Large lakes harbor streamlined free-living nitrifiers

Justin C. Podowski[a], Sara F. Paver[a], Ryan J. Newton[b], Maureen L. Coleman[a#]

[a] *Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois, USA*

[b] *School of Freshwater Sciences, University of Wisconsin Milwaukee, Milwaukee, Wisconsin, USA*

Corresponding author:
Maureen Coleman
University of Chicago
Department of the Geophysical Sciences
Hinds Laboratory
5734 S. Ellis Ave.
Chicago, IL 60637

Telephone: (773) 702-8352
Fax: (773) 702-0207
E-mail: mlcoleman@uchicago.edu

Running title: Nitrifiers across the Laurentian Great Lakes

Abstract

Microbial nitrification is a critical process governing nitrogen availability in aquatic systems and provides energy for aphotic carbon fixation. Nitrifiers are phylogenetically and physiologically diverse, but the factors controlling this diversity, and its biogeochemical implications, are poorly understood, particularly in freshwater systems. Here, we characterized the taxonomic, genomic, and metabolic diversity of free-living nitrifiers across Earth's largest freshwater system, the Laurentian Great Lakes. Surprisingly, ammonia oxidizing Bacteria (AOB) related to *Nitrosospira* dominated over ammonia oxidizing Archaea (AOA) at most stations, even in oligotrophic lakes with extremely low substrate concentrations. Nitrite oxidizing Bacteria (NOB) *Ca.* Nitrotoga and *Nitrospira* each dominated a subset of stations, even switching within a lake. Among four ecotypes of *Nitrosospira* we identified, one encodes proteorhodopsin, which could enhance survival of this AOB in transparent, low nutrient lakes. Genome reconstructions for free-living AOA, AOB, and NOB indicate small genomes across all taxa and diverse adaptations to sunlight and oxidative stress. Our findings expand the known functional diversity of nitrifiers and provide the first integrated picture of freshwater nitrifier genomics and ecology. As lakes worldwide undergo rapid environmental change, understanding the relationship between microbial biodiversity and nitrogen cycling is essential for predicting future ecosystem health and services.

Introduction

The oxidation of ammonia to nitrate powers the growth of nitrifying microorganisms and represents a critical flux in the global nitrogen cycle. In aquatic systems, microbial nitrification of ammonia released from sinking organic matter produces nitrate, which can then be removed from the system by denitrification (1). As chemolithoautotrophs, nitrifiers are also a major source of dark carbon fixation (2), which may contribute significant organic carbon to the microbial food web of the ocean's interior (3–5) and of deep freshwater lakes (6).

Microbial nitrifiers encompass wide diversity in phylogeny, physiology, and ecology, and understanding the drivers and consequences of this diversity across ecosystems is a fundamental challenge. Ammonia-oxidizing Archaea (AOA) in the phylum *Thaumarchaeota* dominate the mesopelagic oceans (7,8), likely due to their high affinity for ammonia (9,10) and streamlined genomes (11). In freshwater systems, AOA have been found to dominate oligotrophic lakes, while ammonia-oxidizing Bacteria (AOB) affiliated with the *Nitrosomonadaceae* (Betaproteobacteria) dominate more eutrophic systems (12–17). Complicating this picture, however, there is considerable genomic and physiological variation within each group. Some lineages, such as the thaumarchaeal *Nitrosopelagicus* (11) and bacterial *Nitrosomonas* cluster 6a (18,19), seem to be more specialized for nutrient scarcity than other related lineages. Within the AOA, there are also distinct ecotypes that appear to segregate with depth in the water column, in both marine (8,20) and freshwater systems (12). In freshwaters especially – which are poorly characterized compared to the oceans – it remains difficult to predict which ammonia-oxidizing taxa are likely to dominate in a given system.

For aquatic nitrite oxidizers, which span the phyla *Nitrospira, Nitrospinae,* and *Proteobacteria,* niche differentiation is even less clear. The oceans are dominated by exclusively marine lineages (2,21,22), consistent with ancient salinity-driven divergence. Cultivated strains of NOB show variation in substrate affinity and physiology (21,23,24), but the phylogenetic conservation of these traits, and their influence on environmental distributions, are

69 poorly understood. Moreover, recent studies have discovered that NOB are capable of
70 alternative energy metabolisms (25,26) and can access nitrogen from cyanate and urea (27–
71 29), expanding their ecological potential. In freshwater systems, the NOB *Ca.* Nitrotoga
72 (Betaproteobacteria) was only recently discovered to be widespread (30), and the diversity of
73 this genus and factors favoring its success are unknown.

74 Here, we use the Laurentian Great Lakes as a model system to examine niche
75 partitioning among planktonic freshwater nitrifiers. The Great Lakes hold 20% of Earth's surface
76 freshwater and more than half of this volume receives little to no light (< 1% surface irradiance).
77 This system, while hydrologically connected, spans strong trophic and chemical gradients:
78 ultraoligotrophic Lake Superior supports rates of primary production and nitrification comparable
79 to the ocean gyres (31,32), while Lake Erie supports greater production (33) and more than 70-
80 fold higher nitrification rate (34). Between these extremes, Lake Ontario has low ambient
81 ammonium concentrations like Lake Superior (35) but nitrification rates up to four times higher
82 (36). While previous studies reported that AOA and AOB dominate Lakes Superior and Erie,
83 respectively (16,32), recent community profiling has revealed broader diversity in both
84 ammonia-oxidizing and nitrite-oxidizing lineages (37–39). We sought to link taxonomic,
85 genomic, and metabolic diversity of nitrifiers with the varied biogeochemistry of the Great Lakes,
86 using genome reconstructions and abundance profiling. Our results uncover novel lineages and
87 metabolic capabilities and provide the first large-scale assessment of freshwater nitrifier
88 genomics.

89 **Materials and Methods**

90 ***Sample collection***

91 Water samples were collected from the Laurentian Great Lakes aboard the R/V *Lake*
92 *Guardian*, during the biannual Water Quality Surveys conducted by the U.S. EPA Great Lakes
93 National Program Office (40). Data presented here were collected in April and August 2012.

94 Samples were collected using a CTD rosette sampler (Sea-Bird Scientific) at the surface (2 m),
95 deep chlorophyll maximum (if present), the mid-hypolimnion, and near the bottom of the water
96 column (10 m above the lake bottom at most stations, 1 m above bottom at shallow stations).
97 For each sample, 5-8L of water was pre-filtered through a GF/A glass fiber filter (Whatman
98 1820-047; nominal pore size 1.6 µm) to exclude eukaryotic phytoplankton and particle
99 associated microbes, and cells were collected on 0.22 µm Sterivex filters (Millipore
100 SVGP01050). Filters were stored at -80°C. For dissolved nutrient analysis, 0.22 µm filtrate was
101 collected in 125 ml acid-clean HDPE bottles (Nalgene) and stored at -20°C. Samples for single
102 cell amplified genomes (SAG) were collected in August 2014. For each sample, 1 ml of raw
103 water was incubated with 100 µl of glycerol-TE buffer (20 ml 100X TE pH 8 + 100 ml glycerol +
104 60 ml water; final concentration after sample addition is 10mM Tris, 1mM EDTA, 5% glycerol)
105 for 10 minutes in the dark, then flash frozen in liquid nitrogen and stored at -80°C until
106 processing.

107 ***Physicochemical data***

108 CTD profiles, water chemistry and chlorophyll-*a* data were collected by the U.S. EPA
109 according to standard protocols (41) and retrieved from the Great Lakes Environmental
110 Database (https://cdx.epa.gov/) for 2012 and 2013. In addition, we measured dissolved nitrogen
111 species from August 2013 samples. Ammonium concentrations were measured using the OPA
112 method in a 96-well plate (42). Nitrate and nitrite concentrations were measured using the
113 Greiss reaction method in a 96-well plate (43).

114 ***16S rRNA analysis***

115 The full 16S rRNA amplicon dataset was described by Paver and colleagues (37). Here
116 we focus on data from the V4-V5 region (primers 515F-Y, 926R; ref. 44), collected in 2012 in
117 tandem with metagenome samples from select stations. We classified sequences using the
118 Silva v. 132 database (45) and the wang method (46) as implemented by mothur (47).

119 Sequences classified to each detected family of nitrifiers (ammonia oxidizer families
120 *Nitrosomonadaceae* and *Nitrosopumilaceae*; nitrite oxidizer families *Gallionellaceae* and
121 *Nitrospiraceae*) with a mothur-assigned confidence score above 90 were delineated into
122 taxonomic units using minimum entropy decomposition with a minimum substantive abundance
123 of 10 (48).

124 ***Metagenome and single-cell genome analysis***

125 One station per lake in Superior, Michigan, Huron, and Ontario, and two stations in Erie,
126 were selected for metagenome sequencing. DNA was extracted using a modified
127 phenol:chloroform extraction protocol (37) and libraries prepared according to the Illumina
128 TruSeq protocol. Samples from spring 2012 were sequenced at the Joint Genome Institute
129 using Illumina HiSeq (2x150bp). Samples from summer 2012 were sequenced at the University
130 of Chicago Functional Genomics Core Facility using Illumina HiSeq 2500 (2x250bp). Separate
131 assemblies of quality filtered reads were carried out for each metagenome using metaSPAdes
132 3.1.11 (49). Detailed QC, assembly, contig binning, and annotation are described in
133 Supplementary Methods.

134 To confirm the presence of proteorhodopsin, we analyzed a single cell amplified genome
135 from *Nitrosospira* collected from Lake Michigan and sequenced by the Joint Genome Institute.
136 Quality filtered reads were downloaded from JGI IMG/ER and normalized using bbnorm.sh with
137 target=100 and mindepth=2. Normalized reads were assembled using SPAdes in single cell
138 mode (50) with flags --sc and --careful. Resulting scaffolds were annotated identically to MAGs
139 as described in Supplementary Methods.

140 Phylogenomic trees and individual gene trees for proteorhodopsin, cyanase, and
141 photolyase were constructed as described in Supplementary Methods.

142

143

Results

Niche partitioning of nitrifiers across the Great Lakes

To map free-living (here defined as cells that pass through a 1.6μm filter) nitrifiers across the Great Lakes, we searched our recent 16S rRNA datasets for known nitrifying taxa (37). We detected putative AOB in the genus *Nitrosospira* (Betaproteobacteria) and AOA similar to *Nitrosoarchaeum*, along with putative NOB in the genera *Ca.* Nitrotoga (Betaproteobacteria) and *Nitrospira* (Nitrospirae). We did not detect 16S rRNA amplicons from *Nitrosococcus*, *Nitrococcus*, *Nitrospina* or *Nitrobacter*. The relative abundance of nitrifiers was negatively correlated with photosynthetically active radiation (PAR; Spearman's rho = -0.89, $p < 2.2e-16$) and reached a maximum below the depth of 1% PAR in each lake, up to 20% of amplicon sequences (Fig S1a). The relative abundances of ammonia- and nitrite-oxidizing taxa were strongly correlated (Spearman's rho = 0.918, $p < 2.2e-16$; Fig S1b). Free-living nitrifiers were rare (<0.1% relative abundance) in bottom water samples from the southern basin of Huron (HU15M) and western basin of Erie (HU91M); these two stations are the shallowest in our dataset and have relatively high light penetration to the bottom (~1% PAR). Chlorophyll-*a* concentration was also negatively correlated with the relative abundance of nitrifiers (Spearman's rho = -0.677, $p < 1.7e-7$; Fig S1c). These findings are consistent with previous work demonstrating photoinhibition of nitrification (51–54), as well as potential competition with phototrophs for ammonium (55).

The taxonomic assemblage of nitrifiers, along with nitrogen availability, differed across lakes and even across stations within a lake (Fig 1, Tables S1-S3). AOB (*Nitrosomonadaceae*) were observed across all lakes. In contrast, AOA (*Nitrosoarchaeum*) sequences only exceeded 0.5% relative abundance at the three deepest stations (SU08M, MI41M, ON55M) where the ratio of AOB:AOA ranged from 10:1 to 1:3 (Table S1). A single dominant oligotype accounted for more than 85% of *Nitrosoarchaeum* sequences in all samples, while four major oligotypes of

169 *Nitrosomonadaceae* were observed (data not shown). We found pronounced shifts in the
170 dominant NOB across stations (Fig 1), and all stations except those in Lake Ontario showed
171 strong dominance (greater than ten-fold) of either *Ca.* Nitrotoga or *Nitrospira*. *Nitrospira* was the
172 only nitrite oxidizer detected in Superior and the dominant nitrite oxidizer in parts of Michigan
173 (MI41M, MI18M). By contrast, *Ca.* Nitrotoga was the only nitrite oxidizer observed in Erie and
174 Huron and the dominant nitrite oxidizer at the shallowest station in Michigan (MI27M). In all but
175 one sample, a single dominant oligotype of *Ca*. Nitrotoga and *Nitrospira* accounted for >74%
176 and >84% of sequences from each taxon, respectively.

177 ***Ecotypic variation in abundant streamlined* Nitrosospira**

178 We reconstructed 15 genomes of the AOB *Nitrosospira*, substantially expanding genome
179 descriptions for this genus (56–58). Based on a phylogenomic tree, free-living Great Lakes
180 *Nitrosospira* fall into two major clades, both of which are distinct from published species; each of
181 these clades also includes MAGs recovered from lakes Biwa and Baikal, suggesting novel
182 globally distributed freshwater lineages (Fig S2). One clade, which we call NspGL1, has a
183 reduced genome (average 1.4 Mb) and low G+C content (41%) (Fig 2, Table S4), typical of
184 streamlined genomes (59). The second clade was resolved into three subclades (denoted
185 NspGL2a, 2b, and 3; Fig S2) based on phylogeny and average nucleotide identity (ANI), all with
186 small genome sizes of 1.5-1.8 Mb and 50% G+C content (Fig 2, Table S4). Based on short read
187 mapping, these subclades are ecologically distinct: NspGL1 and NspGL2b -- with the smallest
188 genomes -- are the dominant AOB in the upper oligotrophic lakes, while NspGL2a is only
189 abundant in Ontario and NspGL3 is only abundant in Erie (Fig 2). Hereafter we refer to these
190 subclades as ecotypes due to their phylogenetic and ecological separation.

191 We compared the genomes of our Great Lakes *Nitrosospira* to 87 reference AOB
192 genomes belonging to the family *Nitrosomonadaceae,* revealing distinctive patterns of gene
193 content. Great Lakes *Nitrosospira* genomes encode fewer two-component signal transduction

194 systems, transposases, motility genes, pilus and secretion genes, and defense-related genes,
195 and their complement of transporters is distinct from most reference genomes (Table S5). They
196 also lack functions related to biofilm formation such as polysaccharide matrix production (e.g.
197 *pel* genes) and extracellular protein targeting (exosortase and PEP-CTERM motifs). These
198 features are in sharp contrast to *Nitrosospira* isolates from soil (56,58) and to oligotrophic
199 *Nitrosomonas* isolates (60) and are consistent with a passive planktonic lifestyle in extremely
200 low-nutrient systems.

201 We next compared metabolic potential among Great Lakes AOB ecotypes to understand
202 their ecological preferences for upper lakes (NspGL1, NspGL2b), Ontario (NspGL2a), and Erie
203 (NspGL3). Surprisingly, all seven NspGL1 MAGs contain a cluster of genes encoding
204 proteorhodopsin and retinal biosynthesis, including 15,15'-Beta-carotene dioxygenase (*blh*),
205 lycopene cyclase (*crtY*), phytoene dehydrogenase (*crtI*), phytoene synthase (*crtB*), and GGPP
206 synthase (*crtE*) (Fig 3a). We also identified proteorhodopsin in a single cell amplified genome
207 (SAG) representing NspGL1 from Lake Michigan (Fig 3a), demonstrating that it is not an artifact
208 of metagenome assembly. To our knowledge, this is the first example of a nitrifier with
209 proteorhodopsin. All NspGL1 proteorhodopsins share residues H95, D127 and E138 along with
210 a short beta-turn (G111-P116) between helices B and C, which are characteristic features of
211 proteorhodopsin as distinct from sensory and other rhodopsins (61), and the presence of
212 leucine at position 135 suggests green light tuning (62) (Fig 3b). All of the genes in this module
213 have highest similarity to homologs from *Polynucleobacter*, but are flanked by
214 Nitrosomonadaceae-like genes, suggesting recent horizontal gene transfer (Fig 3a). The
215 predicted NspGL1 proteorhodopsins cluster with *Polynucleobacter*, *Methylopumilus*, and other
216 freshwater Betaproteobacteria in Supercluster III as defined by MicRhoDE (63) (Fig S3). We
217 compared the homologous genome region in two highly similar MAGs from Lakes Biwa and
218 Baikal (Fig S2); these contigs lack the proteorhodopsin module but appear to flank a variable
219 region where the contig assembly ends (Fig S4). A proteorhodopsin photosystem could support

220 survival of NspGL1 in the presence of sunlight, which has been shown to inhibit ammonia
221 oxidation (51,64). In the upper lakes where NspGL1 is abundant, light penetration is high well
222 below the thermocline in stratified periods (65), and deep water taxa are seasonally advected to
223 the surface by water column mixing (37). In addition to proteorhodopsin, NspGL1 – but not the
224 other three ecotypes of Great Lakes *Nitrosospira* – encode a class I cyclopyrimidine dimer
225 photolyase, which uses light energy to repair UV-induced DNA damage, and the catalase-
226 peroxidase *katG*, suggesting that the NspGL1 ecotype is adapted to relatively shallow depths in
227 the water column (Fig 2).

228 Great Lakes *Nitrosospira* carry a reduced, ecotype-specific complement of nitrogen
229 metabolism genes compared to reference AOB (Fig 2; Table S5). All are presumed to have the
230 core ammonia oxidation enzymes ammonia monooxygenase and hydroxylamine
231 dehydrogenase; these genes were assembled and binned as expected in some MAGs and
232 were manually identified on short unbinned contigs in other cases. Surprisingly, all Great Lakes
233 *Nitrosospira* lack the copper protein nitrosocyanin, whose precise function is unknown but so far
234 has been found in all AOB except one member of the *N. oligotropha* clade (60). The lack of
235 nitrosocyanin extends beyond the Great Lakes MAGs to closely related freshwater and marine
236 strains, along with five more members of the *N. oligotropha* clade (Fig S2); its absence may be
237 related to the divergence of these clades. Only NspGL1 and NspGL2b encode NO-forming
238 nitrite reductase (NirK), which confers nitrite tolerance (66); this result is surprising given that
239 these two clades dominate the upper lakes where productivity and reduced N are lowest. None
240 of the Great Lakes MAGs encode NO reductase (NorCBQD), and NspGL1 lacks cytochrome
241 P460 family proteins; these genomic patterns may influence N_2O emissions across the Great
242 Lakes. Nitrogen acquisition is also distinct among Great Lakes AOB: NspGL1 lacks an apparent
243 ammonium transporter, but encodes urease structural and accessory genes (*ureABCEFG*) and
244 a high-affinity urea transporter (*urtABCDE)*. Further, all Great Lakes ecotypes encode a high-
245 affinity amino acid transporter (*livFGHM*) and a secreted non-specific nuclease; these genes are

246 rare (<5%) in reference genomes and could supply reduced nitrogen and/or organic carbon.
247 Finally, NspGL1 and NspGL3 have genes for producing cyanophycin, an intracellular storage
248 compound for nitrogen (58,67). Together, the minimal and distinctive gene complements
249 present in Great Lakes *Nitrosospira* illustrate the variability and adaptability of AOB gene
250 content across habitats.

251 As with nitrogen metabolism, carbon metabolism is also distinct between Great Lakes
252 and reference AOB, and among Great Lakes ecotypes (Fig 2; Table S5). Unlike most reference
253 AOB, Great Lakes *Nitrosospira* lack two key enzymes of the oxidative pentose phosphate
254 pathway, glucose-6-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase. All
255 ecotypes except Erie-specific NspGL3 also lack genes for glycogen synthesis and degradation.
256 These findings suggest continuous autotrophy, rather than “feast or famine” dynamics of carbon
257 storage and utilization, or diel periodicity as in photoautotrophs. The key enzyme for carbon
258 fixation, RuBisCO, has evolved several kinetically distinct forms whose distribution likely reflects
259 ecological pressures (68). NspGL1 and NspGL3 both contain Form IA RuBisCO, while
260 NspGL2a and NspGL2b contain Form IC RuBisCO (68,69) (Fig 2). NspGL1 genomes also
261 possess an alpha carboxysome-like *cso* operon, similar to *Nitrosomonas eutropha* C91 (69),
262 though our draft assembly lacks the expected carbonic anhydrase (*csoS3/csoSCA*).
263 Carboxysome-associated RuBisCO may allow NspGL1, the ecotype most strongly adapted to
264 energy and nutrient limitation, to more efficiently fix CO_2 by minimizing the wasteful oxygenation
265 reaction and reducing the cellular nitrogen allocation to RuBisCO (68). The ranges of kinetic
266 properties observed in other autotrophs for Form IAq (found in NspGL3) and Form IC (found in
267 NspGL2a and NspGL2b) overlap, and therefore more work is needed to understand the fitness
268 advantages, if any, that this RuBisCO diversity confers on Great Lakes nitrifiers.

269 ***Streamlined freshwater Thaumarchaeota***

270 We reconstructed three similar genomes (>99% ANI) of *Nitrosoarchaeum* (NarchGL; Fig
271 S5, Fig S6, Table S4) from three separate samples (two from Superior, one from Ontario),
272 consistent with our low observed 16S rRNA diversity for *Thaumarchaeota*. These NarchGL
273 genomes are very similar (~99% ANI) to two genomes reconstructed from Lake Baikal, located
274 thousands of kilometers away (70). Their next closest relatives are also from freshwater
275 environments, and phylogenetic clustering suggests that salinity is an important driver of
276 divergence throughout the *Nitrosopumilaceae* (Fig S5). Using our reconstructed genomes as
277 probes for metagenomic read recruitment, NarchGL were detected in Superior, Michigan, and
278 Ontario; they represented roughly one-third of ammonia oxidizers in the mid-hypolimnion of
279 station SU08M, reaching relative abundance of approximately 0.5% (Fig 2).
280 NarchGL share nearly 90% of their predicted proteins with close relatives including *Ca.*
281 Nitrosoarchaeum limnia; at the same time, they show distinctive patterns in gene content that
282 pinpoint the key selective pressures of deep lakes (Fig 2, Table S6). All three NarchGL
283 genomes encode urease and a urea transporter, implicating urea as a vital source of nitrogen
284 for energy and/or biosynthesis. Consistent with phosphorus scarcity in much of the Great Lakes
285 (71), NarchGL encode high affinity transport systems for phosphate and potentially
286 phosphonates, though we did not identify a phosphonate lyase. In addition to both CRISPR/Cas
287 enzymes cas1 and cas4, NarchGL genomes contain several phage proteins, suggesting viral
288 infection and integration events may be common. DNA photolyases, which have been found in
289 epipelagic clades of marine *Thaumarchaeota* (72–74), are present in all low salinity
290 *Nitrosoarchaeum* including NarchGL, suggesting NarchGL are exposed to sunlight due to high
291 water clarity (65) and/or annual mixing in the Great Lakes. NarchGL also lack the common
292 tRNA modification 4-thiouridylation (indicated by IscS K04487 and ThiI-NFACT PF02568-
293 PF18297; ref. 75); we propose that the absence of this modification, which is susceptible to
294 near-UV radiation (75), is also related to sunlight exposure.

295 While aquatic AOA are typically adapted to oligotrophic conditions, NarchGL appear
296 especially streamlined. Their genomes reveal striking reduction in environmental sensing,
297 response, and regulatory functions, relative to other *Nitrosoarchaeum* and *Nitrosopumilaceae*
298 (Table S6). NarchGL encode fewer variants of the general archaeal transcription factors, TATA
299 binding protein (TBPs, PF00352) and transcription factor B (TFBs, PF00382 and PF08271),
300 combinations of which influence differential gene expression (76). NarchGL also encode fewer
301 two-component systems that transmit environmental signals to control gene expression or
302 protein activity (domains: cache PF02743, HAMP PF00672, His kinase A PF00512, receiver
303 PF00072). Further, they are depleted in ArsR family transcription factors (PF01022), P-II
304 proteins for regulation of nitrogen metabolism (PF00543), and other potential regulatory
305 domains (CBS PF00571, USP PF00582). This extremely limited regulatory capacity in NarchGL
306 stands in sharp contrast to closely related *Ca.* N. limnia, but instead parallels the oceanic
307 minimalist *Ca.* Nitrosopelagicus brevis (11) (Table S6).

308 ***Expanded diversity of Ca.* Nitrotoga *with reduced genomes***

309 Despite the broad distribution of *Ca.* Nitrotoga in freshwater systems and beyond, only
310 six genomes are available, derived from rivers heavily impacted by urban and agricultural
311 influence, a wastewater treatment plant, and coastal sediment (30,77,78). Hence the metabolic
312 and phylogenetic diversity of this group is virtually unexplored. We reconstructed six new MAGs
313 of *Ca*. Nitrotoga, which form two clusters with >99% ANI within each cluster and ~97% ANI
314 between clusters (NtogaGL1a and NtogaGL1b; Fig S2). These new *Ca*. Nitrotoga MAGs are far
315 smaller than published genomes (1.5 Mb compared to 2.6-3.0 Mb; Fig 4, Table S4), lacking
316 functions such as motility and chemotaxis, pilus biogenesis, and DNA repair (*mutLS*) (Table
317 S7). Great Lakes *Ca.* Nitrotoga also encode markedly fewer two-component systems for
318 sensing and responding to environmental cues, defense-related genes (restriction-modification,
319 toxin-antitoxin, and Crispr-Cas systems), transposases, and efflux transporters (Table S7).

320 While incomplete assembly of hypervariable genome regions may explain some of these
321 absences, the overall genome properties are consistent with a relatively stable low-nutrient
322 environment and planktonic lifestyle.

323 The reduced genomes of NtogaGL1a/b help clarify core features of the genus *Nitrotoga*,
324 along with accessory functions that may enable local adaptation in specific populations. To date,
325 sequenced *Ca.* Nitrotoga including NtogaGL1a/b encode similar electron transport pathways,
326 including NADH dehydrogenase-Complex I, succinate dehydrogenase-Complex II, and
327 Alternative Complex III, along with high-affinity cbb3-type cytochrome oxidases suggesting
328 adaptation to low oxygen conditions. They also share the Calvin cycle for carbon fixation, a
329 complete TCA cycle, and an evolutionarily distinct nitrite oxidoreductase (NXR) from other NOB
330 (30,77,78). All *Ca.* Nitrotoga to date also share transporters for amino acids and peptides,
331 potential sources of C and/or N. *Ca.* Nitrotoga can also potentially access reduced sulfur
332 compounds for energy via sulfite dehydrogenase, suggesting metabolic flexibility beyond nitrite
333 oxidation.

334 Beyond these similarities, the small genomes of NtogaGL1a/b are distinct from
335 previously described *Ca.* Nitrotoga in many ways. NtogaGL1a/b lack NiFe hydrogenase to use
336 hydrogen as an energy source. They also lack nitrogen metabolism functions including
337 assimilatory nitrite reductase (*nirBD*), nitrite reductase to NO (*nirK*), and nitric oxide reductase
338 (*norQD*). Based on gene content, NtogaGL1a/b appear unable to use hexoses like glucose,
339 since they lack the glycolytic enzyme phosphofructokinase and the Entner-Doudoroff pathway,
340 similar to *Nitrobacter winogradskyi* (79). Consistent with this, they also lack genes for storage
341 and breakdown of glycogen (Table S7). All but one of the NtogaGL1a/1b genomes encode
342 cyanate lyase (*cynS*), which is found in other NOB but no *Ca.* Nitrotoga to date (27,80,81). The
343 *cynS* gene, adjacent to *glnK-amtB* for ammonium sensing and transport, likely functions in N
344 assimilation, as recently described for *Nitrospinae* (22). While cyanase has been shown to
345 mediate reciprocal feeding between some NOB and ammonia oxidizers (27), this interaction is

346 unlikely in the free-living (<1.6um) size-fraction and dilute environment sampled here. Notably,
347 cyanase from NtogaGL1a/1b, along with predicted *Nitrospirae* proteins from Lake Baikal and
348 soil, form a distinct phylogenetic cluster from most nitrifier cyanase proteins observed to date
349 (Fig S7).

350 The two ANI-based clusters we detected, NtogaGL1a and NtogaGL1b, appear to be
351 phylogenetically and ecologically distinct ecotypes. Based on short-read mapping, NtogaGL1b
352 dominates Erie, while NtogaGL1a dominates all other *Ca.* Nitrotoga-containing samples (Fig 4).
353 We found several metabolic genes that differentiate the two ecotypes. ER-specific NtogaGL1b
354 genomes share a region encoding thiosulfate dehydrogenase (*tsdA*), cytochromes, transport of
355 sulfur-containing compounds, lactate dehydrogenase (ldh*),* a two-component system, and a
356 Crp-family transcription factor (Fig S8). This region may be involved in oxidizing thiosulfate as
357 an energy source, and sensing and responding to redox changes that accompany seasonal
358 hypoxia in Lake Erie. The corresponding region in NtogaGL1a encodes an integrase and
359 photolyase, consistent with greater DNA photodamage in the more transparent waters of
360 Michigan, Huron, and Ontario where NtogaGL1a is abundant. NtogaGL1a also lacks superoxide
361 oxidase (*cybB*) found in NtogaGL1b and reference strains, suggesting these two ecotypes
362 experience differential oxidative stress.

363 ***Great Lakes* Nitrospira *reveal adaptations to sunlit oxic environment***

364 We reconstructed six closely related genomes of *Nitrospira* (~99% ANI; Fig S9, Table
365 S4), representing the predominant NOB throughout Lake Superior and in parts of Michigan,
366 Ontario, and Huron (Fig. 4; Table S1). These genomes, which we refer to as NspiraGL, fall
367 within lineage II (Fig S9), which is broadly distributed across soil, freshwater, and engineered
368 habitats (21); however, genome analyses to date have focused on strains from wastewater and
369 engineered systems, leaving major blind spots. NspiraGL share core features of *Nitrospira*
370 metabolism, including a periplasmic-facing NXR that is advantageous under substrate-limiting

371 conditions, multiple cytochrome *bd*-like oxidases, and the reverse TCA cycle for carbon fixation
372 (80). However, as with *Ca.* Nitrotoga, the *Nitrospira* genomes we reconstructed in the Great
373 Lakes are markedly smaller (~1.8 Mb) than published reference genomes (Fig. 4; Table S4),
374 with reduced capacity for environmental sensing, transport, defense, and transposition, and a
375 lack of pilus or flagellar motility (Table S8). NspiraGL encode just five sigma factors, compared
376 to 18 in *Nitrospira moscoviensis,* consistent with genome streamlining theory (59). Further,
377 NspiraGL genomes encode a single NXR, while *N. moscoviensis* carries five copies that are
378 differentially regulated (29,82). NspiraGL also lack the *glnE* gene for glutamine synthetase (GS)
379 adenylyltransferase, suggesting that GS activity is not repressed by this mechanism. Together,
380 these features suggest limited regulatory and ecological flexibility, consistent with a relatively
381 constant, oligotrophic environment.
382 Compared to other *Nitrospira,* NspiraGL exhibit limited energetic flexibility, but can
383 access diverse nitrogen sources (Fig. 4, Table S8). We predict that NspiraGL are unable to
384 grow on hydrogen or formate as alternative energy sources (25,29), as they lack NiFe-
385 hydogenase and formate dehydrogenase. The glycolysis and oxidative TCA cycles appear to be
386 incomplete, lacking phosphofructokinase and citrate synthase, respectively; this suggests a
387 limited capacity for organic carbon utilization. Regarding nitrogen metabolism, NspiraGL lack
388 *nirK,* encoding NO-forming nitrite reductase, but maintain nitric oxide reductase (*norQD*),
389 implying that cellular NO is formed by other mechanisms. To obtain nitrogen for biosynthesis,
390 NspiraGL encode a high-affinity nitrate/nitrite/cyanate transporter (*nrtABC*), assimilatory nitrite
391 reductase (*nirA*), and cyanase (*cynS*), along with two *amt* family ammonium transporters.
392 Although none of the NspiraGL MAGs include urease (*ureCBA*), one does contain urease
393 accessory proteins (*ureEFGD*) and two contain a urea transporter (*urtABCD*), suggesting
394 incomplete assembly of the urea utilization pathway. As with *Ca.* Nitrotoga, we suggest that
395 cyanase, along with urease where present, functions in nitrogen assimilation rather than cross-
396 feeding, given the dilute environment and free-living planktonic cells.

397 Beyond energy, carbon, and nitrogen metabolism, we discovered striking differences
398 between NspiraGL and reference *Nitrospira* related to DNA repair. NspiraGL encode two
399 additional photolyase-related proteins, along with a class I cyclopyrimidine dimer (CPD)
400 photolyase found in most reference *Nitrospira* (Fig 5, Fig S10). Photolyases use blue light
401 energy to repair DNA lesions caused by UV radiation (83). The two additional genes in
402 NspiraGL are adjacent and share best hits with Betaproteobacteria, suggesting recent
403 horizontal transfer (Fig 5). One likely encodes an FeS-BCP photolyase, which repairs (6-4)
404 dipyrimidine lesions (84,85). The other shares an FAD-binding domain with photolyases but the
405 C-terminal region has no recognizable domains (Fig 5b). This protein is widespread in aquatic
406 bacteria and has not been functionally characterized, though an actinobacterial homolog was
407 suggested to be involved in light sensing and regulation (86). Beyond photolyases, NspiraGL
408 also encode uracil-DNA glycosylase (UNG), which removes misincorporated uracil from DNA.
409 Uracil results from deamination of cytosine, which can occur spontaneously or be induced by
410 NO (87). In addition to the photolyases and UNG that repair DNA lesions, NspiraGL encode
411 translesion DNA polymerase V (*umuCD*) which enables replication to proceed past lesions.
412 Together, these genes indicate that *Nitrospira* in the Great Lakes experience significant DNA
413 damage, including UV-induced damage that also requires light for the repair process, in
414 hypolimnion waters with high transparency (65) and/or during seasonal mixing.
415 Other major differences between NspiraGL and reference *Nitrospira* are related to
416 reactive oxygen species (ROS) (Table S8). Surprisingly, despite their oxic habitat, NspiraGL
417 lack superoxide dismutase (SOD), monofunctional catalase (*katE*), classic regulators of
418 oxidative stress response (*oxyR* and *perR*), and bacterioferritin, which limits the Fenton reaction
419 by sequestering free iron. However, all six NspiraGL MAGs, but few reference genomes (7% of
420 75), have recently acquired bifunctional catalase-peroxidase *katG*; interestingly we also
421 observed *katG* in Great Lakes *Ca.* Nitrotoga and *Nitrosospira* (Fig 2, Fig 4). The absence of
422 SOD suggests that NspiraGL does not produce damaging levels of endogenous superoxide,

423 perhaps because NspiraGL lack the major respiratory and non-respiratory flavoproteins that
424 produce ROS in other SOD-containing *Nitrospira* (88). Unlike superoxide, H_2O_2 can cross
425 membranes, and is known to be produced by both photooxidation of dissolved organic matter
426 and dark heterotrophic activity (89). The lakes where NspiraGL dominate have high water clarity
427 (65) and low productivity, consistent with abiotic photochemistry as the primary source of
428 exogenous ROS; this stress may have selected for *katG* as a defense, though without the
429 canonical *oxyR* regulation. The absence of bacterioferritin could be related to iron limitation (71).
430 NspiraGL also lack cytochrome *c* peroxidase, which is found in 70 of 75 reference genomes;
431 this protein is proposed to function in anaerobic respiration of H_2O_2 (90) and therefore its
432 absence in NspiraGL is consistent with a constant oxic environment. Together, these results
433 indicate that *Nitrospira* in the Great Lakes face distinct ROS pressures that have shaped their
434 gene content.
435
436 **Discussion**
437 The Laurentian Great Lakes harbor nitrifiers that are phylogenetically related, but
438 markedly different in genome size and functional capacity, from their well-studied relatives
439 inhabiting wastewater systems, soils, and even other freshwater systems. By examining the
440 entire nitrifier assemblage at once, we detected common features across taxa that illuminate the
441 selective pressures faced by microbes in deep lakes. All the lineages we describe show small
442 genome sizes (1.3-1.8 Mb), reduced capacity for environmental sensing and response, and
443 adaptation to a passive (i.e. non-motile) planktonic lifestyle, expanding the paradigm of
444 streamlining observed for oceanic AOA (11) to AOB, *Nitrospira*, and *Ca.* Nitrotoga in
445 oligotrophic freshwaters. Within the cosmopolitan *Nitrosospira*, we found ecotypes with a
446 gradient of genome reduction that maps onto their habitats' trophic gradient: from NspGL1 (1.4
447 Mb, low GC, upper lakes) to NspGL2b (1.5 Mb, upper lakes) to NspGL2a (1.6 Mb, Ontario) to
448 NspGL3 (1.8 Mb, Erie) (Fig 2). The thaumarchaeal NarchGL have markedly reduced regulatory

449 capacity like the open ocean strain *Nitrosopelagicus brevis* (11). The NOB NspiraGL have
450 genomes 50-60% smaller than described *Nitrospira* and dominate, along with the AOB NspGL1
451 and the AOA NarchGL, the deeper more oligotrophic basins, while *Ca.* Nitrotoga favor shallower
452 more productive basins. The emergence of Erie-specific ecotypes of both *Nitrosospira*
453 (NspGL3) and *Ca.* Nitrotoga (NtogaGL1b) demonstrates how distinct this habitat is compared to
454 the other lakes.

455 Nitrifiers inhabiting the transparent waters of the upper Great Lakes show distinctive
456 adaptations to light including diverse photolyases, ROS detoxification, and even
457 proteorhodopsin. This discovery is surprising, given that nitrifiers are rare in the surface mixed
458 layer of the Great Lakes (Fig. S1) and that photoinhibition of ammonia oxidation and nitrifier
459 growth is well documented (51-54,64). We propose that proteorhodopsin could be used to
460 augment energy metabolism when ammonia oxidation is photoinhibited. Water clarity has
461 increased over the past several decades in Lakes Michigan and Huron, now surpassing that of
462 Lake Superior (65). High light penetration along with seasonal mixing likely exposes deep water
463 cells to damaging levels of light and oxidative stress. Future cultivation and physiological studies
464 will document photoinhibition and potential phototrophy in Great Lakes nitriifiers.

465 The capacity for nitrification is found across multiple phyla, and our work unveils new
466 clues to understanding the ecological and evolutionary potential of these diverse lineages. This
467 collective nitrifier diversity undoubtedly influences the cycling of carbon and nitrogen across this
468 ecosystem, and future work will explore the differential contributions to nitrification by the distinct
469 lineages we described here. Understanding what controls the diversity of nitrifiers and other key
470 functional groups, and the consequences of this diversity for biogeochemistry, are essential for
471 forecasting the effects of rapid environmental change across the large lakes of the world (e.g.
472 ref. 91) and predicting impacts on the critical ecosystem services they provide (92).
473

References

1. Canfield DE, Glazer AN, Falkowski PG. The evolution and future of Earth's nitrogen cycle. *Science* 2010; 330:192–6.
2. Pachiadaki MG, Sintes E, Bergauer K, Brown JM, Record NR, Swan BK, et al. Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. *Science* 2017; 358:1046–51.
3. Reinthaler T, van Aken HM, Herndl GJ. Major contribution of autotrophy to microbial carbon cycling in the deep North Atlantic's interior. *Deep Sea Res Part II: Topical Studies in Oceanography* 2010; 57:1572–80.
4. Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, et al. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* 2011; 333:1296–300.
5. Baltar F, Herndl GJ. Ideas and perspectives: Is dark carbon fixation relevant for oceanic primary production estimates? *Biogeosciences* 2019; 16:3793–9.
6. Callieri C, Coci M, Eckert EM, Salcher MM, Bertoni R. Archaea and Bacteria in deep lake hypolimnion: in situ dark inorganic carbon uptake. *J Limnol* 2014; 73:31-38.
7. Wuchter C, Abbas B, Coolen MJL, Herfort L, van Bleijswijk J, Timmers P, et al. Archaeal nitrification in the ocean. *Proc Natl Acad Sci USA* 2006; 103:12317–22.
8. Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci USA* 2005; 102:14683–8.
9. Schleper C. Ammonia oxidation: different niches for bacteria and archaea? *ISME J* 2010; 4:1092–4.
10. Martens-Habbena W, Berube PM, Urakawa H, de la Torre JR, Stahl DA. Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* 2009; 461:976–9.
11. Santoro AE, Dupont CL, Richter RA, Craig MT, Carini P, McIlvin MR, et al. Genomic and proteomic characterization of "*Candidatus* Nitrosopelagicus brevis": An ammonia-oxidizing archaeon from the open ocean. *Proc Natl Acad Sci USA* 2015; 112:1173–8.
12. Auguet J-C, Triadó-Margarit X, Nomokonova N, Camarero L, Casamayor EO. Vertical segregation and phylogenetic characterization of ammonia-oxidizing Archaea in a deep oligotrophic lake. *ISME J* 2012; 6:1786–97.
13. Herber J, Klotz F, Frommeyer B, Weis S, Straile D, Kolar A, et al. A single Thaumarchaeon drives nitrification in deep oligotrophic Lake Constance. *Environ Microbiol* 2020; 22:212–28.
14. Urbach E, Vergin KL, Young L, Morse A, Larson GL, Giovannoni SJ. Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnol Oceanogr* 2001; 46:557–72.
15. Okazaki Y, Fujinaga S, Tanaka A, Kohzu A, Oyagi H, Nakano S. Ubiquity and quantitative significance of bacterioplankton lineages inhabiting the oxygenated hypolimnion of deep freshwater lakes. *ISME J* 2017; 11:2279–93.
16. Mukherjee M, Ray A, Post AF, McKay RM, Bullerjahn GS. Identification, enumeration and diversity of nitrifying planktonic archaea and bacteria in trophic end members of the Laurentian Great Lakes. *J Great Lakes Res* 2016; 42:39–49.
17. Hugoni M, Etien S, Bourges A, Lepère C, Domaizon I, Mallet C, et al. Dynamics of ammonia-oxidizing Archaea and Bacteria in contrasted freshwater ecosystems. *Res Microbiol* 2013; 164:360–70.
18. Bollmann A, Bar-Gilissen M-J, Laanbroek HJ. Growth at low ammonium concentrations and starvation response as potential factors involved in niche differentiation among ammonia-oxidizing bacteria. *Appl Environ Microbiol* 2002; 68:4751–7.
19. Sedlacek CJ, McGowan B, Suwa Y, Sayavedra-Soto L, Laanbroek HJ, Stein LY, et al. A

physiological and genomic comparison of *Nitrosomonas* cluster 6a and 7 ammonia-oxidizing bacteria. *Microb Ecol* 2019; 78:985–94.

20. Beman JM, Popp BN, Francis CA. Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. *ISME J* 2008; 2:429–41.
21. Daims H, Lücker S, Wagner M. A new perspective on microbes formerly known as nitrite-oxidizing bacteria. *Trends Microbiol* 2016; 24:699–712.
22. Kitzinger K, Marchant HK, Bristow LA, Herbold CW, Padilla CC, Kidane AT, et al. Single cell analyses reveal contrasting life strategies of the two main nitrifiers in the ocean. *Nature Comm* 2020; 11:767.
23. Nowka B, Daims H, Spieck E. Comparison of oxidation kinetics of nitrite-oxidizing bacteria: nitrite availability as a key factor in niche differentiation. *Appl Environ Microbiol* 2015; 81:745–53.
24. Wegen S, Nowka B, Spieck E. Low temperature and neutral pH define "*Candidatus* Nitrotoga sp." as a competitive nitrite oxidizer in coculture with *Nitrospira defluvii*. *Appl Environ Microbiol* 2019; 85:e02569-18.
25. Koch H, Galushko A, Albertsen M, Schintlmeister A, Gruber-Dorninger C, Lucker S, et al. Growth of nitrite-oxidizing bacteria by aerobic hydrogen oxidation. *Science* 2014; 345:1052–4.
26. Füssel J, Lücker S, Yilmaz P, Nowka B, van Kessel MAHJ, Bourceau P, et al. Adaptability as the key to success for the ubiquitous marine nitrite oxidizer *Nitrococcus*. *Science Advances* 2017; 3:e1700807.
27. Palatinszky M, Herbold C, Jehmlich N, Pogoda M, Han P, von Bergen M, et al. Cyanate as an energy source for nitrifiers. *Nature* 2015; 524:105–8.
28. Kitzinger K, Padilla CC, Marchant HK, Hach PF, Herbold CW, Kidane AT, et al. Cyanate and urea are substrates for nitrification by Thaumarchaeota in the marine environment. *Nat Microbiol* 2019; 4:234.
29. Koch H, Lücker S, Albertsen M, Kitzinger K, Herbold C, Spieck E, et al. Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus *Nitrospira*. *Proc Natl Acad Sci USA* 2015; 112:11371–6.
30. Boddicker AM, Mosier AC. Genomic profiling of four cultivated *Candidatus* Nitrotoga spp. predicts broad metabolic potential and environmental distribution. *ISME J* 2018; 12:2864–82.
31. Sterner RW. In situ-measured primary production in Lake Superior. *J Great Lakes Res* 2010; 36:139–49.
32. Small GE, Bullerjahn GS, Sterner RW, Beall BFN, Brovold S, Finlay JC, et al. Rates and controls of nitrification in a large oligotrophic lake. *Limnol Oceanogr* 2013; 58:276–86.
33. Vollenweider RA, Munawar M, Stadelmann P. A comparative review of phytoplankton and primary production in the Laurentian Great Lakes. *J Fish Res Bd Can* 1974; 31:739–62.
34. Clevinger CC, Heath RT, Bade DL. Oxygen use by nitrification in the hypolimnion and sediments of Lake Erie. *J Great Lakes Res* 2014; 40:202–7.
35. Neilson MA, Stevens RJJ. Spatial heterogeneity of nutrients and organic matter in Lake Ontario. *Can J Fish Aquat Sci* 1987; 44:2192–203.
36. Lean DRS, Knowles R. Nitrogen transformations in Lake Ontario. *Can J Fish Aquat Sci* 1987; 44:2133–43.
37. Paver SF, Newton RJ, Coleman ML. Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environ Microbiol* 2020; 22:433–46.
38. Rozmarynowycz MJ, Beall BFN, Bullerjahn GS, Small GE, Sterner RW, Brovold SS, et al. Transitions in microbial communities along a 1600 km freshwater trophic gradient. *J Great Lakes Res* 2019; 45:263–76.
39. Fujimoto M, Cavaletto J, Liebig JR, McCarthy A, Vanderploeg HA, Denef VJ. Spatiotemporal distribution of bacterioplankton functional groups along a freshwater

575 estuary to pelagic gradient in Lake Michigan. *J Great Lakes Res* 2016; 42:1036–48.
576 40. Barbiero RP, Lesht BM, Hinchey EK, Nettesheim TG. A brief history of the U.S. EPA Great
577 Lakes National Program Office's water quality survey. *J Great Lakes Res* 2018; 44:539–
578 46.
579 41. US Environmental Protection Agency. Sampling and analytical procedures for GLNPO's
580 open lake water quality survey of the Great Lakes. Great Lakes National Program Office,
581 Illinois, USA; 2003.
582 42. Holmes RM, Aminot A, Kérouel R, Hooker BA, Peterson BJ. A simple and precise method
583 for measuring ammonium in marine and freshwater ecosystems. *Can J Fish Aquat Sci*
584 1999; 56:1801–8.
585 43. Miranda KM, Espey MG, Wink DA. A rapid, simple spectrophotometric method for
586 simultaneous detection of nitrate and nitrite. *Nitric Oxide* 2001; 5:62–71.
587 44. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit
588 rRNA primers for marine microbiomes with mock communities, time series and global field
589 samples. *Environ Microbiol* 2016; 18:1403–14.
590 45. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal
591 RNA gene database project: improved data processing and web-based tools. *Nucleic*
592 *Acids Res* 2013; 41:D590–6.
593 46. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian classifier for rapid assignment
594 of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;
595 73:5261–7.
596 47. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-
597 index sequencing strategy and curation pipeline for analyzing amplicon sequence data on
598 the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* 2013; 79:5112–20.
599 48. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping:
600 differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods*
601 *Ecol Evol* 2013; 4:1111–9.
602 49. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile
603 metagenomic assembler. *Genome Res* 2017; 27:824–34.
604 50. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a
605 new genome assembly algorithm and its applications to single-cell sequencing. *J Comput*
606 *Biol* 2012; 19:455–77.
607 51. Hooper AB, Terry KR. Photoinactivation of ammonia oxidation in *Nitrosomonas*. *J Bacteriol*
608 1974; 119:899-906.
609 52. Horrigan SG, Springer AL. Oceanic and estuarine ammonium oxidation: Effects of light.
610 *Limnol Oceanogr* 1990; 35:479–82.
611 53. Guerrero M, Jones R. Photoinhibition of marine nitrifying bacteria. I. Wavelength-
612 dependent response. *Mar Ecol Prog Ser* 1996; 141:183–92.
613 54. Merbt SN, Stahl DA, Casamayor EO, Martí E, Nicol GW, Prosser JI. Differential
614 photoinhibition of bacterial and archaeal ammonia oxidation. *FEMS Microbiol Lett* 2012;
615 327:41–6.
616 55. Smith JM, Chavez FP, Francis CA. Ammonium uptake by phytoplankton regulates
617 nitrification in the sunlit ocean. *PLoS ONE* 2014; 9:e108173.
618 56. Rice MC, Norton JM, Valois F, Bollmann A, Bottomley PJ, Klotz MG, et al. Complete
619 genome of *Nitrosospira briensis* C-128, an ammonia-oxidizing bacterium from agricultural
620 soil. *Stand Genomic Sci* 2016; 11:46.
621 57. Garcia JC, Urakawa H, Le VQ, Stein LY, Klotz MG, Nielsen JL. Draft genome sequence of
622 *Nitrosospira sp*. Strain APG3, a psychrotolerant ammonia-oxidizing bacterium isolated
623 from sandy lake sediment. *Genome Announcements* 2013; 1:e00930-13.
624 58. Norton JM, Klotz MG, Stein LY, Arp DJ, Bottomley PJ, Chain PSG, et al. Complete
625 genome sequence of *Nitrosospira multiformis*, an ammonia-oxidizing bacterium from the

626 soil environment. *Appl Environ Microbiol* 2008; 74:3559–72.
627 59. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for
628 microbial ecology. *ISME J* 2014; 8:1553–65.
629 60. Bollmann A, Sedlacek CJ, Norton J, Laanbroek HJ, Suwa Y, Stein LY, et al. Complete
630 genome sequence of *Nitrosomonas sp*. Is79, an ammonia oxidizing bacterium adapted to
631 low ammonium concentrations. *Stand Genomic Sci* 2013; 7:469–82.
632 61. Reckel S, Gottstein D, Stehle J, Löhr F, Verhoefen M-K, Takeda M, et al. Solution NMR
633 structure of proteorhodopsin. *Angewandte Chemie* 2011; 50:11942–6.
634 62. Kralj JM, Bergo VB, Amsden JJ, Spudich EN, Spudich JL, Rothschild KJ. Protonation state
635 of Glu142 differs in the green- and blue-absorbing variants of proteorhodopsin. *Biochem*
636 2008; 47:3447–53.
637 63. Boeuf D, Audic S, Brillet-Guéguen L, Caron C, Jeanthon C. MicRhoDE: a curated
638 database for the analysis of microbial rhodopsin diversity and evolution. *Database* 2015;
639 bav080.
640 64. Hyman MR, Arp DJ. $^{14}C_2H_2$- and $^{14}CO_2$-labeling studies of the de novo synthesis of
641 polypeptides by *Nitrosomonas europaea* during recovery from acetylene and light
642 inactivation of ammonia monooxygenase. *J Biol Chem* 1992; 267:1534–45.
643 65. Yousef F, Shuchman R, Sayers M, Fahnenstiel G, Henareh A. Water clarity of the Upper
644 Great Lakes: Tracking changes between 1998–2012. *J Great Lakes Res* 2017; 43:239–47.
645 66. Beaumont HJE, Hommes NG, Sayavedra-Soto LA, Arp DJ, Arciero DM, Hooper AB, et al.
646 Nitrite reductase of *Nitrosomonas europaea* is not essential for production of gaseous
647 nitrogen oxides and confers tolerance to nitrite. *J Bacteriol* 2002; 184:2557–60.
648 67. Watzer B, Forchhammer K. Cyanophycin synthesis optimizes nitrogen utilization in the
649 unicellular cyanobacterium Synechocystis sp. strain PCC 6803. *Appl Environ Microbiol*
650 2018; 84:e01298-18.
651 68. Badger MR, Bek EJ. Multiple Rubisco forms in proteobacteria: their functional significance
652 in relation to CO_2 acquisition by the CBB cycle. *J Exp Bot* 2008; 59:1525–41.
653 69. Rae BD, Long BM, Badger MR, Price GD. Functions, compositions, and evolution of the
654 two types of carboxysomes: polyhedral microcompartments that facilitate CO_2 fixation in
655 cyanobacteria and some Proteobacteria. *Microbiol Mol Biol Rev* 2013; 77:357–79.
656 70. Cabello-Yeves PJ, Zemskaya TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov VV, et
657 al. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake
658 Baikal. *Appl Environ Microbiol* 2018; 84: e02132-17.
659 71. Sterner RW, Smutka TM, McKay RML, Xiaoming Q, Brown ET, Sherrell RM. Phosphorus
660 and trace metal limitation of algae and bacteria in Lake Superior. *Limnol Oceanogr* 2004;
661 49:495–507.
662 72. Bayer B, Vojvoda J, Offre P, Alves RJE, Elisabeth NH, Garcia JA, et al. Physiological and
663 genomic characterization of two novel marine thaumarchaeal strains indicates niche
664 differentiation. *ISME J* 2016; 10:1051–63.
665 73. Luo H, Tolar BB, Swan BK, Zhang CL, Stepanauskas R, Ann Moran M, et al. Single-cell
666 genomics shedding light on marine Thaumarchaeota diversification. *ISME J* 2014; 8:732–
667 6.
668 74. Ahlgren NA, Chen Y, Needham DM, Parada AE, Sachdeva R, Trinh V, et al. Genome and
669 epigenome of a novel marine Thaumarchaeota strain suggest viral infection,
670 phosphorothioation DNA modification and multiple restriction systems. *Environ Microbiol*
671 2017; 19:2434–52.
672 75. Ryals J, Hsu RY, Lipsett MN, Bremer H. Isolation of single-site *Escherichia coli* mutants
673 deficient in thiamine and 4-thiouridine syntheses: identification of a nuvC mutant. *J*
674 *Bacteriol* 1982; 151:899–904.
675 76. Karr EA. Transcription Regulation in the Third Domain. In: Sariaslani S, Gadd GM (eds).
676 *Advances in Applied Microbiology*. Academic Press, 2014, pp. 101–33.

77. Kitzinger K, Koch H, Lücker S, Sedlacek CJ, Herbold C, Schwarz J, et al. Characterization of the first "*Candidatus* Nitrotoga" isolate reveals metabolic versatility and separate evolution of widespread nitrite-oxidizing bacteria. *mBio* 2018; 9:e01186-18.
78. Ishii K, Fujitani H, Sekiguchi Y, Tsuneda S. Physiological and genomic characterization of a new "*Candidatus* Nitrotoga" isolate. *Environ Microbiol* 2020; 22:2365–82.
79. Starkenburg SR, Chain PSG, Sayavedra-Soto LA, Hauser L, Land ML, Larimer FW, et al. Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. *Appl Environ Microbiol* 2006; 72:2050–63.
80. Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B, et al. A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci USA* 2010; 107:13479–84.
81. Lücker S, Nowka B, Rattei T, Spieck E, Daims H. The genome of *Nitrospina gracilis* illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Front Microbiol* 2013; 4:27.
82. Mundinger AB, Lawson CE, Jetten MSM, Koch H, Lücker S. Cultivation and transcriptional analysis of a canonical *Nitrospira* under stable growth conditions. *Front Microbiol* 2019; 10:1325.
83. Sancar A. Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. *Chem Rev* 2003; 103:2203–38.
84. Zadow A von, Ignatz E, Pokorny R, Essen L-O, Klug G. *Rhodobacter sphaeroides* CryB is a bacterial cryptochrome with (6–4) photolyase activity. *FEBS J* 2016; 283:4291–309.
85. Zhang F, Scheerer P, Oberpichler I, Lamparter T, Krauß N. Crystal structure of a prokaryotic (6-4) photolyase with an Fe-S cluster and a 6,7-dimethyl-8-ribityllumazine antenna chromophore. *Proc Natl Acad Sci USA* 2013; 110:7217–22.
86. Maresca JA, Keffer JL, Hempel PP, Polson SW, Shevchenko O, Bhavsar J, et al. Light modulates the physiology of nonphototrophic Actinobacteria. *J Bacteriol* 2019; 201: e00740-18.
87. Wink DA, Kasprzak KS, Maragos CM, Elespuru RK, Misra M, Dunams TM, et al. DNA deaminating ability and genotoxicity of nitric oxide and its progenitors. *Science* 1991; 254:1001–3.
88. Imlay JA. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. *Nat Rev Microbiol* 2013; 11:443–54.
89. Zhang T, Hansel CM, Voelker BM, Lamborg CH. Extensive dark biological production of reactive oxygen species in brackish and freshwater ponds. *Environ Sci Technol* 2016; 50:2983–93.
90. Khademian M, Imlay JA. *Escherichia coli* cytochrome *c* peroxidase is a respiratory oxidase that enables the use of hydrogen peroxide as a terminal electron acceptor. *Proc Natl Acad Sci USA* 2017; 114:E6922–31.
91. O'Reilly CM, Sharma S, Gray DK, Hampton SE, Read JS, Rowley RJ, et al. Rapid and highly variable warming of lake surface waters around the globe. *Geophys Res Lett* 2015; 42:10773-10781.
92. Sterner RW, Keeler B, Polasky S, Poudel R, Rhude K, Rogers M. Ecosystem services of Earth's largest freshwater lakes. *Ecosystem Services* 2020; 41:101046.
93. Murphy TP. Ammonia and nitrate uptake in the lower Great Lakes. *Can J Fish Aq Sci* 1980; 37:1365–72.
94. Gardner WS, Lavrentyev PJ, Cavaletto JF, McCarthy MJ, Eadie BJ, Johengen TH, et al. Distribution and dynamics of nitrogen and microbial plankton in southern Lake Michigan during spring transition 1999-2000. *J Geophys Res* 2004; 109: C03007.

726 **Figure Legends**

727 **Figure 1.** Dissolved inorganic nitrogen availability and distribution of nitrifiers across the Great
728 Lakes. **(a)** Oxidized nitrogen concentrations. Values include NOx concentrations from published
729 studies (n=128; ref. 16,36,38,93,94), US EPA Water Quality Surveys in 2012 and 2013 (n=1626
730 from GLENDA database), and this study (n=20) (Table S2). **(b)** Ammonium concentrations.
731 Values are derived from the literature as in (a) (n=118) and from this study (n=20) (Table S3).
732 **(c)** Distribution of nitrifiers across the Great Lakes. Top panel: map of sampling stations. Stars
733 indicate stations chosen for metagenome analysis. Bottom panel: relative abundance of
734 ammonia-oxidizing (green) and nitrite-oxidizing (pink) families based on 16S rRNA V4-V5
735 amplicon sequencing, sampled in the mid-hypolimnion (except western Erie, sampled 1m from
736 bottom). Data is plotted roughly west to east as indicated on the map.

737

738 **Figure 2.** Genome properties and cross-lake distribution of ammonia oxidizing organisms,
739 showing both Archaea (top) and Betaproteobacteria (bottom). Rows highlighted in green
740 represent clusters of genomes reconstructed from the Great Lakes; expanded phylogenetic
741 trees are shown in Figures S2 and S5. For catalase, "E" indicates monofunctional catalase *katE*;
742 "G" indicates bifunctional catalase-peroxidase *katG*. For carbon fixation, RuBisCO type is shown
743 in parentheses (68); HP/HB, 3-Hydroxypropionate/4-Hydroxybutyrate cycle; CBB, Calvin-
744 Benson-Bassham cycle. Bubble plot shows composition of ammonia oxidizers in hypolimnion
745 samples, using MAGs as probes to recruit metagenomic reads (values sum to 100% for each
746 lake column). Genes identified in only a subset of genomes are shown as (+).

747

748 **Figure 3.** Evidence for proteorhodopsin (PR) in *Nitrosospira* from the Great Lakes. (a) Gene
749 neighborhood surrounding PR in *Nitrosospira* MAG MC17_S15_bin_110 and SAG 207399.
750 Genes are colored according to the best BLAST hit taxonomy in the NCBI nr database. (b)
751 Alignment of predicted *Nitrosospira* PR with reference sequences. Diagnostic features are

752 highlighted (61,62): blue boxes, diagnostic residues for PR; pink box, residue indicative of blue
753 or green tuning; green underline, shorter beta-sheet region in PR. Sequence accession
754 numbers: bacteriorhodopsin P02945, actinorhodopsin A0A1D9E0H1, sensory rhodopsin
755 P42196, *Dokdonia* PR EAQ40507.1, SAR86 blue-tuned PR Q4PP54, SAR86 green-tuned PR
756 Q9F7P4, SAR11 PR A6YQL7.
757
758 **Figure 4**. Genome properties and cross-lake distribution of nitrite oxidizing taxa *Nitrospira* (top)
759 and *Ca.* Nitrotoga (Betaproteobacteria*;* bottom). Rows highlighted in pink represent clusters of
760 genomes reconstructed from the Great Lakes. Expanded phylogenetic trees are shown in
761 Figures S2 and S6. rTCA, reductive tricarboxylic acid cycle; CBB, Calvin-Benson-Bassham
762 cycle; ONR, octaheme nitrite reductase. Values in parentheses indicate RuBisCO type (68).
763 Bubble plot shows composition of NOB per lake based on metagenomic read mapping. Genes
764 identified in only a subset of genomes are indicated by (+).
765
766 **Figure 5.** Distinct photolyase proteins in NspiraGL. **a)** Phylogenetic tree showing families of
767 photolyases. Three families are found in NspiraGL: CPD class I photolyase, FeS-BCP / CRYPro
768 family, and an uncharacterized CPF-related family found in diverse Bacteria. CPD Class I
769 photolyases are also found in other nitrifiers including *Ca*. Nitrotoga NtogaGL1a, *Nitrosospira*
770 NspGL1, and *Nitrosoarchaeum* NarchGL. **b)** Domain structure of the three photolyase families
771 present in NspiraGL.
772

773 **Acknowledgments**

774 Sequencing support for this project was provided by the DOE Joint Genome Institute
775 Community Sequencing Program (CSP #1565, #503460). Computational resources were
776 provided by the UChicago Research Computing Center. Funding for this work was provided by
777 Illinois-Indiana Sea Grant (Grant # NA14OAR170095), UChicago Women's Board, and the
778 National Science Foundation (OCE-1830011 to MLC). We thank Glenn Warren and the science
779 staff in the Great Lakes National Program Office of the US EPA, and the captain and crew of the
780 R/V *Lake Guardian*, for facilitating sample collection. We thank members of the Coleman and
781 Waldbauer labs for assistance with sample collection and processing, and for discussion and
782 comments on the manuscript.

783

784 **Data Availability**

785 The metagenome-assembled genomes presented here are available via NCBI BioProject
786 PRJNA636190. 16S rRNA data is available at NCBI BioProject PRJNA591360. Metagenomes
787 sequenced by JGI are available at http://genome.jgi.doe.gov, Project IDs 1045056, 1045059,
788 1045062, 1045065, 1045068, 1045071. Single cell amplified genome is available at
789 img.jgi.doe.gov, IMG Genome ID 3300033241.

790

791 **Competing Interests**

792 The authors declare no competing interests.

Figure 1

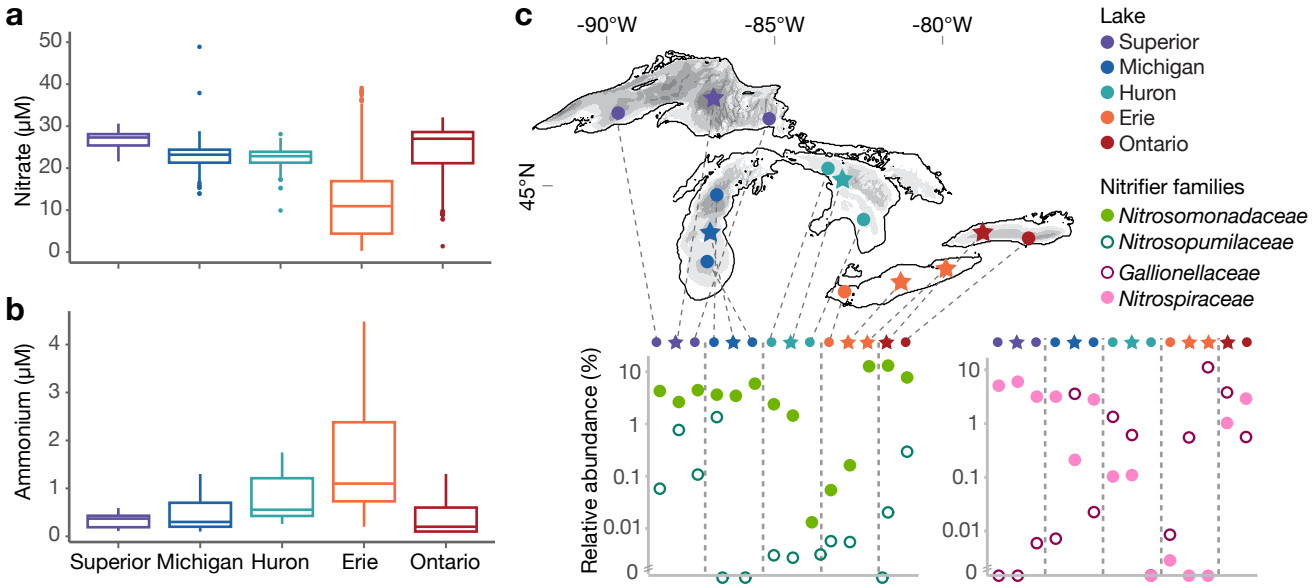

a
Nitrate (μM)
50
40
30
20
10
0
b
Ammonium (μM)
4
3
2
1
0
Superior
Michigan
Huron
Erie
Ontario
c
-90°W
-85°W
-80°W
45°N
Lake
Superior
Michigan
Huron
Erie
Ontario
Nitrifier families
Nitrosomonadaceae
Nitrosopumilaceae
Gallionellaceae
Nitrospiraceae
Relative abundance (%)
10
1
0.1
0.01
0

Figure 2

		% GC	Genome size (Mb)	Coding density (%)	Photophysiology: Catalase	Photophysiology: Photolyase	Photophysiology: Proteorhodopsin	Photophysiology: Retinal synthesis	N physiology: *nirK*	N physiology: *norCBQD*	N physiology: Cytochrome P460	N physiology: Nitrosocyanin	N physiology: Urease	Carbon fixation pathway
Thaumarchaea	NarchGL	33	1.3	90	-	+	-	-	+	-	-	-	+	HP/HB
	Nitrosoarchaeum limnia SFB1	32	1.7	89	-	+	-	-	+	-	-	-	-	HP/HB
	Nitrosopumilus maritimus SCM1	34	1.6	92	-	-	-	-	+	-	-	-	-	HP/HB
	Nitrosopelagicus brevis CN25	33	1.2	94	-	+	-	-	+	-	-	-	-	HP/HB
Betaproteo.	NspGL1	41	1.4	89	G	+	+	+	+	-	-	-	+	CBB (IAc)
	NspGL3	50	1.8	90	-	-	-	-	-	-	+	-	-	CBB (IAq)
	NspGL2a	50	1.6	90	-	-	-	-	-	-	+	-	-	CBB (IC)
	NspGL2b	50	1.5	89	-	-	-	-	(+)	-	+	-	-	CBB (IC)
	Nitrosospira briensis C-128	54	3.2	83	E	-	-	-	+	-	-	+	+	CBB (IC)
	Nitrosospira multiformis ATCC 25196	54	3.2	86	E	-	-	-	+	+	-	+	+	CBB (IC)
	Nitrosomonas ureae Nm10	44	3.3	86	E	+	-	-	+	-	+	+	+	CBB (IC/IAq)
	Nitrosomonas europaea ATCC 19718	50	3.8	88	EG	-	-	-	+	+	+	+	-	CBB (IAq)

Distribution across lakes

Figure 3

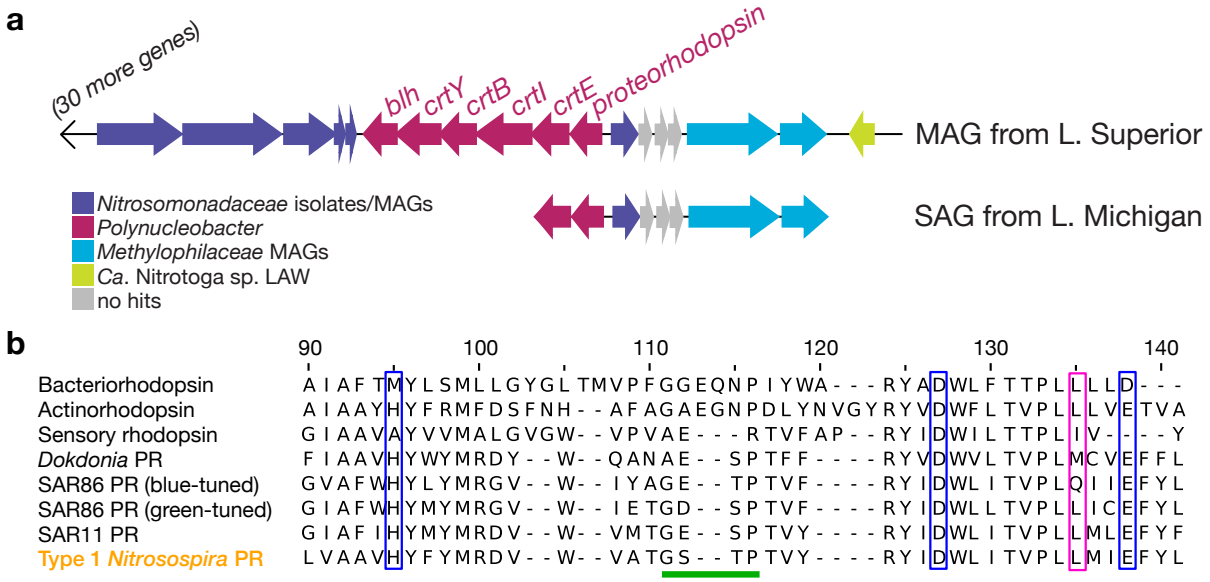

a
(30 more genes)
blh
crtY
crtB
crtI
crtE
proteorhodopsin
MAG from L. Superior
SAG from L. Michigan
Nitrosomonadaceae isolates/MAGs
Polynucleobacter
Methylophilaceae MAGs
Ca. Nitrotoga sp. LAW
no hits
b
90
100
110
120
130
140
Bacteriorhodopsin
Actinorhodopsin
Sensory rhodopsin
Dokdonia PR
SAR86 PR (blue-tuned)
SAR86 PR (green-tuned)
SAR11 PR
Type 1 Nitrosospira PR

Figure 4

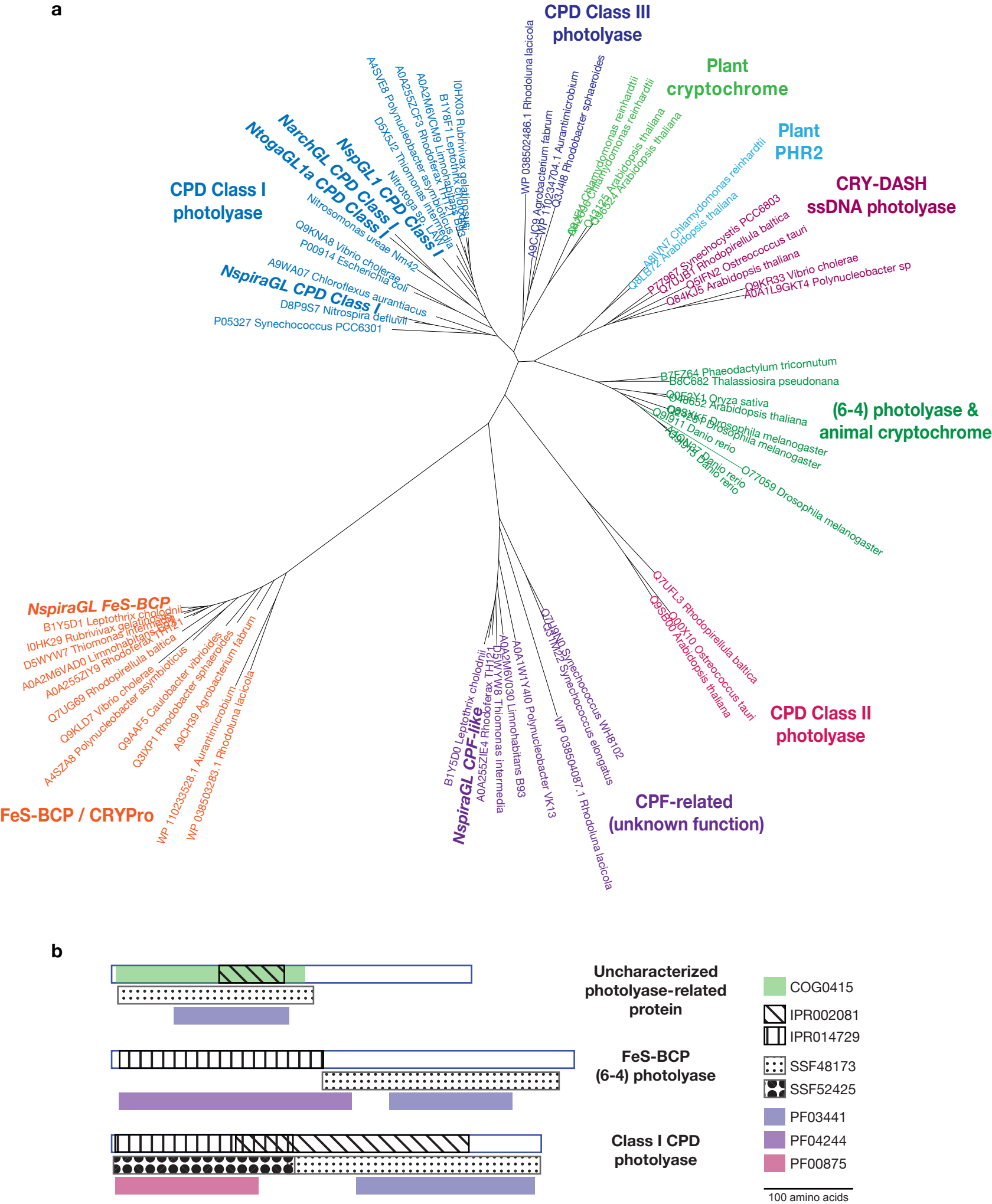
		% GC	Genome size (Mb)	Coding density (%)	# Sigma factors	Photolyase	Catalase	Formate dehydrogenase	NiFe hydrogenase	Thiosulfate dehydrogenase (*tsdA*)	*N physiology*						Carbon fixation pathway
											Assim. nitrite reduction	*nirK*	*norQD*	# *nxrAB* copies	Cyanase	Urease	
Nitrospira	NspiraGL	58	1.8	89	5	+++	G	-	-	-	*nirA*	-	+	1	+	(+)	rTCA
	Nitrospira japonica NJ1	58	4.1	90	27	+	E	-	-	-	ONR	+	+	3	+	+	rTCA
	Nitrospira moscoviensis NSP M-1	61	4.5	89	18	+	EG	+	+	+	ONR	+	+	5	+	+	rTCA
	Nitrospira defluvii	59	4.3	89	17	+	-	+	-	+	*nirA*	+	+	2	+	-	rTCA
Betaproteo.	NtogaGL1a	51	1.4	91	4	(+)	G	-	-	-	-	-	-	1	+	-	CBB (IC)
	NtogaGL1b	52	1.5	91	4	-	G	-	-	+	-	-	-	1	+	-	CBB (IC)
	Ca. Nitrotoga CP45	49	2.8	85	8	+	EG	-	+	-	*nirBD*	+	+	1	-	-	CBB (IC, ID/IA)
	Ca. Nitrotoga KNB	50	2.6	85	6	+	G	-	+	-	*nirBD*	+	+	2	-	-	CBB (IC, ID/IA)

Distribution across lakes

Relative abundance

25 50 75

Figure 5


a
CPD Class III photolyase
Plant cryptochrome
Plant PHR2
CRY-DASH ssDNA photolyase
(6-4) photolyase & animal cryptochrome
CPD Class II photolyase
CPF-related (unknown function)
FeS-BCP / CRYPro
CPD Class I photolyase
NspiraGL FeS-BCP
NspiraGL CPF-like
NspiraGL CPD Class I
NspGL1 CPD Class I
NarchGL CPD Class I
NtogaGL1a CPD Class I
WP 038502486.1 Rhodoluna lacicola
A9CJC9 Agrobacterium fabrum
WP 110234704.1 Aurantimicrobium
Q3J4J8 Rhodobacter sphaeroides
Q42125 Arabidopsis thaliana
Q9SZ24 Arabidopsis thaliana
A8IVN7 Chlamydomonas reinhardtii
Q8LB72 Arabidopsis thaliana
P77967 Synechocystis PCC6803
Q7UJB1 Rhodopirellula baltica
Q5IFN2 Ostreococcus tauri
Q84KJ5 Arabidopsis thaliana
Q9KR33 Vibrio cholerae
A0A1L9GKT4 Polynucleobacter sp
B7F764 Phaeodactylum tricornutum
B8C682 Thalassiosira pseudonana
Q0F2Y1 Oryza sativa
Q9SXK5 Drosophila melanogaster
Q9I911 Danio rero
A9JUJ7 Danio rerio
O77059 Drosophila melanogaster
Q7UFL3 Rhodopirellula baltica
Q9SB00 Arabidopsis thaliana
Q00Y10 Ostreococcus tauri
Q7UN22 Synechococcus WH8102
WP 038504087.1 Rhodoluna lacicola
A0A1W1Y4I0 Polynucleobacter VK13
A0A2M6V030 Limnohabitans B93
A0A255ZIE4 Rhodoferax THI21
B1Y5D0 Leptothrix cholodnii
I0HX03 Rubrivivax gelatinosus
B1Y8F1 Leptothrix cholodnii
A0A2M6VCM9 Limnohabitans
A0A255ZCF3 Rhodoferax
D5X5J2 Thiomonas intermedia
A4SVE8 Polynucleobacter asymbioticus
Nitrosomonas ureae Nm42
Q9KNA8 Vibrio cholerae
P00914 Escherichia coli
A9WA07 Chloroflexus aurantiacus
D8P9S7 Nitrospira defluvii
P05327 Synechococcus PCC6301
B1Y5D1 Leptothrix cholodnii
I0HK29 Rubrivivax gelatinosus
D5WYW7 Thiomonas intermedia
A0A2M6VAD0 Limnohabitans
A0A255ZIY9 Rhodoferax
Q7UG69 Rhodopirellula baltica
Q9KLD7 Vibrio cholerae
A4SZA8 Polynucleobacter asymbioticus
Q9AAF5 Caulobacter vibrioides
Q3IXP1 Rhodobacter sphaeroides
A9CH39 Agrobacterium fabrum
WP 110233528.1 Aurantimicrobium
WP 038503283.1 Rhodoluna lacicola
b
Uncharacterized photolyase-related protein
FeS-BCP (6-4) photolyase
Class I CPD photolyase
COG0415
IPR002081
IPR014729
SSF48173
SSF52425
PF03441
PF04244
PF00875
100 amino acids

Supplementary Information

Extended Methods

Metagenome QC and Assembly

Raw reads for spring surface samples were quality controlled at the Joint Genome Institute, using bbduk.sh for adapter trimming (ktrim=r, minlen=40, minlenfraction=0.6,mink=11, tbo, tpe, k=23, hdist=1, hdist2=1, ftm=5) and quality filtering (maq=8, maxns=1, minlen=40, minlenfraction=0.6, k=27, hdist=1, trimq=12, qtrim=rl). Raw reads for summer hypolimnion samples were adapter trimmed, quality filtered, and interleaved using bbduk (parameters: ktrim=r, mink=8, hdist=2, k=21, forcetrimleft=10, forcetrimright=199, minlen=150) using BBTools suite version 35.74 (https://sourceforge.net/projects/bbmap/). Separate assemblies of quality filtered reads were carried out for each metagenome using metaSPAdes 3.1.11 --meta mode using default k sizes of 21, 33, 55 (1). For mapping, forward and reverse reads were merged using bbmerge in BBtools, using qtrim2=r trimq=10,13,16 and adapter=default. Merged short reads from all 12 metagenomes were mapped onto each assembly using bowtie2 in --sensitive mode (2), and together this information was used to bin assembled contigs. Binning was performed using MetaBAT2 (3), Binsanity (4) and CONCOCT (5) using default parameters. The resulting bins were scored, aggregated, and de-replicated using DAS_Tool (6) followed by manual curation using Anvi'o (7). We assessed genome completion and contamination of manually curated bins using CheckM lineage_wf (8), and all genomes presented here are greater than 70% complete and have less than 10% contamination (Table S1). Potential nitrifiers were screened by searching for ammonia monooxygenase, hydroxylamine oxidoreductase and nitrite oxidoreductase within reconstructed genomes using blastp (9). For bins where any of these genes were detected, we identified bacterial single copy core genes (10) or archaeal single copy core genes (11) using HMMER (12), as implemented in Anvi'o. Single copy core genes were queried against proteins predicted from bacterial and archaeal

genomes in RefSeq (13), and taxonomic identity of these core genes was ascertained based on a least common ancestor approach using a 0.1% window around the bit score of the best hit using KronaTools (14). Taxonomic assignment was further validated using GTDB-tk (15). Grouping of MAGs into clades and subclades based on ANI was carried out using fastANI (16).

Annotation and gene clustering

Genes from putative nitrifier genomes were annotated using InterProScan 5.30-69.0 (17) and eggnog-mapper against the bactNOG database (18). Gene cluster analysis was carried out using the Anvi'o pangenome pipeline (19), using blastp to determine sequence similarity, ITEP to eliminate weak similarity (20) and MCL to cluster, using a minbit of 0.5, MCL inflation of 2 and minimum gene occurrence of 1 (21). Assignment of KEGG Orthology to gene clusters was carried out using GhostKOALA (22). Sigma factors were tallied by identifying gene clusters annotated with the following PFAMs: PF00309, PF03979, PF00140, PF04542, PF04539, PF04545, PF08281. Prokka (23) was used to generate GenBank file format files from MAGs and SAGs, and genoPlotR (24) was used to generate gene neighborhood maps

Gene tree construction

The NspGL1 proteorhodopsin sequence was inserted into the MicRhoDE rhodopsin tree using pplacer (25) through the MicRhoDE Galaxy pipeline (26). We then constructed a more targeted phylogenetic tree using aligned reference sequences of Supercluster III from MicRhoDE, filtered to exclude fragments shorter than 220 amino acids. To this alignment, we added NspGL1 sequences using MAFFT (27) along with high similarity sequences from NCBI nr that were not present in MicRhoDE. The tree was inferred using RaxML with model PROTGAMMALG (28). The tree was visualized in iTOL (29) and more distant clusters were collapsed for clarity.

A cyanase phylogenetic tree was created using sequences drawn from querying NtogaGL cyanase against NCBI nr using blastp, as well as sequences from references (30–32).

Sequences were aligned using MAFFT (27) and the tree was inferred using RaxML with model PROTGAMMALG (28). Tree was visualized in iTOL (29) and branches were colored based on the taxonomy of the parent genome.

Photolyase-related proteins in GL MAGs were identified by searching for the following features: K01669, COG0415, PF03441, PF00875, PF04244, SSF48173, SSF52425. Reference proteins (n=56) spanning the previously defined families of photolyases and cryptochromes (33) were obtained from UniProt, along with aquatic bacterial sequences described by Maresca and colleagues (34). The reference sequences were aligned using MAFFT (27), and sequences from GL MAGs were added using the MAFFT --addfragments option. The tree was estimated using IQ-TREE 2 (35) and visualized using iTOL (29).

Phylogenomic tree construction

All *Nitrospirae*, *Thaumarchaeota*, *Gallionellaceae* and *Nitrosomonadaceae* genomes were downloaded from Genbank (36) and included in phylogenetic trees for their respective family. Single copy core genes were extracted as described above and individually aligned at the protein level using muscle (37). Core proteins were then concatenated for each genome. Concatenated alignments were trimmed using Gblocks (38) and analyzed by RAxML 8.2.12 (28) to create a phylogenetic tree using the PROTGAMMALG model and 50 bootstraps. Trees were visualized in iTOL (29).

Proteorhodopsin assembly verification

We used several approaches to validate the presence of proteorhodopsin in assembled *Nitrosospira* genomes, to rule out the possibility of chimeric assemblies from different species. We note that proteorhodopsin-containing contigs were independently assembled and binned together with core *Nitrosospira* contigs from seven different samples (i.e. each sample was

assembled and binned separately, rather than co-assembled). In five of seven cases, proteorhodopsin and retinal biosynthesis genes were assembled together with core *Nitrosospira* genes on the same contig. To rule out a systematic reproducible error in assembly and/or binning, we compared these seven MAGs to a single cell amplified *Nitrosospira* genome (SAG) from Lake Michigan, obtained as part of another project with the JGI. This SAG was processed through JGI's standard decontamination pipeline and manually investigated to ensure lack of contamination. We found no evidence of contaminating core genes, as all core genes had best hits to either *Nitrosospira* or more generally *Nitrosomonadaceae* in nr. SAG contigs were matched to homologous contigs from NspGL1 MAGs to determine if any SAG contigs were unique using FastANI (16) with --visualize flag. All contigs from this *Nitrosospira* SAG were found within an NspGL1 MAG. Bandage (39) was used to manually inspect the assembly graph around the contig that contained the NspGL1 *Nitrosospira* proteorhodopsin to ensure that the assembled contig did not represent a chimeric contig or inappropriate scaffolding. We verified that a single, unique path exists from the beginning to the end of the NspGL1 contig containing proteorhodopsin (Fig 3, Fig S4). Further, we verified that consistent coverage across this contig existed by mapping short reads from the original sample using bowtie2 (2) and viewing results using Integrated Genomics Viewer (40) A closely related assembly of the same genomic region from Lake Biwa did not show evidence of proteorhodopsin; to confirm this difference between the Biwa and Great Lakes MAGs, we mapped reads from Lake Biwa (41) (BioProject PRJDB6644) onto the assembled contig described above (Fig 3, Fig S4) using bowtie2 (2). This analysis demonstrated that while a large fraction of the NspGL1 contig in question recruited reads from Lake Biwa at high identity (98-99%), starting upstream of proteorhodopsin and retinal biosynthesis, this contig no longer recruited reads from Lake Biwa.

Manual identification of key nitrification genes

Despite recovery of 15 high completion MAGs in NspGL1/2a/2b/3, many of these MAGs

lacked key nitrification genes in *amo* and *hao* operons. This was largely due to the fact that *amo* and *hao* operons were often assembled on small contigs below the minimum size cutoff we imposed for binning contigs. Difficulty in assembling these contigs was likely in part due to the several *amo* and *hao* operons with extremely high identity to one another in each genome, a phenomenon which has been observed in other *Nitrosospira* (42). Manual assembly graph inspection with Bandage (39) supported this hypothesis, as did assessment of abundance of short reads associated with *amo* operons from NspGL and comparison of abundance of short reads associated with core gene *rpoB* from NspGL, using ROCker (43). Still, an exemplar MAG from at least one representative of each ecotype (NspGL1/2a/2b/3) was found with both *amo* and *hao* operons. Further, manual inspection of unbinned contigs confirmed that *amo* and *hao* operons existed on contigs in every sample from which a MAG for a particular ecotype was recovered. That is, for every time that an NspGL1 MAG was recovered from a sample, we were able to determine that an *amo* and *hao* operon which could be affiliated with NspGL1 existed, even if it was not correctly binned. Affiliation for these unbinned key nitrification genes was carried out by alignment of *amoAB* and *haoAB* sequences to *amoAB* and *haoAB* sequences correctly binned in NspGL ecotypes. This process was also carried out for two NtogaGL1a MAGs for *nxrAB*, which were poorly assembled in those two samples.

Data analysis and plots

All data were analyzed in R version 3.5.3 (44) and plots were generated using ggplot2 (45).

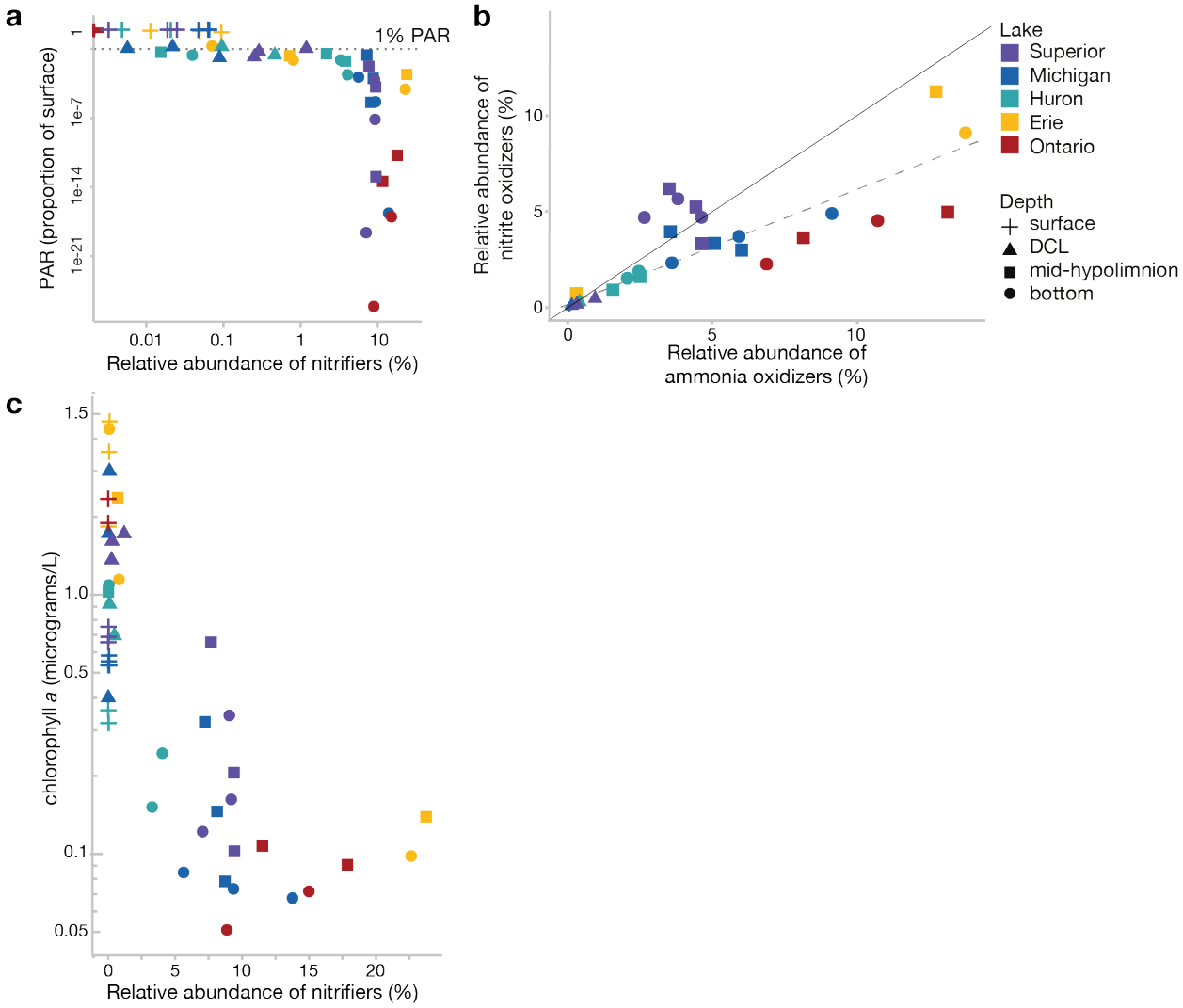


Figure S1. Relationships between environmental factors and relative abundance of nitrifiers in the Great Lakes. Relative abundance based on 16S rRNA V4-V5 region amplicon sequencing from Summer 2012. Nitrifier abundance includes sequences assigned to the familes *Nitrosopumilaceae* (AOA), *Nitrosomonadaceae* (AOB), *Nitrospiraceae* (NOB), and *Gallionellaceae* (NOB). (a) Relative abundance of nitrifiers (% of total community) compared to photosynthetically active radiation (PAR). (b) Correlation between the relative abundances of nitrite oxidizers and ammonia oxidizers in each sample. A 1:1 line (solid) and linear regression fit (dashed) are shown for reference. (c) Relative abundance of nitrifiers (% of total community) compared to chlorophyll *a* concentration. PAR and chlorophyll data from the EPA Great Lakes Environmental Database (https://cdx.epa.gov/).

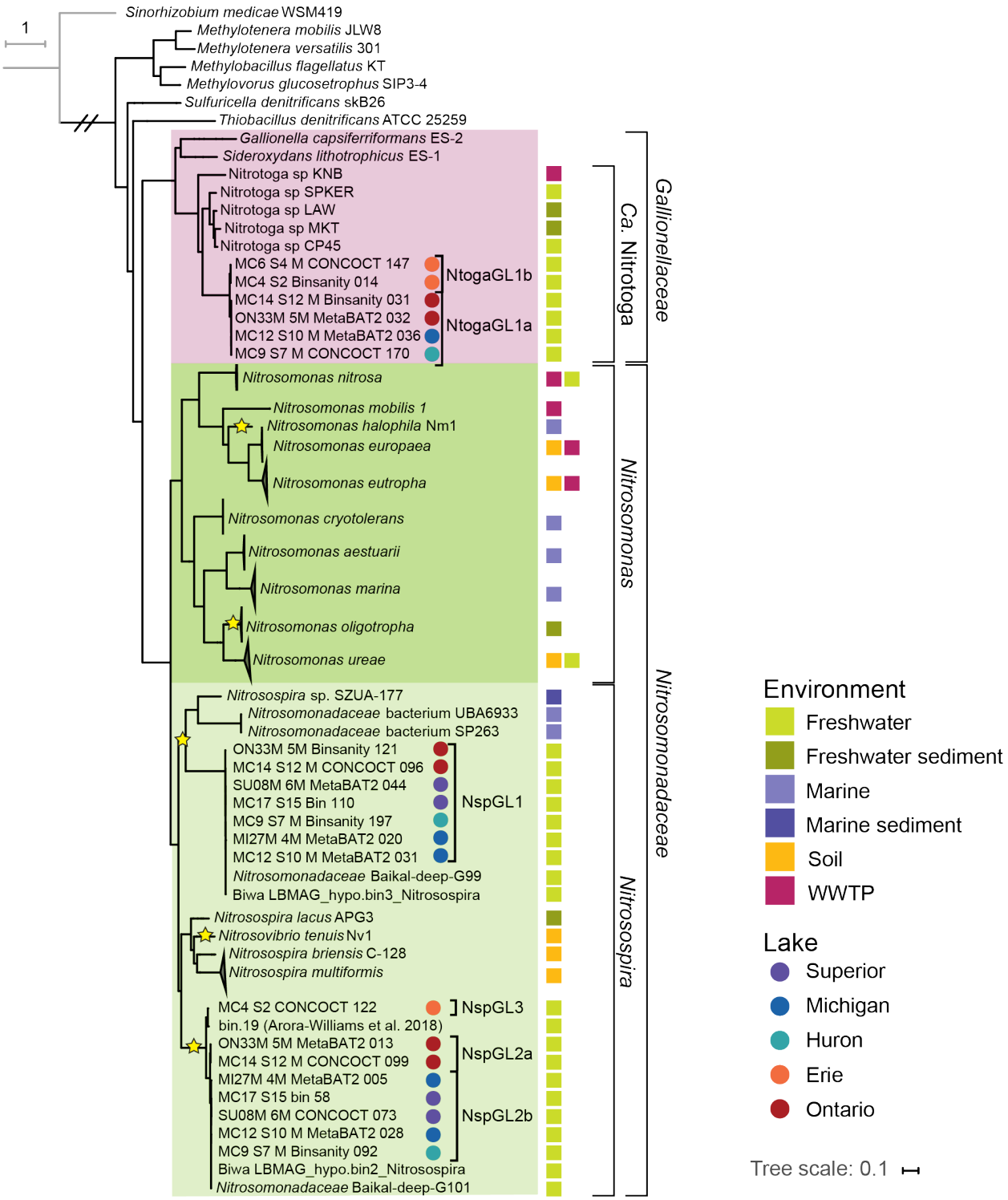


Figure S2. Phylogenomic tree of the *Nitrosomonadales* order of *Betaproteobacteria*, focusing on the *Nitrosomonadaceae* and *Gallionellaceae* families. Tree was constructed using 139 concatenated single-copy core genes (10). Colored squares indicate the environment of origin; colored circles denote the lake of origin for genomes from the Laurentian Great Lakes; yellow stars indicate clades containing strains that lack nitrosocyanin.

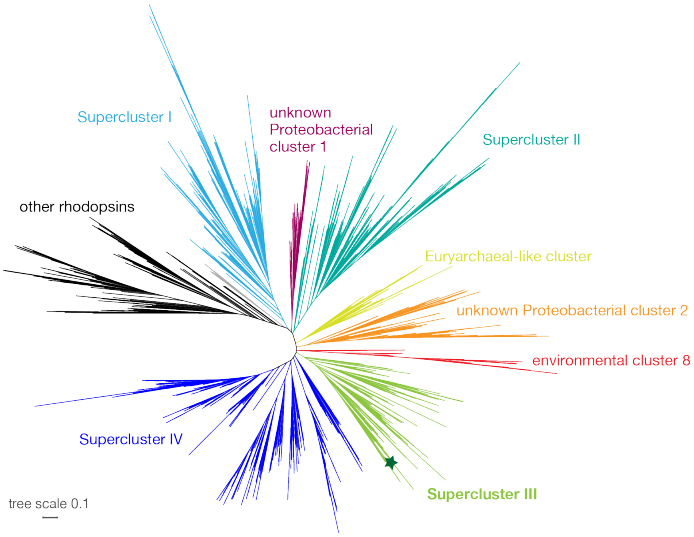
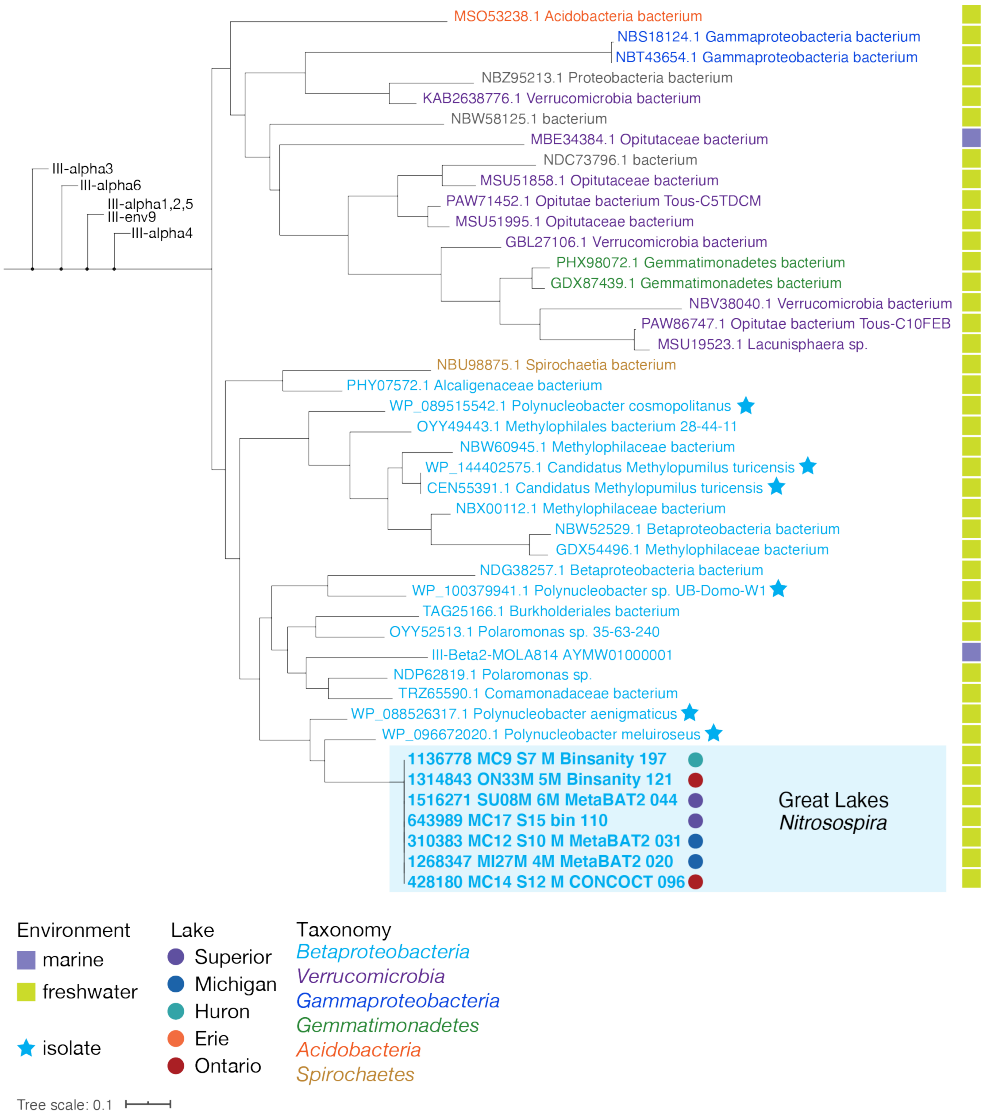


Figure S3. Proteorhodopsin phylogeny. Top: Phylogenetic placement of *Nitrosospira* PR. Rhodopsin sequences and classifications from MicRhoDE database (26); star shows NspGL1 PR. Bottom: Close relatives of *Nitrosospira* PR within Supercluster III. Neighboring clusters have been collapsed for clarity.



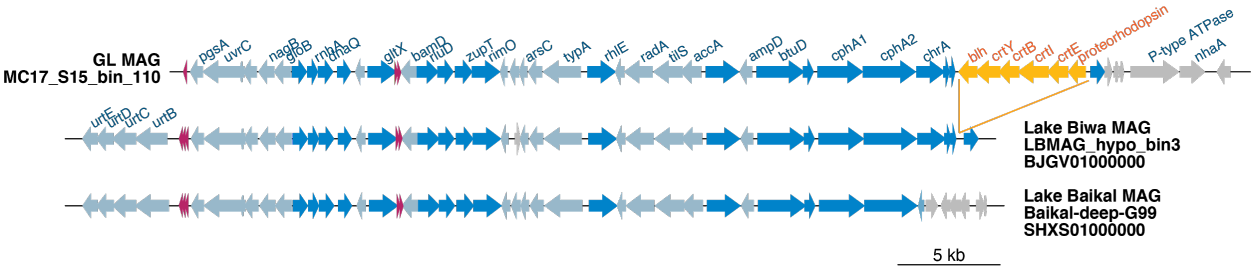


Figure S4. Shared synteny surrounding the proteorhodopsin module in *Nitrosospira* draft genomes from deep lakes. Light blue, shared genes on reverse strand; dark blue, shared genes on forward strand; pink, tRNA genes; orange, PR module in NspGL1 MAG; gray, genes unique to one MAG.

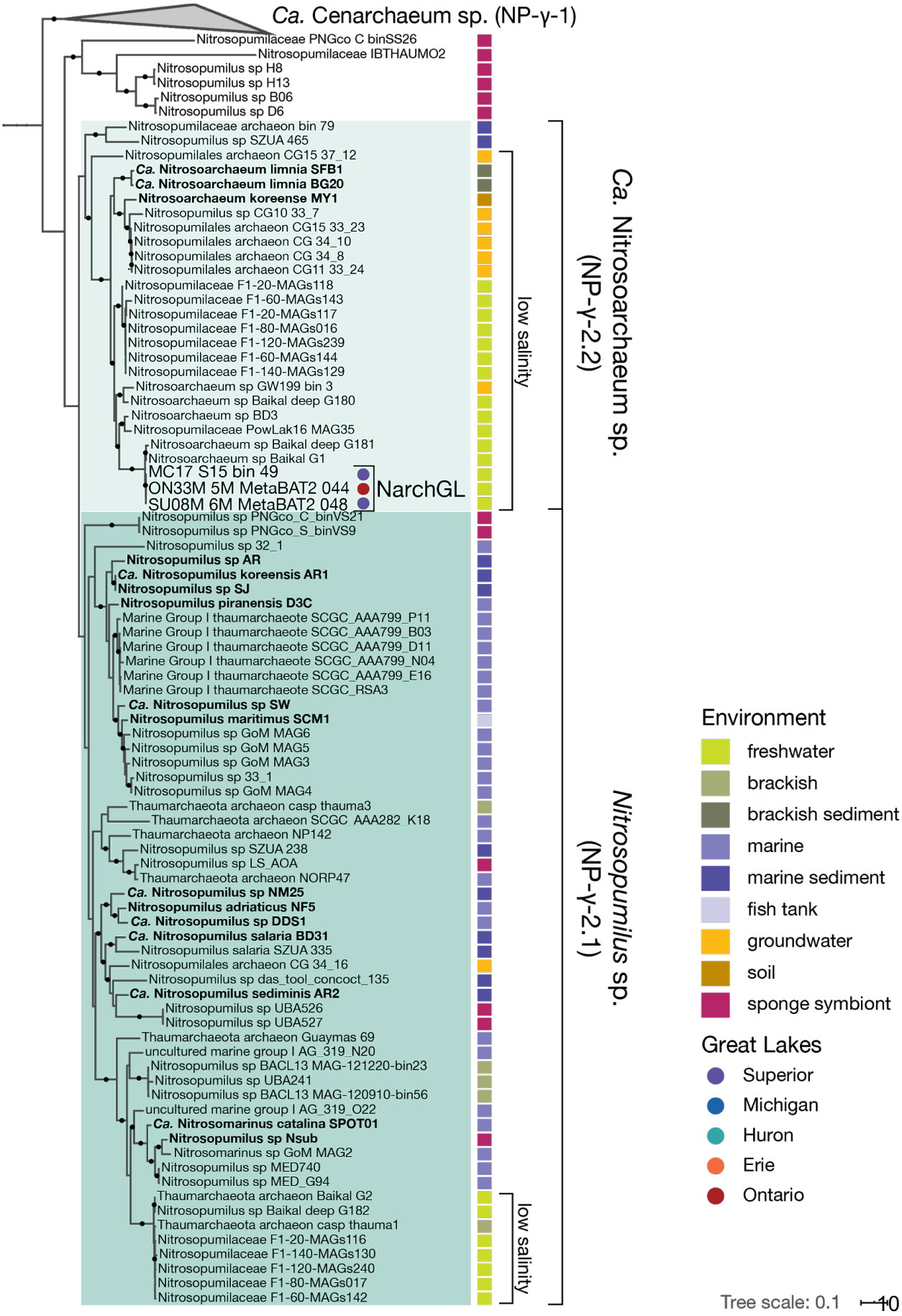

Ca. Cenarchaeum sp. (NP-γ-1)
Nitrosopumilaceae PNGco C binSS26
Nitrosopumilaceae IBTHAUMO2
Nitrosopumilus sp H8
Nitrosopumilus sp H13
Nitrosopumilus sp B06
Nitrosopumilus sp D6
Nitrosopumilaceae archaeon bin 79
Nitrosopumilus sp SZUA 465
Nitrosopumilales archaeon CG15 37_12
Ca. Nitrosoarchaeum limnia SFB1
Ca. Nitrosoarchaeum limnia BG20
Nitrosoarchaeum koreense MY1
Nitrosopumilus sp CG10 33_7
Nitrosopumilales archaeon CG15 33_23
Nitrosopumilales archaeon CG 34_10
Nitrosopumilales archaeon CG 34_8
Nitrosopumilales archaeon CG11 33_24
Nitrosopumilaceae F1-20-MAGs118
Nitrosopumilaceae F1-60-MAGs143
Nitrosopumilaceae F1-20-MAGs117
Nitrosopumilaceae F1-80-MAGs016
Nitrosopumilaceae F1-120-MAGs239
Nitrosopumilaceae F1-60-MAGs144
Nitrosopumilaceae F1-140-MAGs129
Nitrosoarchaeum sp GW199 bin 3
Nitrosoarchaeum sp Baikal deep G180
Nitrosoarchaeum sp BD3
Nitrosopumilaceae PowLak16 MAG35
Nitrosoarchaeum sp Baikal deep G181
Nitrosoarchaeum sp Baikal G1
MC17 S15 bin 49
ON33M 5M MetaBAT2 044
SU08M 6M MetaBAT2 048
NarchGL
Nitrosopumilus sp PNGco_C_binVS21
Nitrosopumilus sp PNGco_S_binVS9
Nitrosopumilus sp 32_1
Nitrosopumilus sp AR
Ca. Nitrosopumilus koreensis AR1
Nitrosopumilus sp SJ
Nitrosopumilus piranensis D3C
Marine Group I thaumarchaeote SCGC_AAA799_P11
Marine Group I thaumarchaeote SCGC_AAA799_B03
Marine Group I thaumarchaeote SCGC_AAA799_D11
Marine Group I thaumarchaeote SCGC_AAA799_N04
Marine Group I thaumarchaeote SCGC_AAA799_E16
Marine Group I thaumarchaeote SCGC_RSA3
Ca. Nitrosopumilus sp SW
Nitrosopumilus maritimus SCM1
Nitrosopumilus sp GoM MAG6
Nitrosopumilus sp GoM MAG5
Nitrosopumilus sp GoM MAG3
Nitrosopumilus sp 33_1
Nitrosopumilus sp GoM MAG4
Thaumarchaeota archaeon casp thauma3
Thaumarchaeota archaeon SCGC_AAA282_K18
Thaumarchaeota archaeon NP142
Nitrosopumilus sp SZUA 238
Nitrosopumilus sp LS_AOA
Thaumarchaeota archaeon NORP47
Ca. Nitrosopumilus sp NM25
Nitrosopumilus adriaticus NF5
Ca. Nitrosopumilus sp DDS1
Ca. Nitrosopumilus salaria BD31
Nitrosopumilus salaria SZUA 335
Nitrosopumilales archaeon CG 34_16
Nitrosopumilus sp das_tool_concoct_135
Ca. Nitrosopumilus sediminis AR2
Nitrosopumilus sp UBA526
Nitrosopumilus sp UBA527
Thaumarchaeota archaeon Guaymas 69
uncultured marine group I AG_319_N20
Nitrosopumilus sp BACL13 MAG-121220-bin23
Nitrosopumilus sp UBA241
Nitrosopumilus sp BACL13 MAG-120910-bin56
uncultured marine group I AG_319_O22
Ca. Nitrosomarinus catalina SPOT01
Nitrosopumilus sp Nsub
Nitrosomarinus sp GoM MAG2
Nitrosopumilus sp MED740
Nitrosopumilus sp MED_G94
Thaumarchaeota archaeon Baikal G2
Nitrosopumilus sp Baikal deep G182
Thaumarchaeota archaeon casp thauma1
Nitrosopumilaceae F1-20-MAGs116
Nitrosopumilaceae F1-140-MAGs130
Nitrosopumilaceae F1-120-MAGs240
Nitrosopumilaceae F1-80-MAGs017
Nitrosopumilaceae F1-60-MAGs142
low salinity
low salinity
Ca. Nitrosoarchaeum sp. (NP-γ-2.2)
Nitrosopumilus sp. (NP-γ-2.1)
Environment
freshwater
brackish
brackish sediment
marine
marine sediment
fish tank
groundwater
soil
sponge symbiont
Great Lakes
Superior
Michigan
Huron
Erie
Ontario
Tree scale: 0.1
10

Figure S5. Phylogenomic tree of *Nitrosopumilaceae* and *Cenarchaeum*. Tree was constructed using concatenated single-copy core genes as defined in (11). This tree represents a zoomed in view of *Nitrosopumilaceae* and *Cenarchaeum* from larger tree of all *Thaumarchaeota* and *Aigarchaeota*. Colored boxes correspond to the environment of origin; genomes from the Laurentian Great Lakes are indicated by colored circles. Names in bold indicate cultured isolates or enrichments.

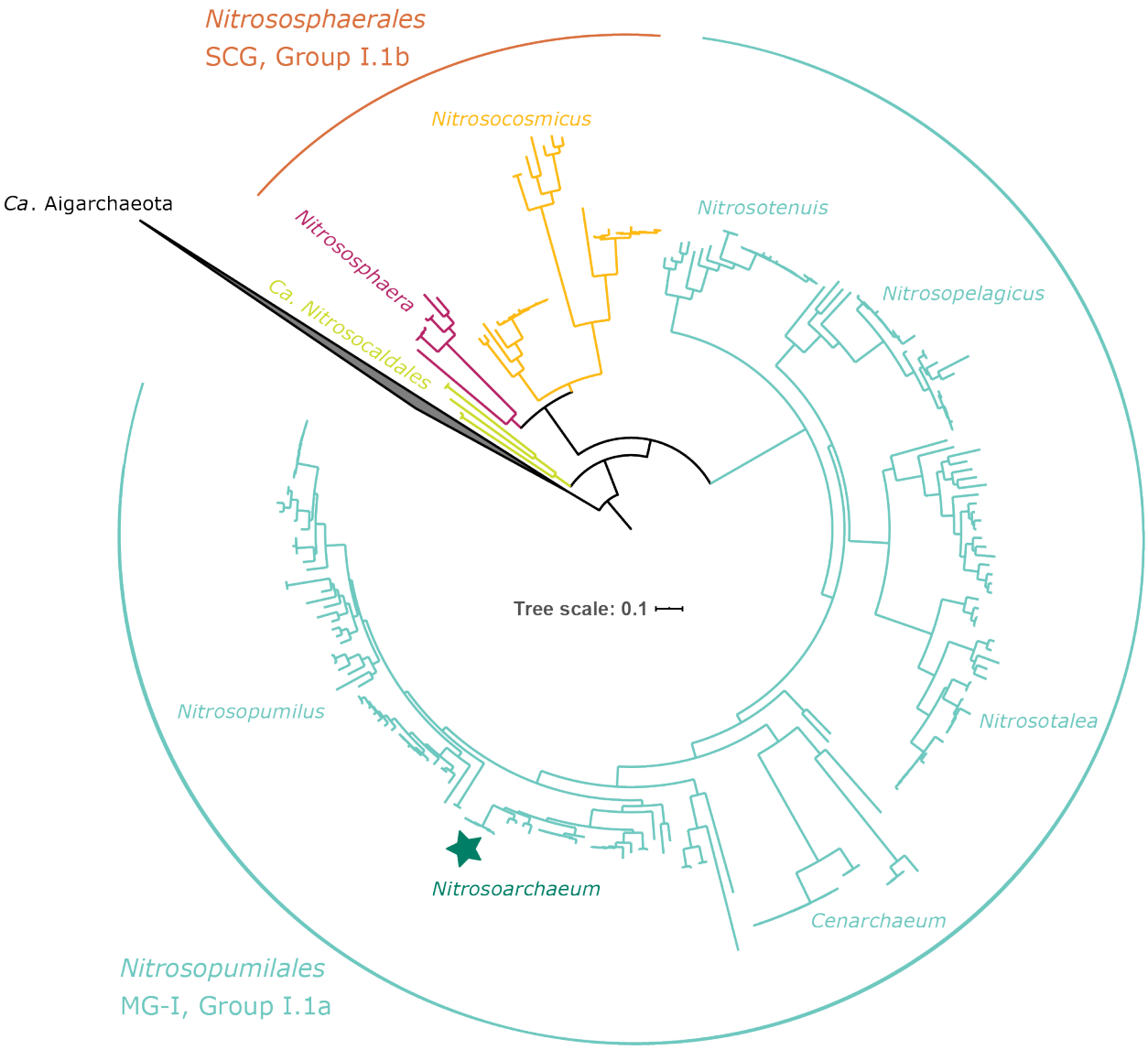


Figure S6. Phylogenomic tree of *Thaumarchaeota* and *Aigarchaeota*. Tree was constructed using concatenated single-copy core genes from (11), and all available genomes in Genbank as of February 6 2020. Colors denote clades described in (46). Star indicates placement of NarchGL.

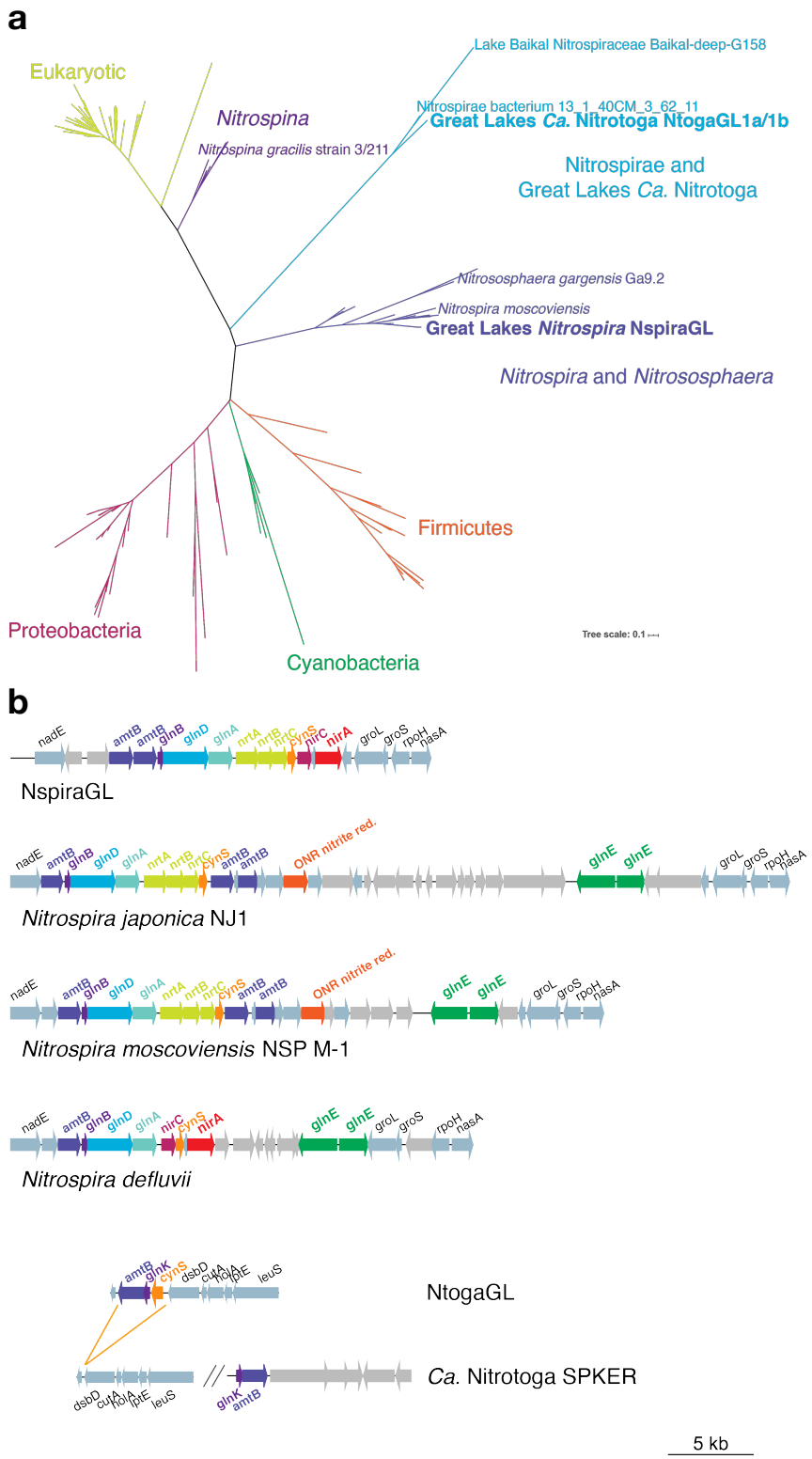


Figure S7. Cyanase diversity and distribution among nitrite oxidizing bacteria. a) Phylogenetic tree showing distinct clades of cyanase, highlighting Great Lakes *Nitrospira* and *Ca.* Nitrotoga. b) Gene synteny surrounding the cyanase gene in NspiraGL and reference *Nitrospira*, and a cyanase insertion in NtogaGL compared to reference *Ca.* Nitrotoga.

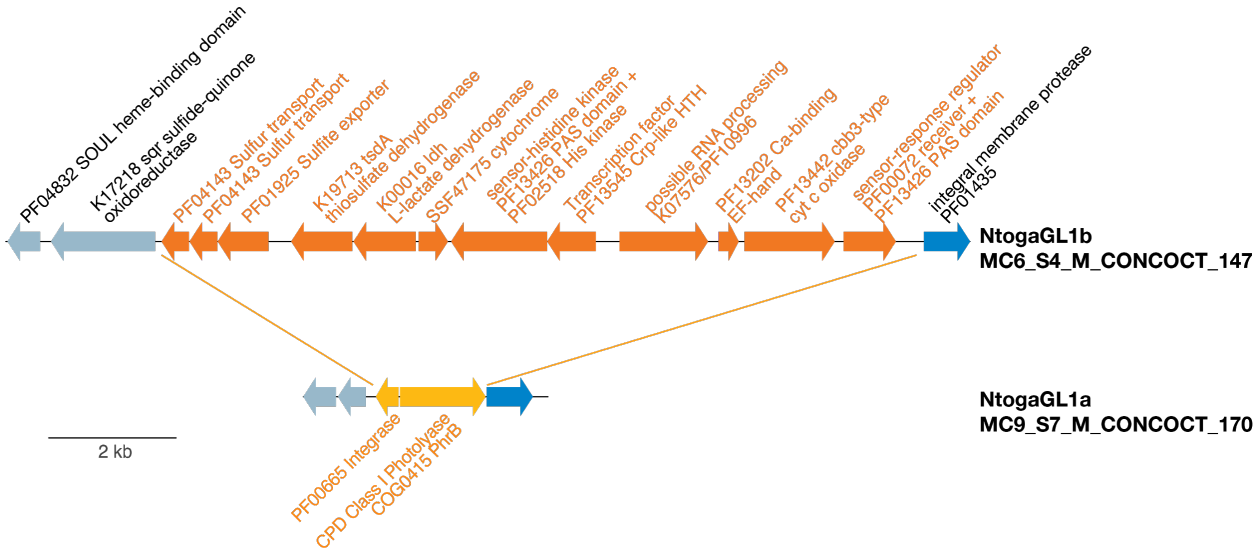


Figure S8. Gene content distinguishing NtogaGL1a and NtogaGL1b. Genes colored in blue are shared between both genomes; orange colored genes are unique to either genome.

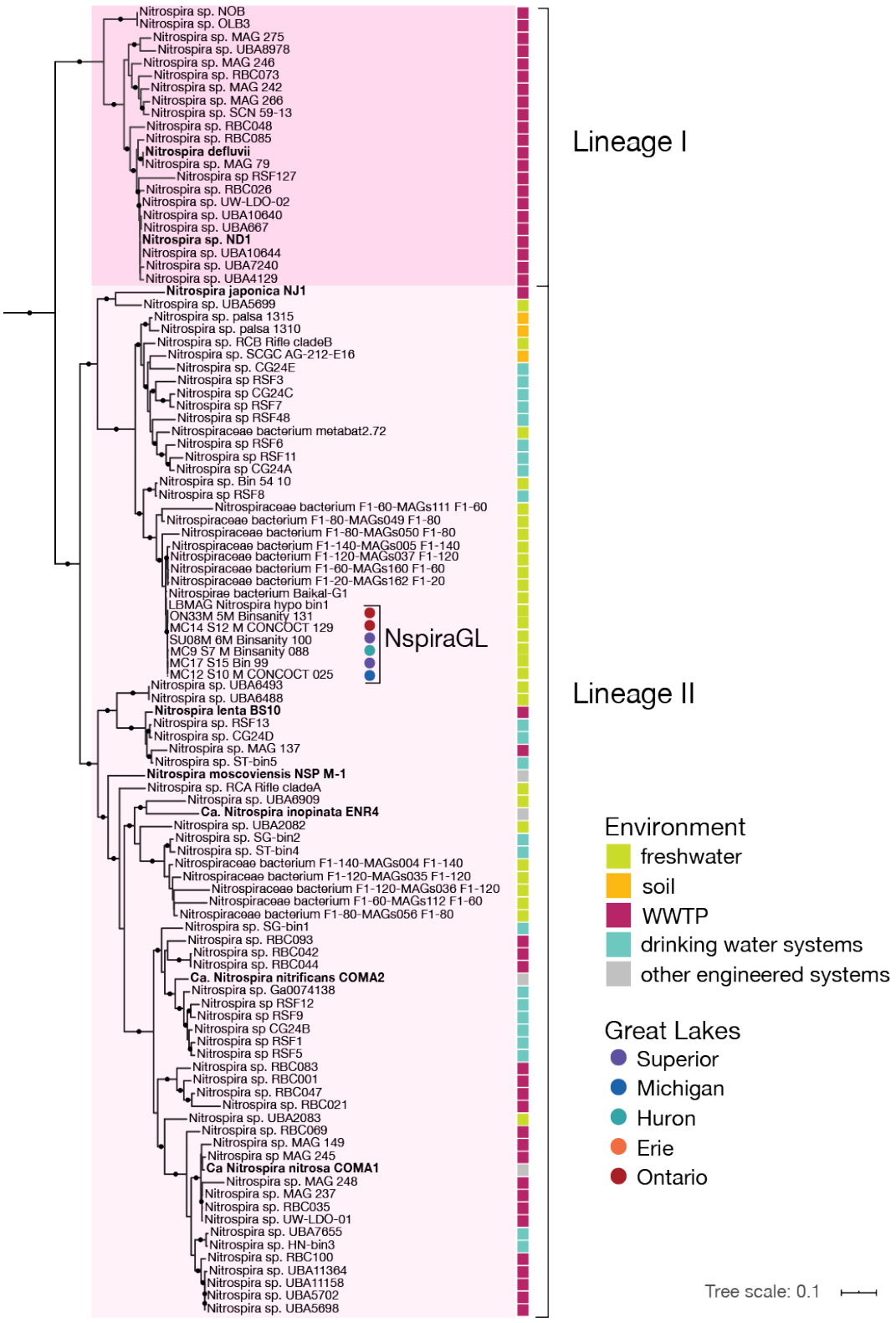


Figure S9. Phylogenomic tree of Lineage I and II *Nitrospira*. Tree was constructed using concatenated single-copy core genes described in (10). Colored boxes correspond to the environment of origin; genomes from the Laurentian Great Lakes are indicated by circles color coded by lake of origin.

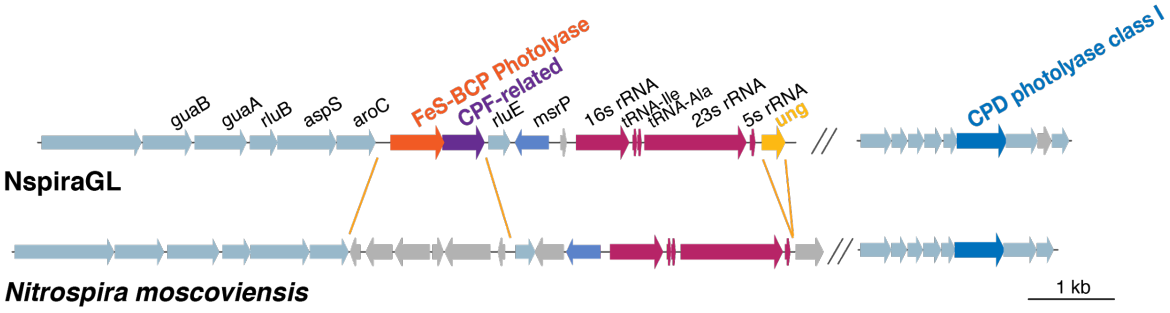


Figure S10. Gene neighborhood surrounding photolyase genes in NspiraGL, compared to reference *N. moscoviensis*. Genes shared between the two genomes are colored light blue (protein coding) or magenta (rRNA, tRNA); non-shared genes are gray; photolyase genes are colored to match the tree in Figure 5.

References

1. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017; 27:824–34.
2. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; 9:357–9.
3. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 2015; 3:e1165.
4. Graham ED, Heidelberg JF, Tully BJ. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* 2017; 5:e3035.
5. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014; 11:1144–6.
6. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018; 3:836–43.
7. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 2015; 3:e1319.
8. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; 25:1043–55.
9. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinform* 2009; 10:421.
10. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci USA* 2011; 108:12776–81.
11. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013; 499:431–7.
12. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 2013; 41:e121–e121.
13. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016; 44:D733–45.
14. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. *BMC Bioinform* 2011; 12:385.
15. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020; 36:1925-1927.
16. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018; 9:1–8.
17. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014; 30:1236–40.
18. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol Biol Evol* 2017; 34:2115–22.
19. Delmont TO, Eren AM. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* 2018; 6:e4320.
20. Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND. ITEP: An integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* 2014; 15:8.
21. van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. *Methods Mol Biol* 2012; 804:281–95.
22. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for

functional characterization of genome and metagenome sequences. *J Mol Biol* 2016; 428:726–31.
23. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014; 30:2068–9.
24. Guy L, Roat Kultima J, Andersson SGE. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 2010; 26:2334–5.
25. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform* 2010; 11:538.
26. Boeuf D, Audic S, Brillet-Guéguen L, Caron C, Jeanthon C. MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. *Database (Oxford)* 2015; bav080.
27. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002; 30:3059–66.
28. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; 30:1312–3.
29. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 2016; 44:W242–5.
30. Pachiadaki MG, Sintes E, Bergauer K, Brown JM, Record NR, Swan BK, et al. Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. *Science* 2017; 358:1046–51.
31. Spang A, Poehlein A, Offre P, Zumbrägel S, Haider S, Rychlik N, et al. The genome of the ammonia-oxidizing *Candidatus* Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations. *Environ Microbiol* 2012; 14:3122–45.
32. Palatinszky M, Herbold C, Jehmlich N, Pogoda M, Han P, von Bergen M, et al. Cyanate as an energy source for nitrifiers. *Nature* 2015;524:105–8.
33. Vechtomova YuL, Telegina TA, Kritsky MS. Evolution of proteins of the DNA photolyase/cryptochrome family. *Biochemistry Moscow* 2020; 85:131–53.
34. Maresca JA, Keffer JL, Hempel PP, Polson SW, Shevchenko O, Bhavsar J, et al. Light modulates the physiology of nonphototrophic Actinobacteria. *J Bacteriol* 2019; 201: e00740-18.
35. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020; 37:1530–4.
36. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res* 2005; 33:D34–8.
37. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; 32:1792–7.
38. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000; 17:540–52.
39. Wick RR, Schultz MB, Zobel J, Holt KE. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 2015; 31:3350–2.
40. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013; 14:178–92.
41. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S-I. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. *Environ Microbiol* 2019; 21:4740-4754.
42. Sedlacek CJ, McGowan B, Suwa Y, Sayavedra-Soto L, Laanbroek HJ, Stein LY, et al. A physiological and genomic comparison of *Nitrosomonas* cluster 6a and 7 ammonia-oxidizing bacteria. *Microb Ecol* 2019; 78:985–94.

43. Orellana LH, Rodriguez-R LM, Konstantinidis KT. ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. *Nucleic Acids Res* 2017; 45:e14.
44. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing V, Austria, 2018.
45. Wickham H. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag; 2009.
46. Alves RJE, Minh BQ, Urich T, von Haeseler A, Schleper C. Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on *amoA* genes. *Nature Commun* 2018; 9:1517.