculture (Hatch Project 1002996). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

Atamna-Ismaeel N, Sabehi G, Sharon I, *et al.* (2008) Widespread distribution of proteorhodopsins in freshwater and brackish ecosystems. *ISME J* **2**, 656-662.

Balashov SP, Imasheva ES, Boichenko VA, *et al.* (2005) Xanthorhodopsin: A Proton Pump with a Light-Harvesting Carotenoid Antenna. *Science* **309**, 2061-2064.

Balashov SP, Petrovskaya LE, Imasheva ES, *et al.* (2013) Breaking the carboxyl rule: lysine 96 facilitates reprotonation of the Schiff base in the photocycle of a retinal protein from Exiguobacterium sibiricum. *J Biol Chem* **288**, 21254-21265.

Béjà O, Aravind L, Koonin EV, *et al.* (2000) Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science* **289**, 1902-1906.

Beja O, Lanyi JK (2014) Nature's toolkit for microbial rhodopsin ion pumps. *Proc Natl Acad Sci U S A* **111**, 6538-6539.

Bendall ML, Stevens SLR, Chan L-K, *et al.* (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J*.

Bratanov D, Kovalev K, Machtens JP, *et al.* (2019) Unique structure and function of viral rhodopsins. *Nature Communications* **10**.

Brindefalk B, Ekman M, Ininbergs K, *et al.* (2016) Distribution and expression of microbial rhodopsins in the Baltic Sea and adjacent waters. *Environ Microbiol* **18**, 4442-4455.

Buck M, Garcia SL, Fernandez L, *et al.* (2021) Comprehensive dataset of shotgun metagenomes from oxygen stratified freshwater lakes and ponds. *Scientific Data* **8**.

Cabello-Yeves PJ, Ghai R, Mehrshad M, *et al.* (2017) Reconstruction of Diverse Verrucomicrobial Genomes from Metagenome Datasets of Freshwater Reservoirs. *Frontiers in Microbiology* **8**.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**.

Chazan A, Rozenberg A, Mannen K, *et al.* (2022) Diverse heliorhodopsins detected via functional metagenomics in freshwater Actinobacteria, Chloroflexi and Archaea. *Environ Microbiol* **24**, 110-121.

Cory RM, Kling GW (2018) Interactions between sunlight and microorganisms influence dissolved organic matter degradation along the aquatic continuum. *Limnology and Oceanography Letters* **3**, 102-116.

Dam P, Rodriguez-R LM, Luo CW, *et al.* (2020) Model-based Comparisons of the Abundance Dynamics of Bacterial Communities in Two Lakes. *Scientific Reports* **10**.

Darling AE, Jospin G, Lowe E, *et al.* (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243.

Denef VJ, Mueller RS, Chiang E, Liebig JR, Vanderploeg HA (2016) Chloroflexi CL500-11 Populations That Predominate Deep-Lake Hypolimnion Bacterioplankton Rely on Nitrogen-Rich Dissolved Organic Matter Metabolism and C(1) Compound Oxidation. *Applied and Environmental Microbiology* **82**, 1423-1432.

Durán-Viseras A, Andrei A-S, Ghai R, Sánchez-Porro C, Ventosa A (2019) New Halonotius Species Provide Genomics-Based Insights Into Cobalamin Synthesis in Haloarchaea. *Frontiers in Microbiology* **10**.

Dwulit-Smith JR, Hamilton JJ, Stevenson DM, *et al.* (2018) acI Actinobacteria Assemble a Functional Actinorhodopsin with Natively Synthesized Retinal. *Applied and Environmental Microbiology* **84**.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113-113.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461.

Finkel OM, Beja O, Belkin S (2013) Global abundance of microbial rhodopsins. *ISME J* **7**, 448-451.

Flores-Uribe J, Hevroni G, Ghai R, *et al.* (2019) Heliorhodopsins are absent in diderm (Gram-negative) bacteria: Some thoughts and possible implications for activity. *Environ Microbiol Rep* **11**, 419-424.

Galachyants AD, Krasnopeev AY, Podlesnaya GV, *et al.* (2021) Diversity of Aerobic Anoxygenic Phototrophs and Rhodopsin-Containing Bacteria in the Surface Microlayer, Water Column and Epilithic Biofilms of Lake Baikal. *Microorganisms* **9**.

Garcia SL, McMahon KD, Grossart H-P, Warnecke F (2014) Successful enrichment of the ubiquitous freshwater acI Actinobacteria. *Environ Microbiol Rep* **6**, 21-27.

Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F (2014) Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molecular Ecology* **23**, 6073-6090.

Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *Isme Journal* **5**, 461-472.

Gómez-Consarnau L, Akram N, Lindell K, *et al.* (2010) Proteorhodopsin Phototrophy Promotes Survival of Marine Bacteria during Starvation. *PLoS Biology* **8**, e1000358.

Gómez-Consarnau L, Gonzalez JM, Coll-Llado M, *et al.* (2007) Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. *Nature* **445**, 210-213.

Gomez-Consarnau L, Gonzalez JM, Riedel T, *et al.* (2016) Proteorhodopsin light-enhanced growth linked to vitamin-B1 acquisition in marine Flavobacteria. *ISME J* **10**, 1102-1112.

Gomez-Consarnau L, Raven JA, Levine NM, *et al.* (2019) Microbial rhodopsins are major contributors to the solar energy captured in the sea. *Science Advances* **5**.

Guindon S, Dufayard JF, Lefort V, *et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321.

Harris A, Ljumovic M, Bondar AN, *et al.* (2015) A new group of eubacterial light-driven retinal-binding proton pumps with an unusual cytoplasmic proton donor. *Biochim Biophys Acta* **1847**, 1518-1529.

He S, Malfatti SA, McFarland JW, *et al.* (2015) Patterns in wetland microbial community composition and functional gene repertoire associated with methane emissions. *mBio* **6**, e00066-00015.

Imasheva ES, Balashov SP, Choi AR, Jung K-H, Lanyi JK (2009) Reconstitution of Gloeobacter violaceus Rhodopsin with a Light-Harvesting Carotenoid Antenna. *Biochemistry* **48**, 10948-10955.

Jaffe AL, Konno M, Kawasaki Y, *et al.* (2022) Saccharibacteria harness light energy using type-1 rhodopsins that may rely on retinal sourced from microbial hosts. *The ISME Journal* **16**, 2056-2059.

Johnson ET, Baron DB, Naranjo B, *et al.* (2010) Enhancement of survival and electricity production in an engineered bacterium by light-driven proton pumping. *Appl Environ Microbiol* **76**, 4123-4129.

Jonsson A, Karlsson J, Jansson M (2003) Sources of Carbon Dioxide Supersaturation in Clearwater and Humic Lakes in Northern Sweden. *Ecosystems* **6**, 224-235.

Keffer JL, Hahn MW, Maresca JA (2015) Characterization of an Unconventional Rhodopsin from the Freshwater Actinobacterium Rhodoluna lacicola. *J Bacteriol* **197**, 2704-2712.

Kwon SK, Kim BK, Song JY, *et al.* (2013) Genomic makeup of the marine flavobacterium Nonlabens (Donghaeana) dokdonensis and identification of a novel class of rhodopsins. *Genome Biol Evol* **5**, 187-199.

Lami R, Cottrell MT, Campbell BJ, Kirchman DL (2009) Light-dependent growth and proteorhodopsin expression by Flavobacteria and SAR11 in experiments with Delaware coastal waters. *Environ Microbiol* **11**, 3201-3209.

Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. *Mol Biol Evol* **25**, 1307-1320.

Linz AM, Aylward FO, Bertilsson S, McMahon KD (2020) Time-series metatranscriptomes reveal conserved patterns between phototrophic and heterotrophic microbes in diverse freshwater systems. *Limnology and Oceanography* **65**, S101-S112.

Linz AM, He SM, Stevens SLR, *et al.* (2018) Freshwater carbon and nutrient cycles revealed through reconstructed population genomes. *Peerj* **6**.

Maizel AC, Li J, Remucal CK (2017) Relationships Between Dissolved Organic Matter Composition and Photochemistry in Lakes of Diverse Trophic Status. *Environmental Science & Technology* **51**, 9624-9632.

Maliar N, Okhrimenko IS, Petrovskaya LE, *et al.* (2020) Novel pH-Sensitive Microbial Rhodopsin from Sphingomonas paucimobilis. *Doklady Biochemistry and Biophysics* **495**, 342-346.

Man D, Wang W, Sabehi G, *et al.* (2003) Diversification and spectral tuning in marine proteorhodopsins. *The EMBO Journal* **22**, 1725-1731.

Maresca JA, Miller KJ, Keffer JL, Sabanayagam CR, Campbell BJ (2018) Distribution and Diversity of Rhodopsin-Producing Microbes in the Chesapeake Bay. *Applied and Environmental Microbiology* **84**.

Markager S, Vincent WF (2000) Spectral light attenuation and the absorption of UV and blue light in natural waters. *Limnology and Oceanography* **45**, 642-650.

Markowitz VM, Chen I-MA, Palaniappan K, *et al.* (2013) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research*.

907 Martinez-Garcia M, Swan BK, Poulton NJ, *et al.* (2012) High-throughput single-cell
908 sequencing identifies photoheterotrophs and chemoautotrophs in freshwater
909 bacterioplankton. *ISME J* **6**.
910 Matsen FA, Kodner RB, Armbrust EV (2010) pplacer: linear time maximum-likelihood and
911 Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC*
912 *Bioinformatics* **11**, 538.
913 McCarren J, DeLong EF (2007) Proteorhodopsin photosystem gene clusters exhibit co-
914 evolutionary trends and shared ancestry among diverse marine microbial phyla.
915 *Environ Microbiol* **9**, 846-858.
916 Mizuno CM, Rodriguez-Valera F, Ghai R (2015) Genomes of Planktonic Acidimicrobiales:
917 Widening Horizons for Marine Actinobacteria by Metagenomics. *mBio* **6**.
918 Needham DM, Yoshizawa S, Hosaka T, *et al.* (2019) A distinct lineage of giant viruses brings
919 a rhodopsin photosystem to unicellular marine predators. *Proc Natl Acad Sci U S A*
920 **116**, 20574-20583.
921 Newton RJ, Kent AD, Triplett EW, McMahon KD (2006) Microbial community dynamics in a
922 humic lake: differential persistence of common freshwater phylotypes. *Environ*
923 *Microbiol* **8**, 956-970.
924 Oh S, Caro-Quintero A, Tsementzi D, *et al.* (2011) Metagenomic Insights into the Evolution,
925 Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a
926 Temperate Freshwater Ecosystem. *Applied and Environmental Microbiology* **77**,
927 6000-6011.
928 Olson DK, Yoshizawa S, Boeuf D, Iwasaki W, DeLong EF (2018) Proteorhodopsin variability
929 and distribution in the North Pacific Subtropical Gyre. *Isme Journal* **12**, 1047-1060.
930 Ottesen EA, Young CR, Gifford SM, *et al.* (2014) Multispecies diel transcriptional oscillations
931 in open ocean heterotrophic bacterial assemblages. *Science* **345**, 207-212.
932 Philosof A, Beja O (2013) Bacterial, archaeal and viral-like rhodopsins from the Red Sea.
933 *Environ Microbiol Rep* **5**, 475-482.
934 Pinhassi J, DeLong EF, Beja O, Gonzalez JM, Pedros-Alio C (2016) Marine Bacterial and
935 Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology, and Ecology.
936 *Microbiology and Molecular Biology Reviews* **80**.
937 Podowski JC, Paver SF, Newton RJ, Coleman ML (2021) Genome streamlining,
938 proteorhodopsin, and organic nitrogen metabolism in freshwater nitrifiers. *bioRxiv*,
939 2021.2001.2019.427344.
940 Pushkarev A, Inoue K, Larom S, *et al.* (2018) A distinct abundant group of microbial
941 rhodopsins discovered using functional metagenomics. *Nature* **558**, 595-+.
942 Read JS, Rose KC (2013) Physical responses of small temperate lakes to variation in dissolved
943 organic carbon concentrations. *Limnology and Oceanography* **58**, 921-931.
944 Riedel T, Tomasch J, Buchholz I, *et al.* (2010) Constitutive Expression of the Proteorhodopsin
945 Gene by a Flavobacterium Strain Representative of the Proteorhodopsin-Producing
946 Microbial Community in the North Sea. *Applied and Environmental Microbiology* **76**,
947 3187-3197.
948 Rozenberg A, Inoue K, Kandori H, Beja O (2021) Microbial Rhodopsins: The Last Two
949 Decades. In: *Annual Review of Microbiology, Vol 75, 2021* (ed. Gottesman S), pp. 427-
950 447.
951 Rusch DB, Halpern AL, Sutton G, *et al.* (2007) The Sorcerer II Global Ocean Sampling
952 expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**.

953 Sharma AK, Sommerfeld K, Bullerjahn GS*, et al.* (2009) Actinorhodopsin genes discovered in
954 diverse freshwater habitats and among cultivated freshwater Actinobacteria. *ISME J*
955 **3**, 726-737.
956 Sharma AK, Zhaxybayeva O, Papke RT, Doolittle WF (2008) Actinorhodopsins:
957 proteorhodopsin-like gene sequences found predominantly in non-marine
958 environments. *Environ Microbiol* **10**, 1039-1056.
959 Shi YM, Tyson GW, Eppley JM, DeLong EF (2011) Integrated metatranscriptomic and
960 metagenomic analyses of stratified microbial assemblages in the open ocean. *Isme*
961 *Journal* **5**, 999-1013.
962 Sieradzki ET, Fuhrman JA, Rivero-Calle S, Gomez-Consamau L (2018) Proteorhodopsins
963 dominate the expression of phototrophic mechanisms in seasonal and dynamic
964 marine picoplankton communities. *Peerj* **6**.
965 Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ (2011) Energy starved
966 Candidatus Pelagibacter ubique substitutes light-mediated ATP production for
967 endogenous carbon respiration. *PLoS One* **6**, e19725.
968 Stomp M, Huisman J, Stal LJ, Matthijs HC (2007) Colorful niches of phototrophic
969 microorganisms shaped by vibrations of the water molecule. *ISME J* **1**, 271-282.
970 Sudo Y, Yoshizawa S (2016) Functional and Photochemical Characterization of a Light-
971 Driven Proton Pump from the Gammaproteobacterium Pantoea vagans. *Photochem*
972 *Photobiol* **92**, 420-427.
973 Suzuki K, Del Carmen Marín M, Konno M*, et al.* (2022) Structural characterization of proton-
974 pumping rhodopsin lacking a cytoplasmic proton donor residue by X-ray
975 crystallography. *J Biol Chem* **298**, 101722.
976 Thaben PF, Westermark PO (2014) Detecting Rhythms in Time Series with RAIN. *Journal of*
977 *Biological Rhythms* **29**, 391-400.
978 Tran P, Ramachandran A, Khawasik O*, et al.* (2018) Microbial life under ice: Metagenome
979 diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes.
980 *Environ Microbiol* **20**, 2568-2584.
981 Tran PQ, Bachand SC, McIntyre PB*, et al.* (2021) Depth-discrete metagenomics reveals the
982 roles of microbes in biogeochemical cycling in the tropical freshwater Lake
983 Tanganyika. *ISME J* **15**, 1971-1986.
984 Vila X, Cristina XP, Abella CA, Hurley JP (1998) Effects of gilvin on the composition and
985 dynamics of metalimnetic communities of phototrophic bacteria in freshwater North-
986 American lakes. *J Appl Microbiol* **85 Suppl 1**, 138S-150S.
987 Vollmers J, Voget S, Dietrich S*, et al.* (2013) Poles apart: Arctic and Antarctic Octadecabacter
988 strains share high genome plasticity and a new type of xanthorhodopsin. *PLoS One* **8**,
989 e63422.
990 Wang Z, O'Shaughnessy TJ, Soto CM*, et al.* (2012) Function and Regulation of Vibrio
991 campbellii Proteorhodopsin: Acquired Phototrophy in a Classical Organoheterotroph.
992 *PLoS One* **7**.
993 Wurzbacher C, Salka I, Grossart HP (2012) Environmental actinorhodopsin expression
994 revealed by a new in situ filtration and fixation sampler. *Environ Microbiol Rep* **4**, 491-
995 497.
996 Yoshizawa S, Kumagai Y, Kim H*, et al.* (2014) Functional characterization of flavobacteria
997 rhodopsins reveals a unique class of light-driven chloride pump in bacteria. *Proc Natl*
998 *Acad Sci U S A* **111**, 6732-6737.

999 Yutin N, Koonin EV (2012) Proteorhodopsin genes in giant viruses. *Biology Direct* **7**, 34-34.
1000 Zabelskii D, Alekseev A, Kovalev K, *et al.* (2020) Viral rhodopsins 1 are an unique family of
1001 light-gated cation channels. *Nature Communications* **11**.
1002 Zehnpfennig JR, Hansel CM, Wankel SD, *et al.* (2022) Diel Patterns in Marine Microbial
1003 Metatranscriptomes Reflect Differences in Community Metabolic Activity Over Depth
1004 on the Continental Shelf of the North Atlantic. *Frontiers in Marine Science* **9**.
1005
1006

1007 **Data Accessibility and Benefit-Sharing**

1008 All metagenome data are publicly available at the DOE Joint Genome Institute's Integrated
1009 Microbial Genome (IMG) database or European Nucleotide Archive, with their Accession No.
1010 listed in Tables S2, S3, and S4. All metatranscriptome data are from the study by Linz *et al.* (2020).
1011 Amino acid sequences of opsin genes used in this study are listed in Supplementary Files 1, 2, 3,
1012 and 4.

1013

1014 **Author Contributions**

1015 K.D.M and S.H. designed and initiated the study; A.M.L. generated the metatranscriptome and
1016 part of the metagenome data; S.S. created the metagenome bins; P.T. generated part of the
1017 metagenome data; F.M.-F., B.O.O., J.R.D.-S., K.T.F., K.D.M., S.H., A.M.L. and S.S. participated
1018 in study discussion and metatranscriptome sampling; S.H. performed data analysis and interpreted
1019 the results; S.H. and K.D.M. wrote the manuscript. All authors contributed to the discussion and
1020 improvement of the manuscript.

1021

1022 **Table 1 and Figure Legends**

1023

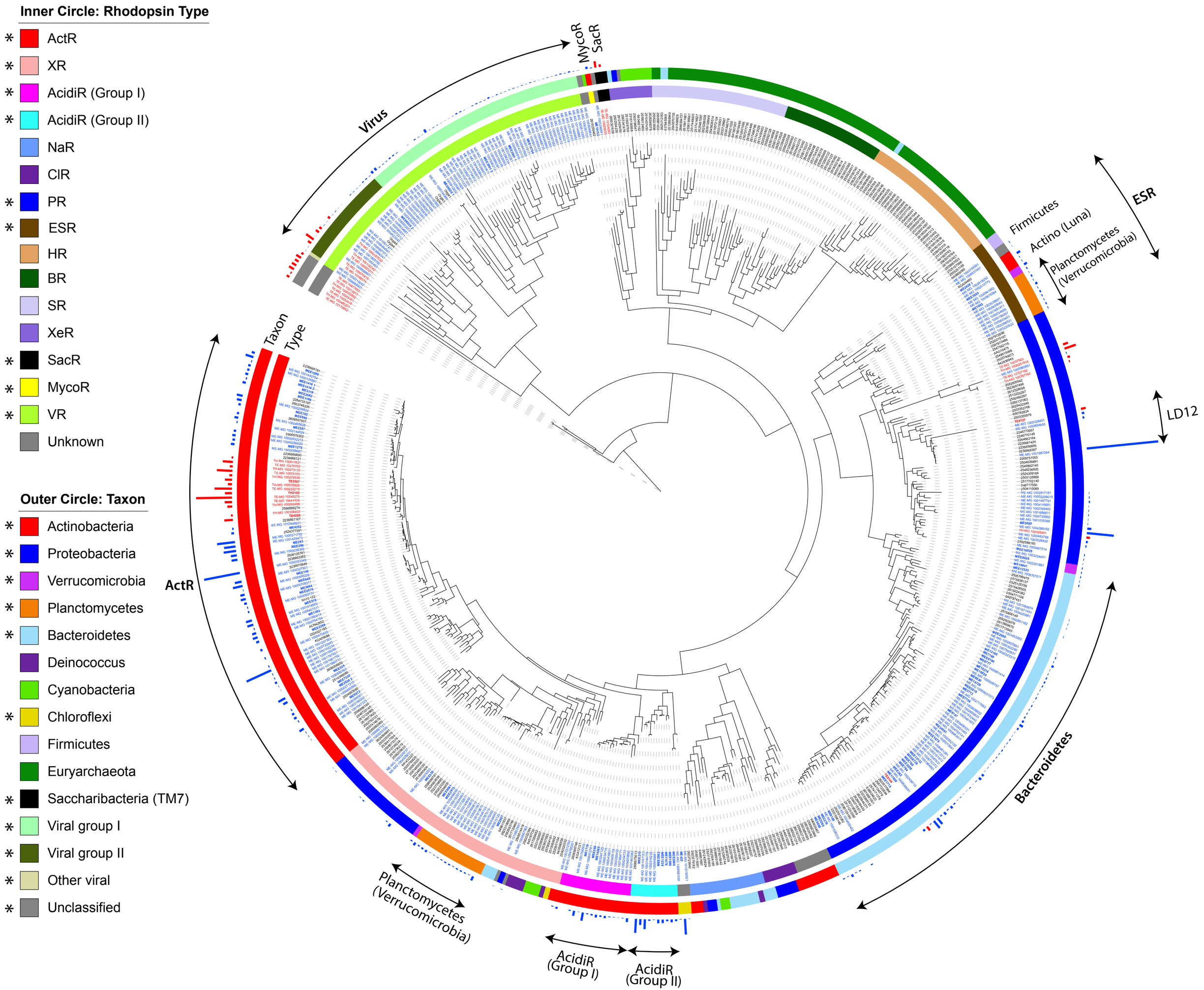
1024 **Table 1.** Mendota and Trout Bog MAGs containing opsin genes and the occurrence of key genes
1025 in their retinal biosynthesis pathway

IMG genome ID	Genome name	Phylum	Class/Lineage	Phytoene synthesis	Phytoene desaturation	Lycopene cyclization	β-carotene cleavage
2582580528	ME1363	Actinobacteria	acI-A	+	+	+	-
2582580550	ME2297	Actinobacteria	acI-A	+	+	+	+
2582580564	ME3119	Actinobacteria	acI-A	+	+	+	+
2582580555	ME283	Actinobacteria	acI-B1	+	+	+	-
2582580575	ME4252	Actinobacteria	acI-B4	+	+	+	+
2582580630	TE3207	Actinobacteria	acI-B	+	+	+	+
2582580635	TE4208	Actinobacteria	acI-B	+	+	+	+
2582580657	TH2152	Actinobacteria	acI-B	+	+	+	+
2582580572	ME3864	Actinobacteria	acI-C	-	+	-	-
2582580553	ME2538	Actinobacteria	Luna1-A	-	+	-	-
2582580513	ME10152	Actinobacteria	Myco	+	+	+	-
2582580520	ME11576	Actinobacteria	acIV	-	-	+	-
2582580530	ME14260	Actinobacteria	acIV	-	-	-	-
2582580539	ME1719	Actinobacteria	Iluma-B1	-	-	-	-
2582580603	ME7484	Bacteroidetes	Cytophagia	+	+	+	+
2582580526	ME13165	Bacteroidetes	Flavobacteriia	+	+	+	+
2582580549	ME2278	Bacteroidetes	Flavobacteriia	+	+	-	+
2582580568	ME34551	Bacteroidetes	Flavobacteriia	+	+	+	+
2582580592	ME6180	Bacteroidetes	Flavobacteriia	+	+	+	+
2582580602	ME7257	Bacteroidetes	Flavobacteriia	+	+	+	-
2582580532	ME14726	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580533	ME15216	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580535	ME15824	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580541	ME17509	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580563	ME3085	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580571	ME3713	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580585	ME5493	Bacteroidetes	Sphigobacteriia	+	+	+	-
2582580587	ME5895	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580596	ME6777	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580601	ME7198	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580610	ME9147	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580614	ME9830	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580640	TE5144	Bacteroidetes	Sphigobacteriia	+	+	+	+
2582580669	TH3063	Bacteroidetes	Sphigobacteriia	-	+	+	+
2582580613	ME9495	Bacteroidetes	Unclassified	+	+	+	-
2582580574	ME422	Chloroflexi	Unclassified	+	+	-	-
2582580581	ME5338	Proteobacteria	Alphaproteobacteria	+	+	+	+
2582580650	TE9727	Proteobacteria	Alphaproteobacteria	-	-	-	-
2582580561	ME3080	Proteobacteria	Betaproteobacteria	+	+	+	+
2582580558	ME2907	Planctomycetes	Planctomycetia	+	+	+	+
2582580606	ME7803	Planctomycetes	Planctomycetia	+	+	-	+
2556921353[a]	MEE2912	Planctomycetes	Unclassified	+	+	+	+
2556921396[a]	MEE3553	Verrucomicrobia	Unclassified	+	+	-	+
2556921562[a]	MEE8025	Verrucomicrobia	Unclassified	+	+	-	+
2556921572[a]	MEE8561	Verrucomicrobia	Unclassified	+	+	-	-
2582580545	ME19941	Unclassified	Unclassified	+	+	+	+
2556921260[a]	MEE15704	Virus		-	-	-	-
2556921414[a]	MEE3901	Virus		-	-	-	-
2556921518[a]	MEE6525	Virus		-	-	-	-

1026 [a]Although we mostly restricted this analysis to high quality MAGs, a few bins were also reported here, because these
1027 taxa were not represented in the high quality MAGs.

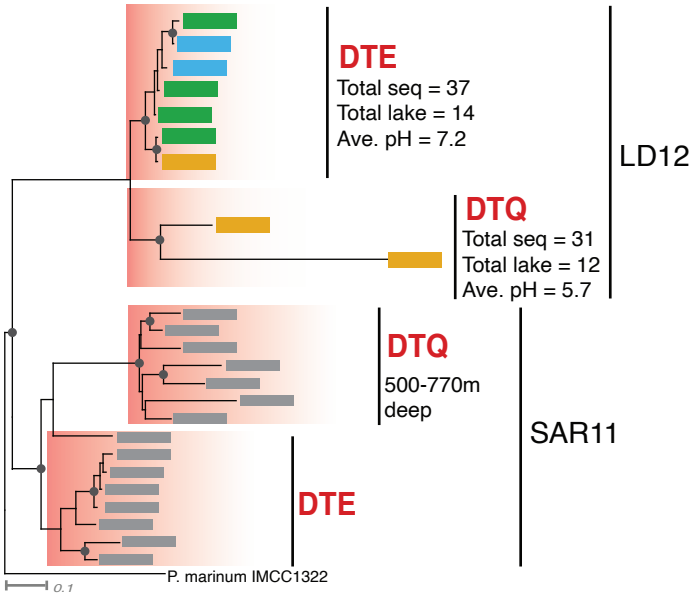
1028
1029 **Figure Legends**
1030
1031
1032 **Figure 1.** Phylogenetic tree of opsin protein sequences. This tree was viewed and annotated using
1033 the Interactive Tree of Life tool (http://itol.embl.de). Opsin gene names are colored as following:
1034 blue for ME, red for TE and TH metagenomes, and black for references from the database. For
1035 sequences in MAGs/bins (highlighted with bold fonts), gene names start with "ME", "MEE", "TE",
1036 and "TH" respectively. For the rest (un-binned) metagenome sequences, gene names start with
1037 "ME-MG", "TE-MG", and "TH-MG", respectively. The inner circle color strip indicates
1038 Rhodopsin types, and the outer circle color strip indicates taxon. Rhodopsin types and taxon
1039 present in our metagenomes were indicated with "*" in the legend. The outmost bar indicates
1040 normalized coverage-weighted abundance of opsin genes in the metagenome (blue for ME, and
1041 red for TE and TH). Some branches were labeled to highlight points discussed in the manuscript.
1042 Rhodopsin types include ActR (actinorhodopsin), XR (xanthorhodopsin), AcidiR (acidirhodopsin),
1043 NaR (sodium-pumping rhodopsins), CIR (chloride-pumping rhodopsins), PR (proteorhodopsin),
1044 ESR (*Exiguobacterium sibiricum* rhodopsin), HR (halorhodopsin), BR (bacteriorhodopsin), SR
1045 (sensory rhodopsin), XeR (xenorhodopsin), SacR (Saccharibacteria rhodopsin), MycoR (DTG
1046 rhodopsins represented by the freshwater Myco tribe), and VR (viral rhodopsin).
1047
1048 **Figure 2.** Variants of the cytoplasmic ion-transferring motif. (**a**) Distribution of major variants of
1049 the ion-transferring motif, listing together with the spectral-tuning residue, the number of opsin
1050 sequences with that motif from Mendota, Trout Bog and the stratfreshDB MAGs, the occurrence
1051 of these MAGs in clear or humic waters, and the taxa that contributed to these variants. The asterisk
1052 symbol ahead of the taxa name indicates that the DTE motif is also present in members of that taxa
1053 (from different MAGs). (**b**) Phylogenetic tree of opsins from LD12 and SAR11, with *Ca.*
1054 Puniceispirillum marinum (SAR116) as the outgroup. (**c**) Phylogenetic tree of opsins from the
1055 MycoR and SacR groups (DTG/DTS cluster), with a xenorhodopsin from *Nostoc* sp. as the
1056 outgroup. In both (**b**) and (**c**), sequences from Mendota, Trout Bog and the stratfreshDB MAGs
1057 were clustered based on 99% amino acid identity, and representative sequences were used to
1058 construct maximum likelihood phylogenetic trees using PhyML 3.0, with the LG substitution
1059 model and the gamma distribution parameter estimated by PhyML. Nodes labeled with a solid
1060 circle have bootstrap support values higher than 50%. For freshwater clades, the total number of
1061 sequences, the total number of their occurring lakes and the average pH of their occurring
1062 environments were listed for each motif variant-defined group. For the DTG group, as most (121
1063 out of 130) sequences were from four clearwater lakes, these values listed for clear and humic
1064 waters respectively.
1065
1066
1067 **Figure 3.** The significance of opsin genes in ME, TE and TH. (**a**) Percentage of opsin gene
1068 containing MAGs among the total of 103, 36, and 63 bacterial MAGs recovered in ME, TE and
1069 TH metagenomes, respectively. (**b**) Average of opsin gene relative abundance in metagenomes
1070 estimated by opsin gene coverage normalized by the coverage of single-copy genes. Error bars
1071 indicate standard deviations derived from the 95, 29 and 29 individual metagenomes generated
1072 from ME, TE and TH (listed in Table S3).
1073

1074 F**igure 4.** Opsin gene relative abundance in the surface layer of the 33 lakes versus the lake DOC
1075 levels. The 33 lakes, DOC levels, Accession No. of metagenomes included in this analysis and
1076 corresponding references for published studies were listed in Table S2. For lakes with multiple
1077 metagenomes (derived from different dates, depths, or studies), opsin gene abundance was
1078 estimated from individual metagenomes from the surface layer and averaged for that lake. Among
1079 the 33 lakes, four small boreal lakes from Buck *et al*. (2021) only had total organic carbon (TOC)
1080 measurements. As the difference between DOC and TOC was generally small for lakes in that
1081 study, we use TOC in place of DOC for those four lakes on this plot.
1082
1083 **Figure 5.** Depth profiles of opsin gene relative abundance. Lakes on the top panels are clearwater
1084 lakes, and on the bottom panels are humic lakes. For Lake Tanganyika, we only plotted abundances
1085 from the top 250 meters. Opsin gene abundance for samples at the depths of 300 m and 1200 m
1086 was 0.28 and 0.30 respectively. Metagenomes included in this analysis and corresponding
1087 reference papers were listed in Table S4.
1088
1089 **Figure 6.** Total opsin transcripts per liter of lake water averaged from all time points included in
1090 this analysis. The values from Lake Mendota and Sparkling Lake are about an order of magnitude
1091 higher than that in Trout Bog.
1092
1093 **Figure 7.** Diel cyclic trends of opsin gene expression. Cyclic genes were determined based on
1094 expression data collected from two consecutive days using the RAIN test with the *p*-value of <
1095 0.05. For both lakes, the top panel shows all cyclic genes, and the bottom three panels show cyclic
1096 genes in Actinobacteria, Bacteroidetes and Proteobacteria respectively. Read counts in transcripts
1097 per liter were z-score transformed and are shown on y-axis. Genes were colored by the peak hour
1098 of their expression as determined in the RAIN test, with warm and cold colors for peaks at "day"
1099 and "night" (as determined by photosynthetically active radiation), respectively.
1100
1101 **Figure 8.** Difference on the significance of rhodopsin-based phototrophy in carbon metabolism
1102 and energy flux between clearwater lakes (such as Lake Mendota) and humic lakes (such as Trout
1103 Bog). Prevalent of reactions is qualitatively indicated by the thickness of lines.
1104
1105

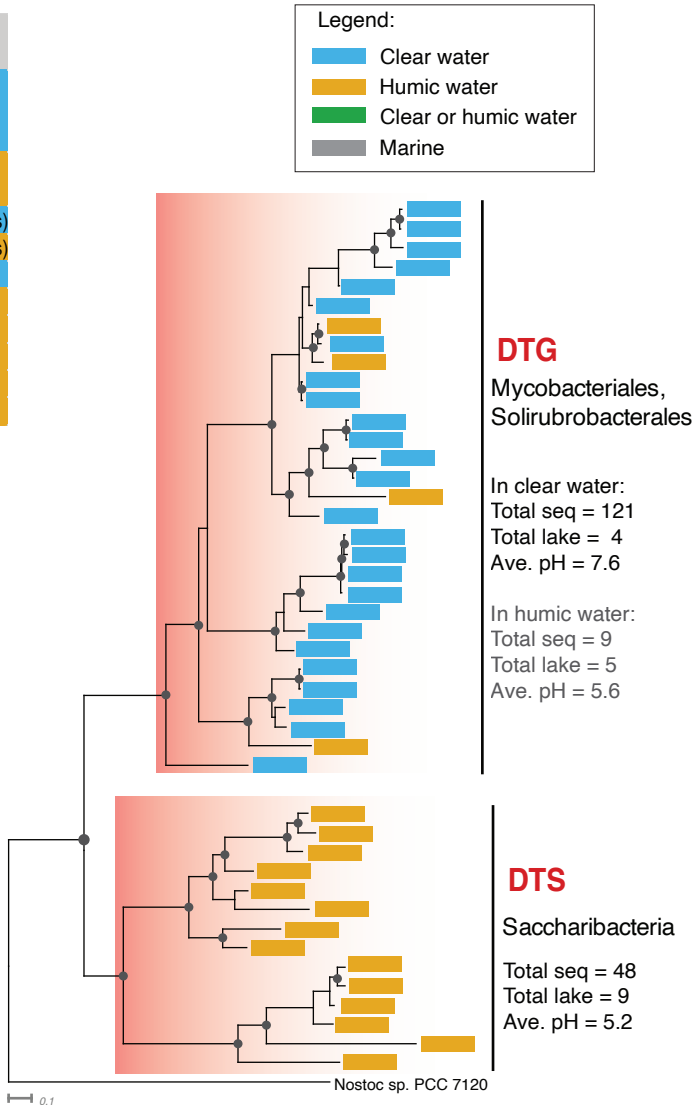

Inner Circle: Rhodopsin Type
ActR
XR
AcidiR (Group I)
AcidiR (Group II)
NaR
ClR
PR
ESR
HR
BR
SR
XeR
SacR
MycoR
VR
Unknown
Outer Circle: Taxon
Actinobacteria
Proteobacteria
Verrucomicrobia
Planctomycetes
Bacteroidetes
Deinococcus
Cyanobacteria
Chloroflexi
Firmicutes
Euryarchaeota
Saccharibacteria (TM7)
Viral group I
Viral group II
Other viral
Unclassified
Virus
MycoR
SacR
Taxon
Type
ActR
Firmicutes
Actino (Luna)
Planctomycetes
(Verrucomicrobia)
ESR
LD12
Bacteroidetes
Planctomycetes
(Verrucomicrobia)
AcidiR
(Group I)
AcidiR
(Group II)

Ion-transfer	Spectral tuning	No. of MAGs	Freshwater type	Taxa
DTE	L/M	841	Clearwater	Actinobacteria, Bacteroidetes, Proteobacteria, Chloroflexi, Planctomycetes, Verrucomicrobia, Acidobacteria, Bdellovibrionota
DTE	L/M	798	Humic	Actinobacteria, Bacteroidetes, Proteobacteria, Planctomycetes, Verrucomicrobia
DTG	L	121	Clearwater	Actinobacteria (Mycobacteriales, Solirubrobacterales)
DTG	L	9	Humic	Actinobacteria (Mycobacteriales, Solirubrobacterales)
DTK	L/M	17	Clearwater	Planctomycetes
DTS	L	46	Humic	Saccharibacteria
DTQ	L	30	Humic	*Proteobacteria (LD12)
ATV	Q	24	Humic	*Proteobacteria (Rhodoferax)
ATT	Q	15	Humic	*Proteobacteria (Rhodoferax)
EGV	Q	13	Humic	*Actinobacteria (acl-B)

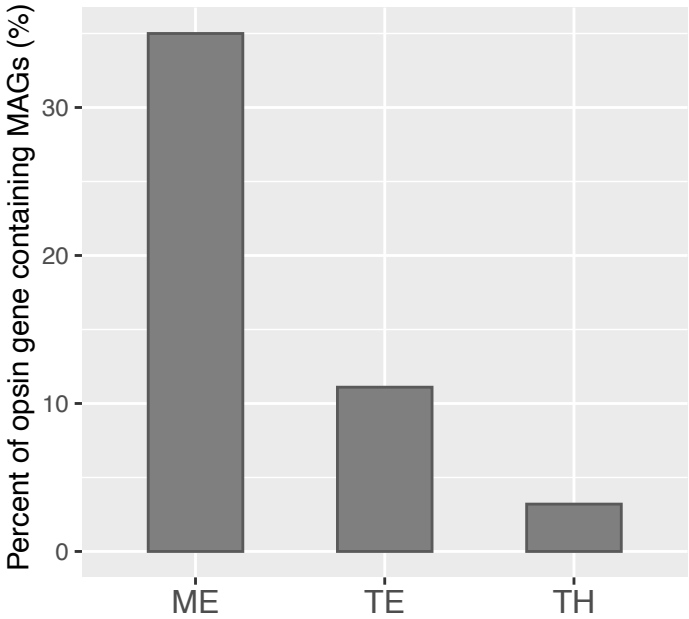
(a)



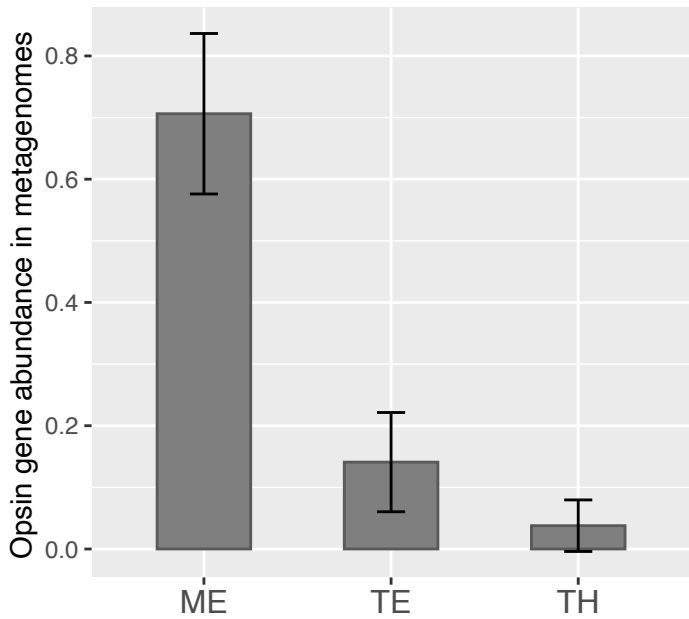
(c)



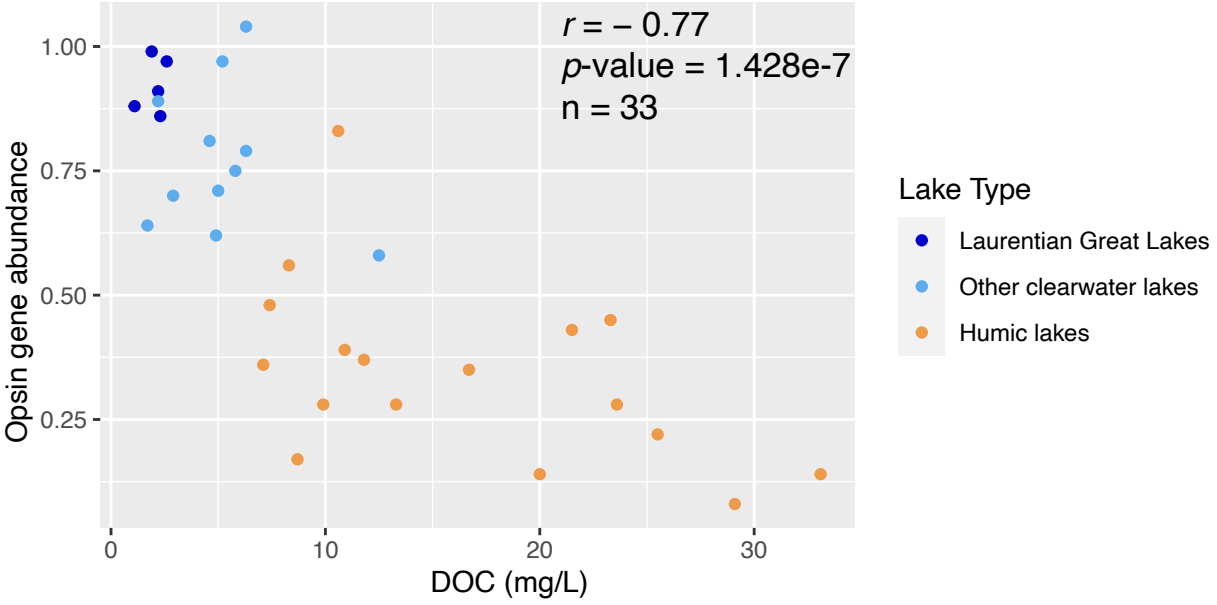
(b)

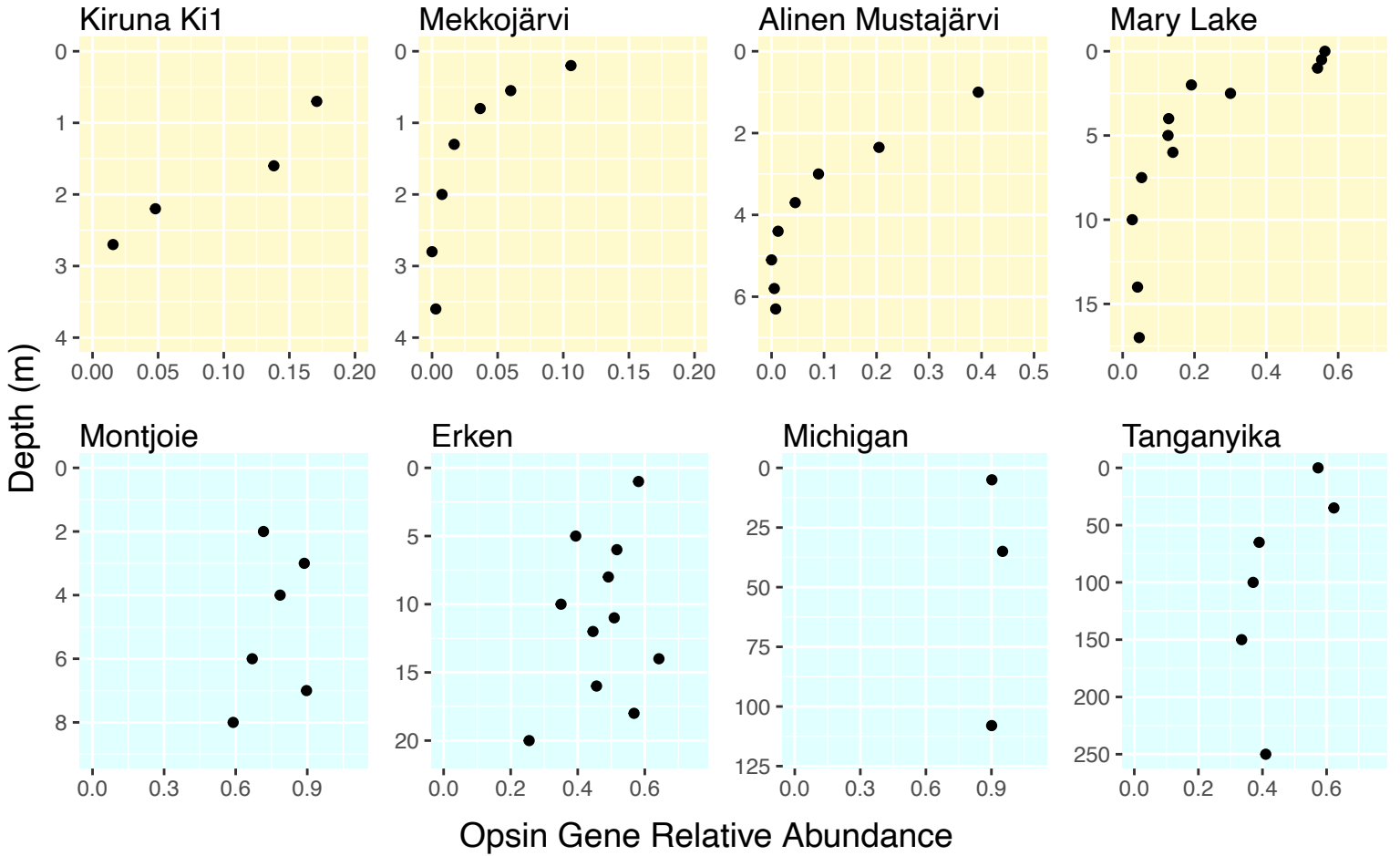

Percent of opsin gene containing MAGs (%)
30
20
10
0
ME
TE
TH

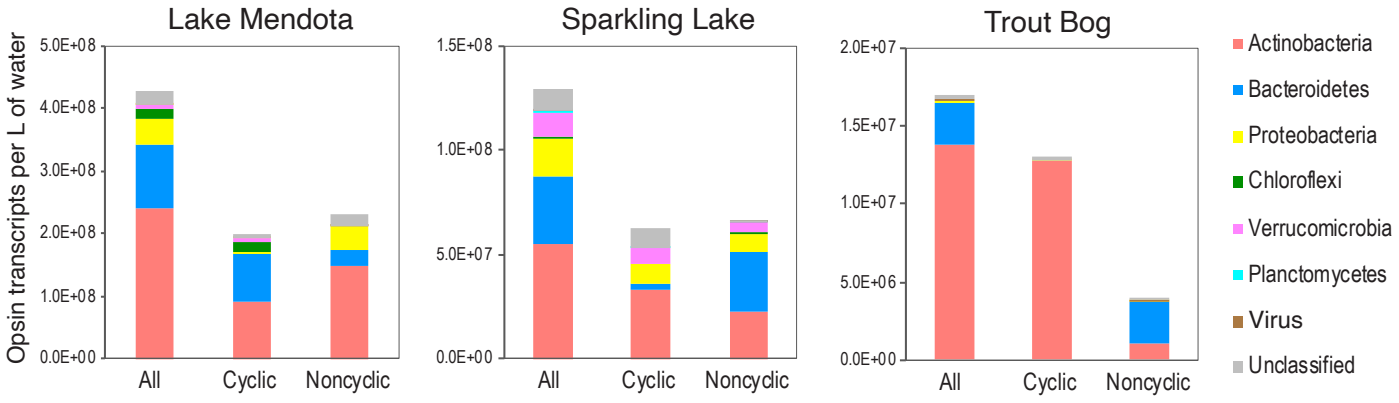
(a)

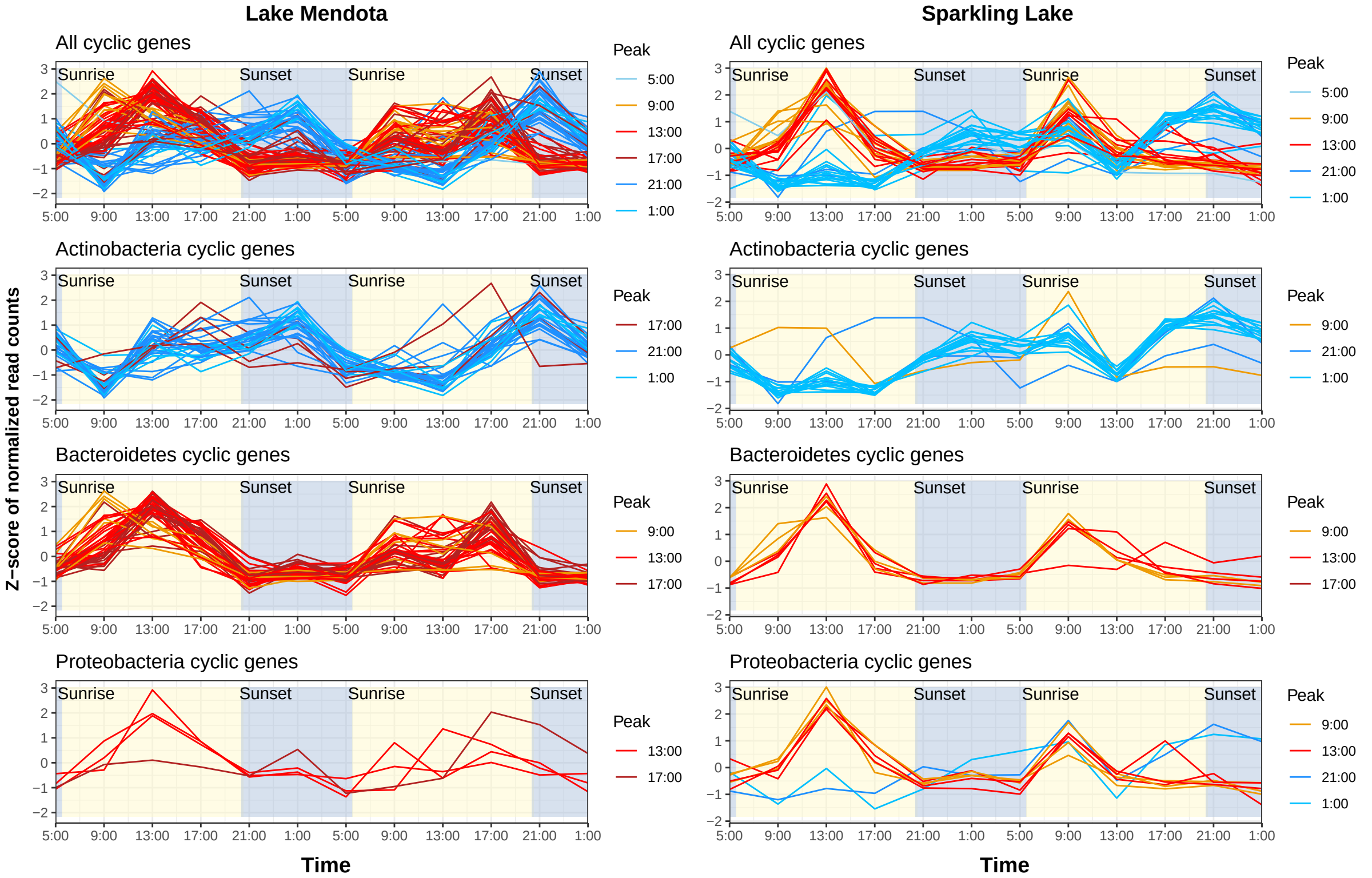

Opsin gene abundance in metagenomes
0.8
0.6
0.4
0.2
0.0
ME
TE
TH

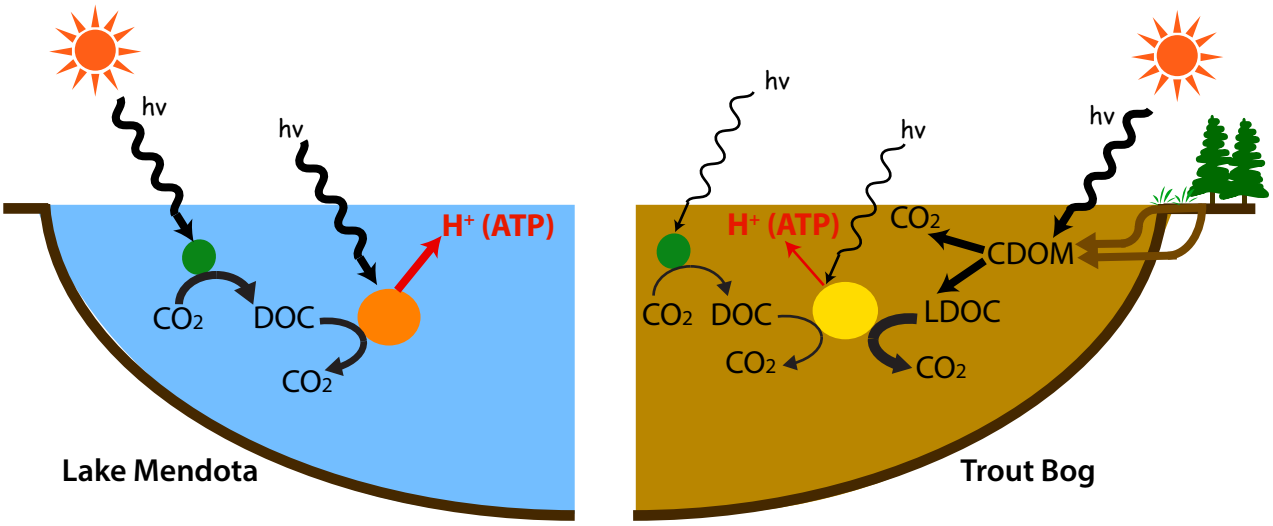
(b)


r = – 0.77
p-value = 1.428e-7
n = 33
Opsin gene abundance
1.00
0.75
0.50
0.25
0
10
20
30
DOC (mg/L)
Lake Type
Laurentian Great Lakes
Other clearwater lakes
Humic lakes


Kiruna Ki1
Mekkojärvi
Alinen Mustajärvi
Mary Lake
Montjoie
Erken
Michigan
Tanganyika
Depth (m)
Opsin Gene Relative Abundance


Lake Mendota
Sparkling Lake
Trout Bog
Opsin transcripts per L of water
5.0E+08
4.0E+08
3.0E+08
2.0E+08
1.0E+08
0.0E+00
1.5E+08
1.0E+08
5.0E+07
0.0E+00
2.0E+07
1.5E+07
1.0E+07
5.0E+06
0.0E+00
All
Cyclic
Noncyclic
Actinobacteria
Bacteroidetes
Proteobacteria
Chloroflexi
Verrucomicrobia
Planctomycetes
Virus
Unclassified


Lake Mendota
Sparkling Lake
All cyclic genes
Actinobacteria cyclic genes
Bacteroidetes cyclic genes
Proteobacteria cyclic genes
Sunrise
Sunset
Peak
5:00
9:00
13:00
17:00
21:00
1:00
Z−score of normalized read counts
Time



Clearwater lake with autochthonous DOC

Humic lake with allochthonous DOC

hv ⟿ Sunlight radiation

CDOM: chromophoric dissolved organic matter; **LDOC**: Labile dissolved organic carbon

Collection of photoautotrophic bacteria

Collection of heterotrophic bacteria with a high propotion of cells containing proton-pumping rhodopsins

Collection of heterotrophic bacteria with a low propotion of cells containing proton-pumping rhodopsins