Genome streamlining, proteorhodopsin, and organic nitrogen metabolism in freshwater 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. Freshwater nitrifiers have received little attention, leaving many unanswered questions about their taxonomic distribution, functional potential, and ecological interactions. Here, we reconstructed genomes to infer the metabolism and ecology of free-living picoplanktonic nitrifiers across the Laurentian Great Lakes, a connected series of five of Earth's largest lakes. Surprisingly, ammonia oxidizing Bacteria (AOB) related to *Nitrosospira* dominated over ammonia oxidizing Archaea (AOA) at nearly all stations, with distinct ecotypes prevailing in the transparent, oligotrophic upper lakes compared to Lakes Erie and Ontario. Unexpectedly, one ecotype of *Nitrosospira* encodes proteorhodopsin, which could enhance survival in conditions where ammonia oxidation is inhibited or substrate limited. Nitrite oxidizing Bacteria (NOB) *Ca.* Nitrotoga and *Nitrospira* fluctuated in dominance, with the latter prevailing in deeper, less productive basins. Genome reconstructions reveal highly reduced genomes and features consistent with genome streamlining, along with diverse adaptations to sunlight and oxidative stress and widespread capacity for organic nitrogen use. Our findings expand the known functional diversity of nitrifiers and establish their ecological genomics in large lake ecosystems. By elucidating links between microbial biodiversity and biogeochemical cycling, our work also informs ecosystem models of the Laurentian Great Lakes, a critical freshwater resource experiencing rapid environmental change.

Importance (150 words)

Microorganisms play critical roles in Earth's nitrogen cycle. In lakes, microorganisms called nitrifiers derive energy from reduced nitrogen compounds. In doing so, they transform nitrogen into a form that can ultimately be lost to the atmosphere by a process called denitrification, which helps mitigate nitrogen pollution from fertilizer runoff and sewage. Despite their importance, freshwater nitrifiers are virtually unexplored. To understand their diversity and

function, we reconstructed genomes of freshwater nitrifiers across some of Earth's largest freshwater lakes, the Laurentian Great Lakes. We discovered several new species of nitrifiers specialized for clear low nutrient waters, and distinct species in comparatively turbid Lake Erie. Surprisingly, one species may be able to harness light energy using a protein called proteorhodopsin, despite the fact that nitrifiers typically live in deep dark water. Our work reveals unique biodiversity of the Great Lakes and fills key gaps in our knowledge of an important microbial group, the nitrifiers.

Introduction

The oxidation of ammonia to nitrate powers the growth of nitrifying microorganisms and represents a critical flux in the global nitrogen cycle. Microbial nitrification of ammonia released from organic matter degradation 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 are found in Archaea and several phyla of Bacteria, spanning diverse physiology and ecology; 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), likely due to their high affinity for ammonia (8) and streamlined genomes (9). In freshwater systems, AOA are abundant in some oligotrophic lakes, while ammonia-oxidizing Bacteria (AOB) affiliated with the *Nitrosomonadaceae* (Betaproteobacteria) are highly variable but tend to dominate more eutrophic systems (10–16). Complicating this picture, however, there is considerable physiological variation within each group, such as low-nutrient-adapted clades of AOB (17, 18) and the ability of some strains to use alternative substrates like urea (18, 19). Within the AOA, there are also distinct ecotypes that appear to segregate with depth in the water column, in both marine (7) and freshwater systems (10). In freshwaters especially — which are poorly characterized compared to the oceans — it remains difficult to predict which AOA and AOB taxa are likely to dominate in a given system (16).

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, 20), consistent with ancient salinity-associated divergence. Cultivated strains of NOB show variation in substrate affinity and physiology (20–22), but the phylogenetic conservation of these traits, and their influence on environmental distributions, are

poorly understood. Moreover, recent studies have discovered that NOB are capable of alternative energy metabolisms (23, 24) and can access nitrogen from cyanate and urea (25, 26), expanding their ecological potential. In freshwater systems, the NOB *Ca.* Nitrotoga (Betaproteobacteria) was only recently discovered to be widespread (27), and the diversity of this genus and factors favoring its success are unknown.

Here, we use the Laurentian Great Lakes as a model system to examine niche partitioning among planktonic freshwater nitrifiers. The Great Lakes hold 20% of Earth's surface freshwater and more than half of this volume receives little to no light (< 1% surface irradiance). This system, while hydrologically connected, spans strong trophic and chemical gradients: ultraoligotrophic Lake Superior supports rates of primary production and nitrification comparable to the ocean gyres (28, 29), while Lake Erie supports greater production (30) and more than 70-fold higher nitrification rate (31). Between these extremes, Lake Ontario has low ambient ammonium concentrations like Lake Superior (32) but nitrification rates up to four times higher (33). While previous studies reported that AOA and AOB dominate Lakes Superior and Erie, respectively (14, 29), recent community profiling has revealed broader diversity in both ammonia-oxidizing and nitrite-oxidizing lineages (34–36). We sought to link taxonomic, genomic, and metabolic diversity of nitrifiers with the varied biogeochemistry of the Great Lakes, using genome reconstructions and abundance profiling. Our results uncover novel lineages and metabolic capabilities, and provide the first large-scale assessment of freshwater nitrifier genomics.

Results and Discussion

Niche partitioning of nitrifiers across the Great Lakes

To map free-living picoplanktonic (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 (34). We detected putative AOB in the genus *Nitrosospira* (Betaproteobacteria, family *Nitrosomonadaceae*) and AOA similar to *Nitrosarchaeum* (family *Nitrosopumilaceae*), along with putative NOB in the genera *Ca.* Nitrotoga (Betaproteobacteria, family *Gallionellaceae*) and *Nitrospira* (family *Nitrospiraceae*). We did not detect 16S rRNA amplicons from *Nitrosococcus*, *Nitrococcus*, *Nitrospina* or *Nitrobacter*. The highest relative abundances of picoplanktonic nitrifiers were observed in deep samples from eastern Erie and Ontario (9-24% of total amplicons), compared to 2-14% in Michigan, Huron and Superior; Erie and Ontario also have higher cell concentrations and higher surface chlorophyll (Dataset S1). 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 (37–40), as well as potential competition with phototrophs for ammonium (41).

The taxonomic assemblage of nitrifiers differed across lakes and even among stations within a lake (Fig 1, Dataset S1), in association with variable productivity and nitrogen availability. Surface ammonium is typically below 300 nM except in Erie, where it is several-fold higher and spatially variable; nitrate, on the other hand, is very high across the lakes but lowest in Erie due to biological uptake (42, 43). Few measurements of urea exist but it can exceed ammonium (44) (Dataset S2). AOB (*Nitrosomonadaceae*) were observed across all lakes. In contrast, AOA (*Nitrosopumilaceae*) 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. We found pronounced shifts in the dominant NOB across stations (Fig 1), and all stations except those in Lake Ontario showed strong dominance (greater than ten-fold) of either *Ca.* Nitrotoga (family *Gallionellaceae*) or *Nitrospira*. *Nitrospira* was the only nitrite oxidizer detected in Superior and the dominant nitrite oxidizer in parts of Michigan (MI41M, MI18M). By contrast, *Ca.* Nitrotoga was the only nitrite oxidizer observed in Erie and Huron and the dominant nitrite oxidizer at the shallowest station in Michigan (MI27M). Within each taxon, a single 16S rRNA oligotype dominated the AOA, *Ca*. Nitrotoga, and *Nitrospira,* while several oligotypes of *Nitrosomonadaceae* shifted abundance across samples (Fig S2), consistent with ecotypic diversity as discussed below.

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

We reconstructed 15 genomes of the AOB *Nitrosospira*, substantially expanding genome descriptions for this genus (45–47). Based on a phylogenomic tree, free-living Great Lakes *Nitrosospira* fall into two major clades, both of which are distinct from published species; each of these clades also includes MAGs recovered from lakes Biwa and Baikal, suggesting novel globally distributed freshwater lineages (Fig S3). One clade, which we call NspGL1, has a highly reduced genome (median 1.42 Mb) and low G+C content (40.7%) (Fig 2, Dataset S3). The second clade was resolved into three subclades (denoted NspGL2a, 2b, and 3; Fig S3) based on phylogeny and average nucleotide identity (ANI), all with small genome sizes of 1.45-1.68 Mb and 50% G+C content (Fig 2, Dataset S3). Compared to 86 reference *Nitrosomonadaceae* genomes, Great Lakes *Nitrosospira* genomes are not only smaller (median estimated complete genome size: reference = 3.21 Mb, GL = 1.45 Mb; Table 1), but also have shorter intergenic spacers, fewer paralogs, fewer pseudogenes, and fewer sigma factors (Table 1, Fig S4, Dataset S4), consistent with genome streamlining to reduce resource demands (48). Based on short read mapping, these subclades are ecologically distinct: NspGL1 and NspGL2b — with the

smallest genomes — are the dominant AOB in the upper oligotrophic lakes, while NspGL2a is only abundant in Ontario and NspGL3 is only abundant in Erie (Fig 2). Hereafter we refer to these subclades as ecotypes due to their phylogenetic and ecological divergence.

We next compared gene content between our Great Lakes *Nitrosospira* and 86 *Nitrosomonadaceae* reference genomes. On average, Great Lakes *Nitrosospira* encode far fewer two-component signal transduction systems (NspGL = 5-8, mean reference = 19), transposases (NspGL = 0-7, mean reference = 39), motility genes (NspGL = 0-4, mean reference = 52), pilus and secretion genes (NspGL = 2-9, mean reference = 27), and defense-related genes (NspGL = 4-11, mean reference = 39) (Dataset S5). They also lack functions related to biofilm formation such as polysaccharide matrix production (e.g. *pel* genes) and extracellular protein targeting (exosortase and PEP-CTERM motifs). We note that while our 15 new *Nitrosospira* MAGs have high estimated completion (median 98.6%; Dataset S3), some gene absences may reflect incomplete assemblies. Nevertheless, this overall picture of gene content in Great Lakes AOB contrasts with *Nitrosospira* isolates from soil (45, 47) and even to oligotrophic *Nitrosomonas* isolates (49), and is consistent with a passive planktonic lifestyle in extremely low-nutrient systems.

We next compared metabolic potential among Great Lakes AOB ecotypes to understand their ecological preferences for upper lakes (NspGL1, NspGL2b), Ontario (NspGL2a), and Erie (NspGL3). Surprisingly, all seven NspGL1 MAGs encode proteorhodopsin, a light-driven proton pump that supports bacterial energy production (50, 51). They also carry the genes necessary to synthesize its chromophore retinal, including 15,15'-Beta-carotene dioxygenase (*blh*), lycopene cyclase (*crtY*), phytoene dehydrogenase (*crtI*), phytoene synthase (*crtB*), and GGPP synthase (*crtE*) ((52, 53); Fig 3a). We also identified proteorhodopsin in a single-cell amplified genome representing NspGL1 from Lake Michigan (Fig 3a; 99.8% ANI with NspGL1 MAGs), demonstrating that it is not an artifact of metagenome assembly. To our knowledge, this is the first example of a nitrifier with proteorhodopsin. All NspGL1 proteorhodopsins share residues

H95, D127 and E138 along with a short beta-turn (G111-P116) between helices B and C, which are characteristic features of proteorhodopsin as distinct from sensory and other rhodopsins (54), and the presence of leucine at position 135 suggests green light tuning (55) (Fig 3b). All of the genes in this module have highest similarity to homologs from *Polynucleobacter*, but are flanked by Nitrosomonadaceae-like genes, suggesting recent horizontal gene transfer (Fig 3a). The predicted NspGL1 proteorhodopsins cluster with *Polynucleobacter*, *Methylopumilus*, and other freshwater Betaproteobacteria in Supercluster III as defined by MicRhoDE ((56); Fig 3c). We compared the homologous genome region in two highly similar MAGs from Lakes Biwa and Baikal (Fig S5); these contigs lack the proteorhodopsin module but appear to flank a variable region where the contig assembly ends. A proteorhodopsin photosystem could support survival of NspGL1 in the presence of sunlight, which has been shown to inhibit ammonia oxidation (37, 57). In the upper lakes where NspGL1 is abundant, light penetration is high well below the thermocline in stratified periods (58), and deep water taxa are seasonally advected to the surface by water column mixing (34). In addition to proteorhodopsin, NspGL1 — but not the other three ecotypes of Great Lakes *Nitrosospira* — encode a class I cyclopyrimidine dimer photolyase, which uses light energy to repair UV-induced DNA damage, and the catalase-peroxidase *katG*, suggesting that the NspGL1 ecotype is adapted to relatively shallow depths in the water column (Fig 2).

Great Lakes *Nitrosospira* carry a reduced, ecotype-specific complement of nitrogen metabolism genes compared to reference AOB (Fig 2, Dataset S5; gene absences were verified as described in Materials & Methods). All are presumed to have the core ammonia oxidation enzymes ammonia monooxygenase and hydroxylamine dehydrogenase; these genes were assembled and binned as expected in some MAGs, and were manually identified on short unbinned contigs in other cases (Dataset S6; see Materials & Methods). Surprisingly, all Great Lakes *Nitrosospira* MAGs lack the copper protein nitrosocyanin, whose precise function is unknown but so far has been found in all described AOB except one member of the *N.*

oligotropha clade (49). Based on the expanded set of genomes analyzed here, the lack of nitrosocyanin likely extends beyond the Great Lakes MAGs to closely related freshwater and marine strains, along with five more members of the *N. oligotropha* clade (Fig S3); its absence may be related to the divergence of these clades. Only NspGL1 and NspGL2b encode NO-forming nitrite reductase (NirK), which confers nitrite tolerance (59); this result is surprising given that these two clades dominate the upper lakes where productivity and reduced N are lowest. None of the Great Lakes ecotypes encode NO reductase (NorCBQD) and NspGL1 lacks cytochrome P460 family proteins, both of which are common in AOB and implicated in nitrogen oxide metabolism (18, 49). Nitrogen acquisition is also distinct among Great Lakes AOB: NspGL1 lacks an apparent ammonium transporter, but encodes urease structural and accessory genes (*ureABCEFG*) and a high-affinity urea transporter (*urtABCDE)*. Further, all Great Lakes ecotypes encode a high-affinity amino acid transporter (*livFGHM*); these genes are rare (<5%) in reference genomes and could supply reduced nitrogen and/or organic carbon. Finally, NspGL1 and NspGL3 have genes for producing cyanophycin, an intracellular storage compound for nitrogen (47, 60). Together, the distinctive gene complements present in Great Lakes *Nitrosospira* illustrate the variability and adaptability of AOB gene content, even across a connected freshwater habitat.

As with nitrogen metabolism, carbon metabolism is also distinct between Great Lakes and reference AOB, and among Great Lakes ecotypes (Fig 2, Dataset S5). Unlike most reference AOB, Great Lakes *Nitrosospira* lack two key enzymes of the oxidative pentose phosphate pathway, glucose-6-phosphate dehydrogenase and 6-phosphogluconate dehydrogenase. All ecotypes except Erie-specific NspGL3 also lack genes for glycogen synthesis and degradation, suggesting that they are unable to store and access this carbon reserve. The key enzyme for carbon fixation, RuBisCO, has evolved several kinetically distinct forms whose distribution likely reflects ecological pressures (61). NspGL1 and NspGL3 both contain Form IA RuBisCO, while NspGL2a and NspGL2b contain Form IC RuBisCO (Fig 2; (61,

62)). NspGL1 genomes also possess an alpha carboxysome-like *cso* operon, similar to *Nitrosomonas eutropha* C91 (62), though our draft assembly lacks the expected carbonic anhydrase (*csoS3*/*csoSCA*). Carboxysome-associated RuBisCO may allow NspGL1, the ecotype most strongly adapted to energy and nutrient limitation, to more efficiently fix CO_2 by minimizing the wasteful oxygenation reaction and reducing the cellular nitrogen allocation to RuBisCO (61). The ranges of kinetic properties observed in other autotrophs for Form IAq (found in NspGL3) and Form IC (found in NspGL2a and NspGL2b) overlap, and therefore more work is needed to understand the fitness advantages, if any, that this RuBisCO diversity confers on Great Lakes nitrifiers.

***Streamlined freshwater* Thaumarchaeota**

We reconstructed three similar genomes (>99% ANI) of *Nitrosarchaeum* (NarchGL; Fig S6, Dataset S3) from three separate samples (two from Superior, one from Ontario), consistent with our low observed 16S rRNA diversity for *Thaumarchaeota*. These NarchGL genomes are very similar (~99% ANI) to two genomes reconstructed from Lake Baikal, located thousands of kilometers away (63). Their next closest relatives are also from freshwater environments, and phylogenetic clustering suggests that salinity is an important driver of divergence throughout the *Nitrosopumilaceae* (Fig S6). As a group, the *Thaumarchaeota* tend to have smaller genomes, lower G+C content, higher coding density, and fewer paralogs and pseudogenes than nitrifying bacterial taxa; even within this group, NarchGL genomes fall below the 30th percentile in size and have significantly fewer paralogs than average (Table1, Fig S4, Dataset S4). Using our reconstructed genomes as probes for metagenomic read recruitment, NarchGL were detected in Superior, Michigan, and Ontario; they represented roughly one-third of ammonia oxidizers in the mid-hypolimnion of station SU08M (Fig 2).

NarchGL share nearly 90% of their predicted proteins with close relatives including *Ca.* Nitrosarchaeum limnia; at the same time, they show distinctive patterns in gene content that

pinpoint the key selective pressures of deep lakes (Fig 2, Dataset S5). All three NarchGL genomes encode urease and a urea transporter, implicating urea as a vital source of nitrogen for energy and/or biosynthesis. Consistent with phosphorus scarcity in much of the Great Lakes (64), NarchGL encode high affinity transport systems for phosphate and potentially phosphonates, though we did not identify a phosphonate lyase. In addition to both CRISPR/Cas enzymes cas1 and cas4, NarchGL genomes contain several phage proteins, suggesting viral infection and integration events may be common. DNA photolyases, which have been found in epipelagic clades of marine *Thaumarchaeota* (7), are present in all low salinity *Nitrosarchaeum* including NarchGL, suggesting NarchGL are exposed to sunlight due to high water clarity (58) and/or annual mixing in the Great Lakes. NarchGL also lack the common tRNA modification 4-thiouridylation (indicated by K04487 and PF02568-PF18297 (65)); we propose that the absence of this modification, which is susceptible to near-UV radiation (65), is also related to sunlight exposure.

Genomes of NarchGL reveal striking reduction in environmental sensing, response, and regulatory functions, relative to most other *Nitrosarchaeum* and *Nitrosopumilaceae* (Dataset S5). NarchGL encode 9-12 domains representing the general archaeal transcription factors, TATA binding protein (TBPs, PF00352) and transcription factor B (TFBs, PF00382 and PF08271), compared to 21 in *Ca.* N. limnia. NarchGL lack common domains found in two-component systems that transmit environmental signals to control gene expression or protein activity (domains PF02743, PF00672, PF00512, PF00072; NarchGL = 0, *Ca.* N. limnia = 53-54 copies per genome). Further, they are depleted in ArsR family transcription factors (PF01022; 0 copies in NarchGL vs. 2-3 in *Ca.* N. limnia), P-II proteins for regulation of nitrogen metabolism (PF00543; 1 copy per NarchGL genome vs. 5 in *Ca.* N. limnia), and other potential regulatory domains (CBS PF00571: 5 in NarchGL vs. 18-19 in *Ca.* N. limnia; USP PF00582: 1 in NarchGL vs. 15 in *Ca.* N. limnia). This extremely limited regulatory capacity in NarchGL stands in sharp

contrast to closely related *Ca.* N. limnia, but instead parallels the oceanic minimalist *Ca.* Nitrosopelagicus brevis (9).

Expanded diversity of Ca.* Nitrotoga *with reduced genomes

Despite the broad distribution of *Ca.* Nitrotoga in freshwater systems and beyond, only six genomes are available, derived from rivers heavily impacted by urban and agricultural influence, a wastewater treatment plant, and coastal sediment (27, 66, 67). Hence the metabolic and phylogenetic diversity of this group is virtually unexplored. We reconstructed six new MAGs of *Ca*. Nitrotoga, which form two clusters with >99% ANI within each cluster and ~97% ANI between clusters (NtogaGL1a and NtogaGL1b; Fig S3). These new *Ca*. Nitrotoga MAGs are far smaller than published genomes (median GL = 1.44 Mb, reference = 2.61-2.98 Mb), have shorter intergenic regions, fewer sigma factors, and fewer paralogs (Table 1, Fig S4, Dataset S4), consistent with genome streamlining (48). They lack functions such as motility and chemotaxis, pilus biogenesis, and DNA repair (*mutLS*) (Dataset S5). Great Lakes *Ca.* Nitrotoga also encode markedly fewer two-component systems for sensing and responding to environmental cues than four river-derived genomes (NtogaGL = 2-6 per genome vs. 30-35 in four reference genomes; reference strain KNB has 7). Compared to reference genomes, NtogaGL have fewer defense-related genes (restriction-modification, toxin-antitoxin, and CRISPR-Cas systems; mean NtogaGL = 11 vs. 39 for references), and transposases (mean NtogaGL = 3 vs. 19 for references) (Dataset S5). While incomplete assembly of hypervariable genome regions may explain some of these absences, the overall genome properties are consistent with a relatively stable low-nutrient environment and planktonic lifestyle.

The reduced genomes of NtogaGL1a/b help clarify core features of the genus *Nitrotoga*, along with accessory functions that may enable local adaptation in specific populations. To date, sequenced *Nitrotoga* including NtogaGL1a/b encode similar electron transport pathways, including NADH dehydrogenase-Complex I, succinate dehydrogenase-Complex II, and

Alternative Complex III, along with high-affinity cbb3-type cytochrome oxidases suggesting adaptation to low oxygen conditions. They also share the Calvin cycle for carbon fixation, a complete TCA cycle, and an evolutionarily distinct nitrite oxidoreductase (NXR) from other NOB (27, 66, 67). All *Nitrotoga* to date also share transporters for amino acids and peptides, potential sources of C and/or N. *Nitrotoga* can also potentially access reduced sulfur compounds for energy via sulfite dehydrogenase, suggesting metabolic flexibility beyond nitrite oxidation.

Beyond these similarities, the small genomes of NtogaGL1a/b are distinct from previously described *Nitrotoga* in many ways. NtogaGL1a/b lack NiFe hydrogenase to use hydrogen as an energy source. They also lack nitrogen metabolism functions including assimilatory nitrite reductase (*nirBD*) and nitrite reductase to NO (*nirK*). Based on gene content, NtogaGL1a/b appear unable to use hexoses like glucose, since they lack the glycolytic enzyme phosphofructokinase and the Entner-Doudoroff pathway, similar to *Nitrobacter winogradskyi* (68). Consistent with this, they also lack genes for storage and breakdown of glycogen (Dataset S5). All but one of the NtogaGL1a/1b genomes encode cyanate lyase (*cynS*), which is found in other NOB but no *Nitrotoga* to date (25, 69, 70). The *cynS* gene, adjacent to *glnK-amtB* for ammonium sensing and transport, likely functions in N assimilation, as recently described for *Nitrospinae* (71). While cyanase has been shown to mediate reciprocal feeding between some NOB and ammonia oxidizers (25), it remains to be seen whether such an interaction occurs in the free-living (<1.6um) size-fraction and dilute environment sampled here. Notably, cyanase from NtogaGL1a/1b, along with predicted *Nitrospirae* proteins from Lake Baikal and soil, form a distinct phylogenetic cluster from most nitrifier cyanase proteins observed to date (Fig S7).

The two ANI-based clusters we detected, NtogaGL1a and NtogaGL1b, appear to be phylogenetically and ecologically distinct ecotypes. Based on short-read mapping, NtogaGL1b dominates Erie, while NtogaGL1a dominates all other *Ca.* Nitrotoga-containing samples (Fig 4). We found several metabolic genes that differentiate the two ecotypes. ER-specific NtogaGL1b genomes share a region encoding thiosulfate dehydrogenase (*tsdA*), cytochromes, transport of

sulfur-containing compounds, lactate dehydrogenase (ldh*),* a two-component system, and a Crp-family transcription factor (Fig S8). This region may be involved in oxidizing thiosulfate as an energy source, and sensing and responding to redox changes that accompany seasonal hypoxia in Lake Erie. The corresponding region in NtogaGL1a encodes an integrase and photolyase, consistent with greater DNA photodamage in the more transparent waters of Michigan, Huron, and Ontario where NtogaGL1a is abundant.

Great Lakes* Nitrospira *reveal adaptations to sunlit oxic environment

We reconstructed six closely related genomes of *Nitrospira* (~99% ANI; Fig S9, Dataset S3), representing the predominant NOB throughout Lake Superior and in parts of Michigan, Ontario, and Huron (Fig. 4; Dataset S1). These genomes, which we refer to as NspiraGL, fall within lineage II (Fig S9), which is broadly distributed across soil, freshwater, and engineered habitats (20); however, genome analyses to date have focused on strains from wastewater and engineered systems, leaving major blind spots. NspiraGL share core features of *Nitrospira* metabolism, including a periplasmic-facing NXR that is advantageous under substrate-limiting conditions, multiple cytochrome *bd*-like oxidases, and the reverse TCA cycle for carbon fixation (69). However, as with *Ca.* Nitrotoga, the *Nitrospira* genomes we reconstructed in the Great Lakes are markedly smaller than published reference genomes (median NspiraGL = 1.83 Mb, median reference = 3.72 Mb), with higher coding density and fewer paralogs, sigma factors and pseudogenes (Fig. S4; Dataset S4), consistent with genome streamlining theory (48). Compared to 75 lineage II *Nitrospira* reference genomes, NspiraGL have reduced capacity for environmental sensing (two-component systems: NspiraGL = 7, mean reference = 26), transport (NspiraGL = 76-83, mean reference = 140), defense (NspiraGL = 7-8, mean reference = 26), and transposition (NspiraGL = 0-2, mean reference = 15), and lack pilus or flagellar motility (Dataset S5). NspiraGL encode just five sigma factors, compared to 18 in *N. moscoviensis.* Further, NspiraGL genomes encode a single NXR, while *N. moscoviensis* carries five copies

that are differentially regulated (26, 72). NspiraGL also lack the *glnE* gene for glutamine synthetase (GS) adenylyltransferase, suggesting that GS activity is not repressed by this mechanism. Together, these features suggest limited regulatory and ecological flexibility, consistent with a relatively constant, oligotrophic environment.

Compared to other *Nitrospira,* NspiraGL exhibit limited energetic flexibility, but can access diverse nitrogen sources (Fig. 4, Dataset S5). We predict that NspiraGL are unable to grow on hydrogen or formate as alternative energy sources (23, 26), as they lack NiFe-hydogenase and formate dehydrogenase. The glycolysis and oxidative TCA cycles appear to be incomplete, lacking phosphofructokinase and citrate synthase, respectively; this suggests a limited capacity for organic carbon utilization. NspiraGL lack *nirK,* encoding NO-forming nitrite reductase, which is found in a majority of reference genomes. To obtain nitrogen for biosynthesis, NspiraGL encode a high-affinity nitrate/nitrite/cyanate transporter (*nrtABC*), assimilatory nitrite reductase (*nirA*), and cyanase (*cynS*), along with *amt* family ammonium transporter. Although none of the NspiraGL MAGs include urease (*ureCBA*), one does contain urease accessory proteins (*ureEFGD*) and two contain a urea transporter (*urtABCD*), suggesting incomplete assembly of the urea utilization pathway. As with *Ca.* Nitrotoga, we suggest that cyanase, along with urease where present, functions in nitrogen assimilation rather than cross-feeding, given the dilute environment and free-living planktonic cells.

Beyond energy, carbon, and nitrogen metabolism, we discovered striking differences between NspiraGL and reference *Nitrospira* related to DNA repair. NspiraGL encode two additional photolyase-related proteins, along with a class I cyclopyrimidine dimer (CPD) photolyase found in most reference *Nitrospira* (Fig 5). Photolyases use blue light energy to repair DNA lesions caused by UV radiation (73). The two additional genes in NspiraGL are adjacent and share best hits with *Betaproteobacteria*, suggesting recent horizontal transfer (Fig S10). One likely encodes an FeS-BCP photolyase, which repairs (6-4) dipyrimidine lesions (74, 75). The other shares an FAD-binding domain with photolyases but the C-terminal region has

no recognizable domains (Fig 5). This protein is widespread in aquatic bacteria and has not been functionally characterized, though an actinobacterial homolog was suggested to be involved in light sensing and regulation (76). Beyond photolyases, NspiraGL also encode uracil-DNA glycosylase (UNG), which removes misincorporated uracil from DNA. Uracil results from deamination of cytosine, which can occur spontaneously or be induced by NO (77). In addition to the photolyases and UNG that repair DNA lesions, NspiraGL encode translesion DNA polymerase V (*umuCD*) which enables replication to proceed past lesions. Together, these genes indicate that *Nitrospira* in the Great Lakes experience significant DNA damage, including UV-induced damage that also requires light for the repair process, in hypolimnion waters with high transparency (58) and/or during seasonal mixing.

Other major differences between NspiraGL and reference *Nitrospira* are related to reactive oxygen species (ROS) (Dataset S5). Surprisingly, despite their oxic habitat, NspiraGL lack superoxide dismutase (SOD), monofunctional catalase (*katE*), and bacterioferritin, which limits the Fenton reaction by sequestering free iron. However, all six NspiraGL MAGs, but few reference genomes (7% of 75), have recently acquired bifunctional catalase-peroxidase *katG*; interestingly we also observed *katG* in Great Lakes *Ca.* Nitrotoga and *Nitrosospira* (Fig 2, Fig 4). The absence of SOD suggests that NspiraGL does not produce damaging levels of endogenous superoxide, perhaps because NspiraGL lack the major respiratory and non-respiratory flavoproteins that produce ROS in other SOD-containing *Nitrospira* (78). Unlike superoxide, H_2O_2 can cross membranes, and is known to be produced by both photooxidation of dissolved organic matter and dark heterotrophic activity (79). The lakes where NspiraGL dominate have high water clarity (58) and low productivity and are fully oxic, consistent with abiotic photochemistry as the primary source of exogenous ROS; this stress may have selected for *katG* as a defense. NspiraGL also lack cytochrome *c* peroxidase, which is found in 70 of 75 reference genomes; this protein is proposed to function in anaerobic respiration of H_2O_2 (80) and therefore its absence in NspiraGL is consistent with a constant oxic environment. Together,

these results indicate that *Nitrospira* in the Great Lakes face distinct ROS pressures that have shaped their gene content.

Conclusions

The Laurentian Great Lakes harbor nitrifiers that are phylogenetically related, but markedly different in genome size and functional capacity, from their well-studied relatives inhabiting wastewater systems, soils, and even other freshwater systems. By examining the entire nitrifier assemblage at once, we detected common features across taxa that illuminate the selective pressures faced by microbes in deep lakes. All the lineages we describe show small genome sizes (1.3-1.7 Mb), reduced capacity for environmental sensing and response, and adaptation to a passive (i.e. non-motile) planktonic lifestyle, features which have not been previously associated with AOB, *Nitrospira*, and *Ca.* Nitrotoga. Within the AOB *Nitrosospira*, we found ecotypes with a gradient of genome reduction that maps onto their habitats' trophic gradient: from NspGL1 (1.4 Mb, low GC, upper lakes) to NspGL2b (1.5 Mb, upper lakes) to NspGL2a (1.6 Mb, Ontario) to NspGL3 (1.7 Mb, Erie) (Fig 2). The thaumarchaeal NarchGL have markedly reduced regulatory capacity like the open ocean strain *Nitrosopelagicus brevis* (9). The NOB NspiraGL have genomes 50-60% smaller than described *Nitrospira* and dominate, along with the AOB NspGL1 and the AOA NarchGL, the deeper more oligotrophic basins, while *Ca.* Nitrotoga favor shallower, more productive basins. The emergence of Erie-specific ecotypes of both *Nitrosospira* (NspGL3) and *Ca.* Nitrotoga (NtogaGL1b) demonstrates how distinct this habitat is compared to the other lakes. Importantly, our findings here represent planktonic cells in the smallest size fraction (<1.6 µm); it is likely, especially in Erie, that particle-associated nitrifiers may be abundant and genetically distinct.

Nitrifiers inhabiting the transparent waters of the upper Great Lakes show distinctive adaptations to light including diverse photolyases, ROS detoxification, and even proteorhodopsin. This discovery is surprising, given that nitrifiers are rare in the surface mixed

layer of the Great Lakes (Fig. S1) and that photoinhibition of ammonia oxidation and nitrifier growth is well documented (37, 40, 57). We propose that proteorhodopsin could be used to augment energy metabolism when ammonia oxidation is photoinhibited and/or ammonia is substrate limited. Water clarity has increased over the past several decades in Lakes Michigan and Huron, now surpassing that of Lake Superior (58). High light penetration along with seasonal mixing likely exposes deep water cells to damaging levels of light and oxidative stress. Future cultivation and physiological studies should examine photoinhibition and potential phototrophy in Great Lakes nitriifiers.

The capacity for nitrification is found across multiple phyla, and our work unveils new clues to understanding the ecological and evolutionary potential of these diverse lineages. This collective nitrifier diversity undoubtedly influences the cycling of carbon and nitrogen across this ecosystem, and future work will explore the differential contributions to nitrification by the distinct lineages we described here. Understanding what controls the diversity of nitrifiers and other key functional groups, and the consequences of this diversity for biogeochemistry, are essential for forecasting the effects of rapid environmental change across the large lakes of the world (e.g.(81)) and predicting impacts on the critical ecosystem services they provide (82).

Materials & Methods

Sample collection

Water samples were collected from the Laurentian Great Lakes aboard the R/V *Lake Guardian*, during the biannual Water Quality Surveys conducted by the U.S. EPA Great Lakes National Program Office (83). Station information is provided in Supplemental Dataset S8. Data presented here were collected in April and August 2012. Samples were collected using a CTD rosette sampler (Sea-Bird Scientific) at the surface (2 m), deep chlorophyll maximum (if present), the mid-hypolimnion (depths ranging from 19 m in Erie to 200 m in Superior; Dataset

S1), and near the bottom of the water column (10 m above the lake bottom at most stations, 1 m above bottom at shallow stations). For each sample, 5-8L of water was pre-filtered through a GF/A glass fiber filter (Whatman 1820-047; nominal pore size 1.6 µm) to exclude eukaryotic phytoplankton and particle associated microbes, and cells were collected on 0.22 µm Sterivex filters (Millipore SVGP01050). Filters were stored at -80°C. For dissolved nutrient analysis, 0.22 µm filtrate was collected in 125 ml acid-clean HDPE bottles (Nalgene) and stored at -20°C. Samples for single cell amplified genomes (SAG) were collected in August 2014. For each sample, 1 ml of raw water was incubated with 100 µl of glycerol-TE buffer (20 ml 100X TE pH 8 + 100 ml glycerol + 60 ml water; final concentration after sample addition is 10mM Tris, 1mM EDTA, 5% glycerol) for 10 minutes in the dark, then flash frozen in liquid nitrogen and stored at -80°C until processing.

Physicochemical data

CTD profiles, water chemistry and chlorophyll-*a* data were collected by the U.S. EPA according to standard protocols (84) and retrieved from the Great Lakes Environmental Database (https://cdx.epa.gov/) for 2012 and 2013. In addition, we measured dissolved nitrogen species from August 2013 samples. Ammonium concentrations were measured using the OPA method in a 96-well plate (85). Nitrate and nitrite concentrations were measured using the Greiss reaction method in a 96-well plate (86). Urea concentrations were measured in a 24 well plate using a colorimetric reaction (87).

16S rRNA analysis

The full 16S rRNA amplicon dataset was described by Paver and colleagues (34). Here we focus on data from the V4□V5 region (primers: 515F□Y, 926R (88)), collected in 2012 in tandem with metagenome samples from select stations. We classified sequences using the Silva v. 132 database (89) and the wang method (90) as implemented by mothur (91). Sequences classified to each detected family of nitrifiers (ammonia oxidizer families

Nitrosomonadaceae and *Nitrosopumilaceae*; nitrite oxidizer families *Gallionellaceae* and *Nitrospiraceae*) with a mothur-assigned confidence score above 90 were delineated into taxonomic units using minimum entropy decomposition with a minimum substantive abundance of 10 (92).

Metagenome and single-cell genome sequencing

One station per lake in Superior, Michigan, Huron, and Ontario, and two stations in Erie, were selected for metagenome sequencing. Spring 2012 metagenome samples were collected from the surface, and Summer 2012 metagenome samples were collected from the mid-hypolimnion (depths listed in Dataset S1). DNA was extracted using a modified phenol:chloroform extraction protocol (34) and libraries prepared according to the Illumina TruSeq protocol. Samples from spring 2012 were sequenced at the Joint Genome Institute using Illumina HiSeq (2x150bp). Samples from summer 2012 were sequenced at the University of Chicago Functional Genomics Core Facility using Illumina HiSeq 2500 (2x250bp).

To confirm the presence of proteorhodopsin, we analyzed a single cell amplified genome from *Nitrosospira* collected from Lake Michigan and sequenced by the Joint Genome Institute. Quality filtered reads were downloaded from JGI IMG/ER and normalized using bbnorm.sh with target=100 and mindepth=2. Normalized reads were assembled using SPAdes 3.1.11 in single cell mode (93) with flags --sc and --careful. Resulting scaffolds were annotated identically to MAGs as described below.

Obtaining metagenome-assembled genomes

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:

522 ktrim=r, mink=8, hdist=2, k=21, forcetrimleft=10, forcetrimright=199, minlen=150) using BBTools
523 suite version 35.74 (https://sourceforge.net/projects/bbmap/). Separate assemblies of quality
524 filtered reads were carried out for each metagenome using metaSPAdes 3.1.11 --meta mode
525 using default k sizes of 21, 33, 55 (94). To enable binning based on sequence coverage,
526 forward and reverse reads were merged using bbmerge in BBtools, using qtrim2=r
527 trimq=10,13,16 and adapter=default. Merged short reads were then mapped onto each
528 assembly using bowtie2 2.2.9 in --sensitive mode (95), and this coverage information was used
529 to bin assembled contigs. Binning was performed using MetaBAT2 2.12.1 (96), Binsanity 0.2.6.3
530 (97) and CONCOCT 1.0.0 (98) using default parameters. The resulting bins were scored,
531 aggregated, and de-replicated using DAS_Tool 1.1.1 (99) followed by manual curation using
532 Anvi'o 4.0 (100). We assessed genome completion and contamination of manually curated bins
533 using CheckM 1.1.0 lineage_wf (101), and all new MAGs presented here are greater than 70%
534 complete with less than 10% contamination (Dataset S3). Potential nitrifiers were screened by
535 searching for ammonia monooxygenase, hydroxylamine oxidoreductase and nitrite
536 oxidoreductase within reconstructed genomes using blastp 2.5.0 (102). For bins where any of
537 these genes were detected, we identified bacterial single copy core genes (103) or archaeal
538 single copy core genes (104) using HMMER (105), as implemented in Anvi'o. Single copy core
539 genes were queried against proteins predicted from bacterial and archaeal genomes in RefSeq
540 (NCBI) (106), and taxonomic identity of these core genes was ascertained based on a least
541 common ancestor approach using a 0.1% window around the bit score of the best hit using
542 KronaTools 2.7.1 (107). Taxonomic assignment was further validated using GTDB-tk 1.0.0
543 (108). Grouping of MAGs into clades and subclades based on ANI was carried out using
544 fastANI 1.1.0 (109). Genome characteristics for each genome group were calculated as the
545 median of those values for the group. Estimated complete genome size was calculated for
546 MAGs and for references in the pangenome analysis using CheckM (101) completion and
547 contamination, as follows: Estimated = Actual*(1-Contamination) / Completion. To quantify the

abundance of each clade/ecotype across samples, we used competitive mapping of merged short reads using bowtie2 in sensitive mode against all nitrifier MAGs, summing up mapped read count across all MAGs in a given clade/ecotype, and dividing by total mapped nitrifier reads in a sample; these values are shown in bubble plots (Fig 2, 4).

Annotation and gene cluster analysis

Reference genomes were obtained from GenBank (accession numbers listed in Supplemental Dataset S4). The full pangenome analyses included all the genomes listed in Supplemental Dataset S4, but we only report results from the subset of genomes most closely related to our MAGs. This subset consists of 86 *Nitrosomonadaceae*, 5 *Ca.* Nitrotoga, 78 *Nitrosopumilaceae* within Thaumarchaeota, and 75 *Nitrospira* that fall within Lineage II. Reference genomes were treated consistently with GL MAGs, with *de novo* gene calling by prodigal 2.6.3 (110) via Anvi'o. Unless otherwise noted, default settings were used for all software. Genes were annotated using InterProScan 5.30-69.0 (111), GhostKOALA (112) and eggnog-mapper 1.0.3 against the bactNOG database (113). Gene cluster analysis was carried out using the Anvi'o pangenome pipeline (114), using blastp to determine sequence similarity, ITEP to eliminate weak similarity (115) and MCL to cluster, using a minbit of 0.5, MCL inflation of 2 and minimum gene occurrence of 1 (116). Sigma factors were tallied by identifying gene clusters annotated with the following PFAMs: PF00309, PF03979, PF00140, PF04542, PF04539, PF04545, PF08281. Pseudogene counts were retrieved where available from NCBI PGAP annotated genomes (117). Paralog counts are reported as the number of gene clusters with more than one gene per genome. Intergenic spacers were calculated using bedtools complementBed function (118). Prokka 1.14.5 (119) was used to generate GenBank-format files from MAGs and SAGs, and genoPlotR 0.8.9 (120) was used to generate initial gene neighborhood maps.

Gene tree construction

The NspGL1 proteorhodopsin sequence was inserted into the MicRhoDE rhodopsin tree using pplacer (121) through the MicRhoDE Galaxy pipeline (56). 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 7.310 (122) along with high similarity sequences from NCBI nr that were not present in MicRhoDE. The tree was inferred using RaxML 8.2.12 with model PROTGAMMALG (123). The tree was visualized in iTOL (124) 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 (2, 25, 125). Sequences were aligned using MAFFT (122) and the tree was inferred using RaxML 8.2.12 with model PROTGAMMALG (123). Tree was visualized in iTOL (124) 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 (126) were obtained from UniProt, along with aquatic bacterial sequences described by Maresca and colleagues (76). The reference sequences were aligned using MAFFT (122), and sequences from GL MAGs were added using the MAFFT --addfragments option. The tree was estimated using IQ-TREE 2 1.6.11 (127) and visualized using iTOL (124).

Phylogenomic tree construction

Nitrospirae, *Thaumarchaeota*, *Gallionellaceae* and *Nitrosomonadaceae* genomes were downloaded from Genbank (NCBI) (128) and included in phylogenomic trees for their respective family. Phylogenomic analyses were carried out within Anvi'o. Briefly, single copy core genes were extracted as described above, individually aligned at the protein level using muscle (129),

and concatenated for each genome. Concatenated alignments were trimmed using Gblocks 0.91b (130) and analyzed by RAxML 8.2.12 (123) to create a phylogenetic tree using the PROTGAMMALG model and 50 bootstraps. Trees were visualized in iTOL (124).

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 1.1.0 (109) with --visualize flag. All contigs from this *Nitrosospira* SAG were found within an NspGL1 MAG. Bandage 0.8.1 (131) 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). Further, we verified that consistent coverage across this contig existed by mapping short reads from the original sample using bowtie2 (95) and viewing results using Integrated Genomics Viewer 2.7.0 (132) 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 (133) (BioProject PRJDB6644) onto the assembled contig described above using bowtie2 (95)**.** 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* (18). Manual assembly graph inspection with Bandage (131) 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 (134). 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. Dataset S6 summarizes the presence of genes related to nitrification and nitrogen metabolism across all our MAGs.

Verification of gene absences

Metagenome-assembled genomes typically comprise tens or even hundreds of contigs, and this fragmented nature makes it impossible to say with certainty whether a particular gene is truly absent. To substantiate our claims of gene absence based on MAGs, we used several lines of evidence. First, we note that our MAGs have high estimated completion (median 96.4%, mean 94.3%), based on the presence of universal core gene markers. Second, for all new lineages described here except NspGL3, we assembled multiple similar MAGs independently from different samples, and we inferred gene absences only if the absence was replicated in multiple assemblies. Together these two factors provide strong support for cases where a missing gene would be expected to occur in a region of predominantly core genes; however these factors are less informative for cases where a missing gene might occur in a genomic island, because we have no way of assessing the completion of regions lacking core genes, and islands tend to have systematic poor assemblies across samples. A third line of evidence that we considered is chromosome organization: if a single gene is deleted from an otherwise conserved region of synteny, then this deletion should be apparent in a gene neighborhood diagram (e.g. Figures S5, S7, S8, S10). Unfortunately in many cases, our MAGs are too dissimilar from reference genomes and share little synteny with them, so this approach is not always informative.

We used a fourth approach based on quantitative analysis of short reads to verify gene absences. If a suspected missing gene were actually present in the population, but failed to assemble and/or bin with the rest of the genome, then it should be detectable in the unassembled short reads. The frequency of a gene in the population can be estimated from its abundance in the short reads, compared to the abundance of core marker genes in the short

reads. We implemented this approach as follows. We searched unassembled short reads for each gene of interest that we identified as absent from MAGs (e.g. nitrosocyanin) using tblastn. Short reads with significant similarity were then filtered by best-hit taxonomy to the appropriate nitrifier group (i.e. *Nitrosomonadaceae*, *Ca.* Nitrotoga, *Nitrospira*, *Thaumarchaea*). These filtered short reads were enumerated and length-normalized (1000 * number of short reads / length of the target gene of interest). The same procedure was repeated for genes expected to be present in every cell (e.g. *amoAB*, *hao*, *nxrAB*, ribosomal protein genes) for comparison. If a putative missing gene (based on MAGs) has near-zero detection in the short reads, we can be confident that the gene is truly missing (or has undetectable sequence similarity, or was so recently acquired from another lineage that its best hit points to a different taxon). By contrast, if a putative missing gene (based on MAGs) does recover short reads, then the gene may be present in genomes related to our MAGs but was unassembled/unbinned, or the gene may be present in another lineage of nitrifiers that is not represented by our MAGs. Short read-based quantification of select genes is presented in Supplemental Dataset S7 and described in Supplemental Text.

Statistical analysis and plots

All statistical comparisons were carried out in R version 3.5.3 (135) and plots were generated using ggplot2 3.2.0 (136). Code and data files are available at bitbucket.org/greatlakes/gl_nitrifiers.

Data Availability

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

img.jgi.doe.gov, IMG Genome ID 3300033241. Raw reads are available in NCBI SRA (SRR14240538-SRR14240543) or through JGI with project IDs listed above.

Acknowledgments

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231. Sequencing support was provided by the DOE JGI Community Sequencing Program (CSP #1565, #503460). Computational resources were provided by the UChicago Research Computing Center. Funding for this work was provided by Illinois-Indiana Sea Grant (Grant # NA14OAR170095), UChicago Women's Board, and the National Science Foundation (OCE-1830011 to MLC). We thank the science staff in the Great Lakes National Program Office of the US EPA, and the captain and crew of the R/V *Lake Guardian*, for facilitating sample collection. We thank members of the Coleman and Waldbauer labs for assistance with sample collection and processing, and for discussion and comments on the manuscript.

Competing Interests

The authors declare no competing interests.

Figure & Table Legends

Table 1. Evidence for genome streamlining in nitrifiers from the Laurentian Great Lakes. Genome features were compared between Great Lakes MAGs and reference genomes using a two-sided Wilcoxon/Mann-Whitney test. n.s., not significant at 0.05 level. Only genomes with >70% completion and <10% contamination are included.

Figure 1. Dissolved inorganic nitrogen availability and distribution of nitrifiers across the Great Lakes. **(a)** Oxidized nitrogen concentrations. Values include NOx concentrations from published studies (n=128; (14, 33, 35, 137, 138)), US EPA Water Quality Surveys in 2012 and 2013 (n=1626 from GLENDA database), and this study (n=20). **(b)** Ammonium concentrations. Values are derived from the literature as in (a) (n=118) and from this study (n=20). **(c)** Distribution of nitrifiers across the Great Lakes. Top panel: map of sampling stations; stars indicate stations chosen for metagenome analysis. Bottom panel: relative abundance of ammonia-oxidizing (green) and nitrite-oxidizing (pink) families based on 16S rRNA V4-V5 amplicon sequencing, sampled in the mid-hypolimnion (except western Erie, sampled 1m from bottom). Data is plotted roughly west to east as indicated on the map.

Figure 2. Genome properties and cross-lake distribution of ammonia oxidizing organisms, showing both Archaea (top) and Betaproteobacteria (bottom). Rows highlighted in green represent clusters of genomes reconstructed from the Great Lakes and median values are shown for genome size, GC content, and coding density. For catalase, “E” indicates monofunctional catalase *katE*; “G” indicates bifunctional catalase-peroxidase *katG*. For carbon fixation, RuBisCO type is shown in parentheses (61); HP/HB, 3-Hydroxypropionate/4-Hydroxybutyrate cycle; CBB, Calvin-Benson-Bassham cycle. Bubble plot shows composition of ammonia oxidizers in hypolimnion samples, using MAGs as probes to recruit metagenomic reads (values sum to 100% for each lake column). Genes identified in only a subset of genomes are shown as (+).

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

highlighted (54, 55): blue boxes, diagnostic residues for PR; pink box, residue indicative of blue or green tuning; green underline, shorter beta-sheet region in PR. Sequence accession numbers: bacteriorhodopsin P02945, actinorhodopsin A0A1D9E0H1, sensory rhodopsin P42196, *Dokdonia* PR EAQ40507.1, SAR86 blue-tuned PR Q4PP54, SAR86 green-tuned PR Q9F7P4, SAR11 PR A6YQL7. (c) Phylogenetic tree showing close relatives of *Nitrosospira* PR within Supercluster III, as defined by MicRhoDE database (56). Neighboring clusters have been collapsed for clarity.

Figure 4. Genome properties and cross-lake distribution of nitrite oxidizing taxa *Nitrospira* (top) and *Ca.* Nitrotoga (Betaproteobacteria*;* bottom). Rows highlighted in pink represent clusters of genomes reconstructed from the Great Lakes and median values are shown for genome size, GC content, and coding density. rTCA, reductive tricarboxylic acid cycle; CBB, Calvin-Benson-Bassham cycle; ONR, octaheme nitrite reductase. Values in parentheses indicate RuBisCO type (61). Bubble plot shows composition of NOB per lake based on metagenomic read mapping. Genes identified in only a subset of genomes are indicated by (+). The asterisk (*) indicates for *Ca.* Nitrotoga, one *nxrAB* copy was recovered in genome assemblies, but short read analysis suggests two copies per genome (Supplemental Text).

Figure 5. Distinct photolyase proteins in NspiraGL. **a)** Phylogenetic tree showing families of photolyases. Three families are found in NspiraGL: CPD class I photolyase, FeS-BCP / CRYPro family, and an uncharacterized CPF-related family found in diverse Bacteria. CPD Class I photolyases are also found in other nitrifiers including *Ca*. Nitrotoga NtogaGL1a, *Nitrosospira* NspGL1, and *Nitrosarchaeum* NarchGL. **b)** Domain structure of the three photolyase families present in NspiraGL.

Supplemental Material

Supplemental Text. Detailed description of short-read analyses to verify gene absences.

10.6084/m9.figshare.15130350

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 families *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/). Values are presented in Supplemental Dataset S1.

Figure S2. Oligotype composition of nitrifiers across the Great Lakes. Each color represents a different 16S rRNA oligotype.

Figure S3. 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 (B.J. Campbell, L. Yu, J.F. Heidelberg, and D.L. Kirchman, PNAS 108:12776-12781, 2011). 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 S4. Histograms of nitrifier genome properties. Rows show each taxonomic group and columns show genome properties. Each circle represents a single genome/MAG. Great Lakes MAGs shown as filled circles.

Figure S5. Gene neighborhood surrounding proteorhodopsin 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.

Figure S6. Phylogenomic tree of *Nitrosopumilaceae* and *Cenarchaeum*. Tree was constructed using concatenated single-copy core genes as defined in (C. Rinke *et al.,* Nature 499:431-437, 2013). 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 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 (B.J. Campbell, L. Yu, J.F. Heidelberg, and D.L. Kirchman, PNAS 108:12776-12781, 2011). 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.

Dataset S1. Relative abundance of nitrifiers and associated environmental parameters. Relative abundance based on 16S rRNA amplicons is reported as proportion of total community reads. The full 16S rRNA dataset and flow cytometry cell counts are presented in Paver et al. (2020)

Environ Microbiol 22(1): 433-446. Chlorophyll a concentrations, irradiance values and dissolved oxygen measurements are from the US EPA Water Quality Survey (available from https://cdx.epa.gov/). Relative abundance based on metagenome read recruitment to MAGs is reported as the proportion of total sequenced reads for that sample. 10.6084/m9.figshare.15130350

Dataset S2. Nitrate (NOx), ammonium, and urea concentrations compiled from literature, EPA Water Quality Survey (available from https://cdx.epa.gov/) or measured in this study. NA, not available; n.d., not detected 10.6084/m9.figshare.15130350

Dataset S3. Assembly statistics and taxonomic information for metagenome-assembled genomes (MAGs) of nitrifiers in the Great Lakes. 10.6084/m9.figshare.15130350

Dataset S4. Genome properties for newly assembled MAGs and reference genomes included in this study. 10.6084/m9.figshare.15130350

Dataset S5. Protein functions differentiating genomes of Great Lakes nitrifiers from previously published genomes. Genome averages are calculated for reference genomes and GL MAGs. Genome-level counts for each individual GL MAG and for selected reference genomes are also shown. 10.6084/m9.figshare.15130350

Dataset S6. Detailed list of presence/absence of key nitrification genes in MAGs presented in this study. NCBI Protein IDs listed when a MAG contains the target gene. 0 indicates the gene was not recovered in the MAG. 10.6084/m9.figshare.15130350

Dataset S7. Short read analysis for verifying gene loss. Each target gene was searched in unassembled short reads as described in Materials & Methods. Values indicate the number of mapped reads, normalized to gene length. n.d., not detected. 10.6084/m9.figshare.15130350

Dataset S8. Stations are from the US EPA Water Quality Survey (available from https://cdx.epa.gov/). 10.6084/m9.figshare.15130350

References

1. Canfield DE, Glazer AN, Falkowski PG. 2010. The evolution and future of Earth's nitrogen cycle. Science 330:192–196.
2. Pachiadaki MG, Sintes E, Bergauer K, Brown JM, Record NR, Swan BK, Mathyer ME, Hallam SJ, Lopez-Garcia P, Takaki Y, Nunoura T, Woyke T, Herndl GJ, Stepanauskas R. 2017. Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. Science 358:1046–1051.
3. Reinthaler T, van Aken HM, Herndl GJ. 2010. Major contribution of autotrophy to microbial carbon cycling in the deep North Atlantic's interior. Deep Sea Res II 57:1572–1580.
4. Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland EDP, Gomez ML, Sieracki ME, DeLong EF, Herndl GJ, Stepanauskas R. 2011. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ccean. Science 333:1296–1300.
5. Baltar F, Herndl GJ. 2019. Ideas and perspectives: Is dark carbon fixation relevant for oceanic primary production estimates? Biogeosciences 16:3793–3799.
6. Callieri C, Coci M, Eckert EM, Salcher MM, Bertoni R. 2014. Archaea and Bacteria in deep lake hypolimnion: in situ dark inorganic carbon uptake. J Limnol 73:31-38.
7. Santoro AE, Richter RA, Dupont CL. 2019. Planktonic marine Archaea. Ann Rev Mar Sci 11:131–158.
8. Schleper C. 2010. Ammonia oxidation: different niches for bacteria and archaea? ISME J 4:1092–1094.
9. Santoro AE, Dupont CL, Richter RA, Craig MT, Carini P, McIlvin MR, Yang Y, Orsi WD, Moran DM, Saito MA. 2015. Genomic and proteomic characterization of "*Candidatus* Nitrosopelagicus brevis": An ammonia-oxidizing archaeon from the open ocean. Proc Natl Acad Sci USA 112:1173–1178.
10. Auguet J-C, Triadó-Margarit X, Nomokonova N, Camarero L, Casamayor EO. 2012.

Vertical segregation and phylogenetic characterization of ammonia-oxidizing Archaea in a deep oligotrophic lake. ISME J 6:1786–1797.

11. Herber J, Klotz F, Frommeyer B, Weis S, Straile D, Kolar A, Sikorski J, Egert M, Dannenmann M, Pester M. 2020. A single Thaumarchaeon drives nitrification in deep oligotrophic Lake Constance. Environ Microbiol 22:212–228.

12. Urbach E, Vergin KL, Young L, Morse A, Larson GL, Giovannoni SJ. 2001. Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. Limnol Oceanogr 46:557–572.

13. Okazaki Y, Fujinaga S, Tanaka A, Kohzu A, Oyagi H, Nakano S. 2017. Ubiquity and quantitative significance of bacterioplankton lineages inhabiting the oxygenated hypolimnion of deep freshwater lakes. ISME J 11:2279–2293.

14. Mukherjee M, Ray A, Post AF, McKay RM, Bullerjahn GS. 2016. Identification, enumeration and diversity of nitrifying planktonic archaea and bacteria in trophic end members of the Laurentian Great Lakes. J Great Lakes Res 42:39–49.

15. Hugoni M, Etien S, Bourges A, Lepère C, Domaizon I, Mallet C, Bronner G, Debroas D, Mary I. 2013. Dynamics of ammonia-oxidizing Archaea and Bacteria in contrasted freshwater ecosystems. Res Microbiol 164:360–370.

16. Hayden CJ, Beman JM. 2014. High abundances of potentially active ammonia-oxidizing Bacteria and Archaea in oligotrophic, high-altitude lakes of the Sierra Nevada, California, USA. PLoS ONE 9:e111560.

17. Bollmann A, Bar-Gilissen M-J, Laanbroek HJ. 2002. Growth at low ammonium concentrations and starvation response as potential factors involved in niche differentiation among ammonia-oxidizing bacteria. Appl Environ Microbiol 68:4751–4757.

18. Sedlacek CJ, McGowan B, Suwa Y, Sayavedra-Soto L, Laanbroek HJ, Stein LY, Norton JM, Klotz MG, Bollmann A. 2019. A physiological and genomic comparison of *Nitrosomonas* cluster 6a and 7 ammonia-oxidizing bacteria. Microb Ecol 78:985–994.

19. Alonso-Sáez L, Waller AS, Mende DR, Bakker K, Farnelid H, Yager PL, Lovejoy C, Tremblay J-É, Potvin M, Heinrich F, Estrada M, Riemann L, Bork P, Pedrós-Alió C, Bertilsson S. 2012. Role for urea in nitrification by polar marine Archaea. Proc Natl Acad Sci USA 109:17989–17994.

20. Daims H, Lücker S, Wagner M. 2016. A new perspective on microbes formerly known as nitrite-oxidizing bacteria. Trends Microbiol 24:699–712.

21. Nowka B, Daims H, Spieck E. 2015. Comparison of oxidation kinetics of nitrite-oxidizing bacteria: nitrite availability as a key factor in niche differentiation. Appl Environ Microbiol 81:745–753.

22. Wegen S, Nowka B, Spieck E. 2019. Low temperature and neutral pH define "*Candidatus* Nitrotoga sp." as a competitive nitrite oxidizer in coculture with *Nitrospira defluvii*. Appl Environ Microbiol 85:e02569-18.

23. Koch H, Galushko A, Albertsen M, Schintlmeister A, Gruber-Dorninger C, Lucker S, Pelletier E, Le Paslier D, Spieck E, Richter A, Nielsen PH, Wagner M, Daims H. 2014. Growth of nitrite-oxidizing bacteria by aerobic hydrogen oxidation. Science 345:1052–1054.

24. Füssel J, Lücker S, Yilmaz P, Nowka B, van Kessel MAHJ, Bourceau P, Hach PF, Littmann S, Berg J, Spieck E, Daims H, Kuypers MMM, Lam P. 2017. Adaptability as the key to success for the ubiquitous marine nitrite oxidizer *Nitrococcus*. Science Adv 3:e1700807.

25. Palatinszky M, Herbold C, Jehmlich N, Pogoda M, Han P, von Bergen M, Lagkouvardos I, Karst SM, Galushko A, Koch H, Berry D, Daims H, Wagner M. 2015. Cyanate as an energy source for nitrifiers. Nature 524:105–108.

26. Koch H, Lücker S, Albertsen M, Kitzinger K, Herbold C, Spieck E, Nielsen PH, Wagner M, Daims H. 2015. Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus *Nitrospira*. Proc Natl Acad Sci USA 112:11371–11376.

27. Boddicker AM, Mosier AC. 2018. Genomic profiling of four cultivated *Candidatus* Nitrotoga spp. predicts broad metabolic potential and environmental distribution. ISME J 12:2864–2882.

28. Sterner RW. 2010. In situ-measured primary production in Lake Superior. J Great Lakes Res 36:139–149.

29. Small GE, Bullerjahn GS, Sterner RW, Beall BFN, Brovold S, Finlay JC, McKay RML, Mukherjee M. 2013. Rates and controls of nitrification in a large oligotrophic lake. Limnol Oceanogr 58:276–286.

30. Vollenweider RA, Munawar M, Stadelmann P. 1974. A comparative review of phytoplankton and primary production in the Laurentian Great Lakes. J Fish Res Bd Can 31:739–762.

31. Clevinger CC, Heath RT, Bade DL. 2014. Oxygen use by nitrification in the hypolimnion and sediments of Lake Erie. J Great Lakes Res 40:202–207.

32. Neilson MA, Stevens RJJ. 1987. Spatial heterogeneity of nutrients and organic matter in Lake Ontario. Can J Fish Aquat Sci 44:2192–2203.

33. Lean DRS, Knowles R. 1987. Nitrogen transformations in Lake Ontario. Can J Fish Aquat Sci 44:2133–2143.

34. Paver SF, Newton RJ, Coleman ML. 2020. Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. Environ Microbiol 22:433–446.

35. Rozmarynowycz MJ, Beall BFN, Bullerjahn GS, Small GE, Sterner RW, Brovold SS, D'souza NA, Watson SB, McKay RML. 2019. Transitions in microbial communities along a 1600 km freshwater trophic gradient. J Great Lakes Res 45:263–276.

36. Fujimoto M, Cavaletto J, Liebig JR, McCarthy A, Vanderploeg HA, Denef VJ. 2016. Spatiotemporal distribution of bacterioplankton functional groups along a freshwater estuary to pelagic gradient in Lake Michigan. J Great Lakes Res 42:1036–1048.

37. Hooper AB, Terry KR. 1974. Photoinactivation of ammonia oxidation in *Nitrosomonas*. J

Bacteriol 119:899-906.

38. Horrigan SG, Springer AL. 1990. Oceanic and estuarine ammonium oxidation: effects of light. Limnol Oceanogr 35:479–482.

39. Guerrero M, Jones R. 1996. Photoinhibition of marine nitrifying bacteria. I. Wavelength-dependent response. Mar Ecol Prog Ser 141:183–192.

40. Merbt SN, Stahl DA, Casamayor EO, Martí E, Nicol GW, Prosser JI. 2012. Differential photoinhibition of bacterial and archaeal ammonia oxidation. FEMS Microbiol Lett 327:41–46.

41. Smith JM, Chavez FP, Francis CA. 2014. Ammonium uptake by phytoplankton regulates nitrification in the sunlit ocean. PLoS ONE 9:e108173.

42. Kumar S, Sterner RW, Finlay JC, Brovold S. 2007. Spatial and temporal variation of ammonium in Lake Superior. J Great Lakes Res 33:581–591.

43. Dove A, Chapra SC. 2015. Long-term trends of nutrients and trophic response variables for the Great Lakes. Limnol Oceanogr 60:696–721.

44. Belisle BS, Steffen MM, Pound HL, Watson SB, DeBruyn JM, Bourbonniere RA, Boyer GL, Wilhelm SW. 2016. Urea in Lake Erie: Organic nutrient sources as potentially important drivers of phytoplankton biomass. J Great Lakes Res 42:599–607.

45. Rice MC, Norton JM, Valois F, Bollmann A, Bottomley PJ, Klotz MG, Laanbroek HJ, Suwa Y, Stein LY, Sayavedra-Soto L, Woyke T, Shapiro N, Goodwin LA, Huntemann M, Clum A, Pillay M, Kyrpides N, Varghese N, Mikhailova N, Markowitz V, Palaniappan K, Ivanova N, Stamatis D, Reddy TBK, Ngan CY, Daum C. 2016. Complete genome of *Nitrosospira briensis* C-128, an ammonia-oxidizing bacterium from agricultural soil. Stand Genomic Sci 11:46.

46. Garcia JC, Urakawa H, Le VQ, Stein LY, Klotz MG, Nielsen JL. 2013. Draft genome sequence of *Nitrosospira* sp. strain APG3, a psychrotolerant ammonia-oxidizing bacterium isolated from sandy lake sediment. Genome Announcements 1:e00930-13.

47. Norton JM, Klotz MG, Stein LY, Arp DJ, Bottomley PJ, Chain PSG, Hauser LJ, Land ML, Larimer FW, Shin MW, Starkenburg SR. 2008. Complete genome sequence of *Nitrosospira multiformis*, an ammonia-oxidizing bacterium from the soil environment. Appl Environ Microbiol 74:3559–3572.

48. Giovannoni SJ, Thrash JC, Temperton B. 2014. Implications of streamlining theory for microbial ecology. ISME J 8:1553–1565.

49. Bollmann A, Sedlacek CJ, Norton J, Laanbroek HJ, Suwa Y, Stein LY, Klotz MG, Arp D, Sayavedra-Soto L, Lu M, Bruce D, Detter C, Tapia R, Han J, Woyke T, Lucas SM, Pitluck S, Pennacchio L, Nolan M, Land ML, Huntemann M, Deshpande S, Han C, Chen A, Kyrpides N, Mavromatis K, Markowitz V, Szeto E, Ivanova N, Mikhailova N, Pagani I, Pati A, Peters L, Ovchinnikova G, Goodwin LA. 2013. Complete genome sequence of *Nitrosomonas* sp. Is79, an ammonia oxidizing bacterium adapted to low ammonium concentrations. Stand Genomic Sci 7:469–482.

50. Beja O. 2000. Bacterial Rhodopsin: Evidence for a new type of phototrophy in the sea. Science 289:1902–1906.

51. Béjà O, Spudich EN, Spudich JL, Leclerc M, DeLong EF. 2001. Proteorhodopsin phototrophy in the ocean. Nature 411:786–789.

52. Sabehi G, Loy A, Jung K-H, Partha R, Spudich JL, Isaacson T, Hirschberg J, Wagner M, Béjà O. 2005. New insights into metabolic properties of marine bacteria encoding proteorhodopsins. PLoS Biol 3:e273.

53. Martinez A, Bradley AS, Waldbauer JR, Summons RE, DeLong EF. 2007. Proteorhodopsin photosystem gene expression enables photophosphorylation in a heterologous host. Proc Natl Acad Sci USA 104:5590–5595.

54. Reckel S, Gottstein D, Stehle J, Löhr F, Verhoefen M-K, Takeda M, Silvers R, Kainosho M, Glaubitz C, Wachtveitl J, Bernhard F, Schwalbe H, Güntert P, Dötsch V. 2011. Solution NMR structure of proteorhodopsin. Angewandte Chemie International Edition 50:11942–

11946.

55. Kralj JM, Bergo VB, Amsden JJ, Spudich EN, Spudich JL, Rothschild KJ. 2008. Protonation state of Glu142 differs in the green- and blue-absorbing variants of proteorhodopsin. Biochemistry 47:3447–3453.

56. Boeuf D, Audic S, Brillet-Guéguen L, Caron C, Jeanthon C. 2015. MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. Database (Oxford) bav080.

57. Hyman MR, Arp DJ. 1992. 14C2H2- and 14CO2-labeling studies of the de novo synthesis of polypeptides by *Nitrosomonas europaea* during recovery from acetylene and light inactivation of ammonia monooxygenase. J Biol Chem 267:1534–1545.

58. Yousef F, Shuchman R, Sayers M, Fahnenstiel G, Henareh A. 2017. Water clarity of the Upper Great Lakes: Tracking changes between 1998–2012. J Great Lakes Res 43:239–247.

59. Beaumont HJE, Hommes NG, Sayavedra-Soto LA, Arp DJ, Arciero DM, Hooper AB, Westerhoff HV, van Spanning RJM. 2002. Nitrite reductase of *Nitrosomonas europaea* is not essential for production of gaseous nitrogen oxides and confers tolerance to nitrite. J Bacteriol 184:2557–2560.

60. Watzer B, Forchhammer K. 2018. Cyanophycin synthesis optimizes nitrogen utilization in the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. Appl Environ Microbiol 84:e01298-18.

61. Badger MR, Bek EJ. 2008. Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO2 acquisition by the CBB cycle. J Exp Bot 59:1525–1541.

62. Rae BD, Long BM, Badger MR, Price GD. 2013. Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO2 fixation in cyanobacteria and some proteobacteria. Microbiol Mol Biol Rev 77:357–379.

63. Cabello-Yeves PJ, Zemskaya TI, Rosselli R, Coutinho FH, Zakharenko AS, Blinov VV,

Rodriguez-Valera F. 2018. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. Appl Environ Microbiol 84:e02132-17.

64. Sterner RW, Smutka TM, McKay RML, Xiaoming Q, Brown ET, Sherrell RM. 2004. Phosphorus and trace metal limitation of algae and bacteria in Lake Superior. Limnol Oceanogr 49:495–507.

65. Ryals J, Hsu RY, Lipsett MN, Bremer H. 1982. Isolation of single-site Escherichia coli mutants deficient in thiamine and 4-thiouridine syntheses: identification of a nuvC mutant. J Bacteriol 151:899–904.

66. Kitzinger K, Koch H, Lücker S, Sedlacek CJ, Herbold C, Schwarz J, Daebeler A, Mueller AJ, Lukumbuzya M, Romano S, Leisch N, Karst SM, Kirkegaard R, Albertsen M, Nielsen PH, Wagner M, Daims H. 2018. Characterization of the first "*Candidatus* Nitrotoga" isolate reveals metabolic versatility and separate evolution of widespread nitrite-oxidizing bacteria. mBio 9:e01186-18.

67. Ishii K, Fujitani H, Sekiguchi Y, Tsuneda S. 2020. Physiological and genomic characterization of a new "*Candidatus* Nitrotoga" isolate. Environ Microbiol 22:2365–2382.

68. Starkenburg SR, Chain PSG, Sayavedra-Soto LA, Hauser L, Land ML, Larimer FW, Malfatti SA, Klotz MG, Bottomley PJ, Arp DJ, Hickey WJ. 2006. Genome sequence of the chemolithoautotrophic nitrite-oxidizing bacterium *Nitrobacter winogradskyi* Nb-255. Appl Environ Microbiol 72:2050–2063.

69. Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B, Rattei T, Damste JSS, Spieck E, Le Paslier D, Daims H. 2010. A *Nitrospira* metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. Proc Natl Acad Sci USA 107:13479–13484.

70. Lücker S, Nowka B, Rattei T, Spieck E, Daims H. 2013. The genome of *Nitrospina gracilis* illuminates the metabolism and evolution of the major marine nitrite oxidizer. Front Microbiol 4:27.

71. Kitzinger K, Marchant HK, Bristow LA, Herbold CW, Padilla CC, Kidane AT, Littmann S, Daims H, Pjevac P, Stewart FJ, Wagner M, Kuypers MMM. 2020. Single cell analyses reveal contrasting life strategies of the two main nitrifiers in the ocean. Nat Commun 11:767.

72. Mundinger AB, Lawson CE, Jetten MSM, Koch H, Lücker S. 2019. Cultivation and transcriptional analysis of a canonical *Nitrospira* under stable growth conditions. Front Microbiol 10:1325.

73. Sancar A. 2003. Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors. Chem Rev 103:2203–2238.

74. Zadow A von, Ignatz E, Pokorny R, Essen L-O, Klug G. 2016. *Rhodobacter sphaeroides* CryB is a bacterial cryptochrome with (6–4) photolyase activity. FEBS J 283:4291–4309.

75. Zhang F, Scheerer P, Oberpichler I, Lamparter T, Krauß N. 2013. 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 110:7217–7222.

76. Maresca JA, Keffer JL, Hempel PP, Polson SW, Shevchenko O, Bhavsar J, Powell D, Miller KJ, Singh A, Hahn MW. 2019. Light modulates the physiology of nonphototrophic Actinobacteria. J Bacteriol 201:e00740-18.

77. Wink DA, Kasprzak KS, Maragos CM, Elespuru RK, Misra M, Dunams TM, Cebula TA, Koch WH, Andrews AW, Allen JS, Et A. 1991. DNA deaminating ability and genotoxicity of nitric oxide and its progenitors. Science 254:1001–1003.

78. Imlay JA. 2013. The molecular mechanisms and physiological consequences of oxidative stress: lessons from a model bacterium. Nat Rev Microbiol 11:443–454.

79. Zhang T, Hansel CM, Voelker BM, Lamborg CH. 2016. Extensive dark biological production of reactive oxygen species in brackish and freshwater ponds. Environ Sci Technol 50:2983–2993.

80. Khademian M, Imlay JA. 2017. *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 114:E6922–E6931.

81. O'Reilly CM, Sharma S, Gray DK, Hampton SE, Read JS, Rowley RJ, Schneider P, Lenters JD, McIntyre PB, Kraemer BM, Weyhenmeyer GA, Straile D, Dong B, Adrian R, Allan MG, Anneville O, Arvola L, Austin J, Bailey JL, Baron JS, Brookes JD, Eyto E de, Dokulil MT, Hamilton DP, Havens K, Hetherington AL, Higgins SN, Hook S, Izmest'eva LR, Joehnk KD, Kangur K, Kasprzak P, Kumagai M, Kuusisto E, Leshkevich G, Livingstone DM, MacIntyre S, May L, Melack JM, Mueller Navarra DC, Naumenko M, Noges P, Noges T, North RP, Plisnier P-D, Rigosi A, Rimmer A, Rogora M, Rudstam LG, Rusak JA, Salmaso N, Samal NR, Schindler DE, Schladow SG, Schmid M, Schmidt SR, Silow E, Soylu ME, Teubner K, Verburg P, Voutilainen A, Watkinson A, Williamson CE, Zhang G. 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophys Res Lett 42:10773-10781.

82. Sterner RW, Keeler B, Polasky S, Poudel R, Rhude K, Rogers M. 2020. Ecosystem services of Earth's largest freshwater lakes. Ecosystem Services 41:101046.

83. Barbiero RP, Lesht BM, Hinchey EK, Nettesheim TG. 2018. A brief history of the U.S. EPA Great Lakes National Program Office's water quality survey. J Great Lakes Res 44:539–546.

84. US Environmental Protection Agency. 2003. Sampling and analytical procedures for GLNPO's open lake water quality survey of the Great Lakes. Great Lakes National Program Office Illinois, USA.

85. Holmes RM, Aminot A, Kérouel R, Hooker BA, Peterson BJ. 1999. A simple and precise method for measuring ammonium in marine and freshwater ecosystems. Can J Fish Aquat Sci 56:1801–1808.

86. Miranda KM, Espey MG, Wink DA. 2001. A rapid, simple spectrophotometric method for simultaneous detection of nitrate and nitrite. Nitric Oxide 5:62–71.

87. Revilla M, Alexander J, Glibert PM. 2005. Urea analysis in coastal waters: comparison of enzymatic and direct methods. Limnol Oceanogr Methods 3:290–299.

88. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol 18:1403–1414.

89. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

90. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267.

91. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79:5112–5120.

92. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Methods Ecol Evol 4:1111–1119.

93. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

94. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834.

95. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

96. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately

reconstructing single genomes from complex microbial communities. PeerJ 3:e1165.

97. Graham ED, Heidelberg JF, Tully BJ. 2017. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. PeerJ 5:e3035.

98. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146.

99. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 3:836–843.

100. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.

101. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

102. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinform 10:421.

103. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. 2011. Activity of abundant and rare bacteria in a coastal ocean. Proc Natl Acad Sci USA 108:12776–12781.

104. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, Dodsworth JA, Hedlund BP, Tsiamis G, Sievert SM, Liu W-T, Eisen JA, Hallam SJ, Kyrpides NC, Stepanauskas R, Rubin EM, Hugenholtz P, Woyke T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499:431–437.

105. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41:e121–

e121.

106. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745.

107. Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinform 12.

108. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925-1927.

109. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun 9:1–8.

110. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform 11.

111. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.

112. Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J Mol Biol 428:726–731.

113. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. Mol Biol Evol 34:2115–2122.
114. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. PeerJ 6:e4320.
115. Benedict MN, Henriksen JR, Metcalf WW, Whitaker RJ, Price ND. 2014. ITEP: An integrated toolkit for exploration of microbial pan-genomes. BMC Genomics 15:8.
116. van Dongen S, Abreu-Goodger C. 2012. Using MCL to extract clusters from networks. Methods Mol Biol 804:281–295.
117. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 44:6614–6624.
118. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
119. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069.
120. Guy L, Roat Kultima J, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. Bioinformatics 26:2334–2335.
121. Matsen FA, Kodner RB, Armbrust EV. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinform 11:538.
122. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066.
123. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
124. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and

annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245.

125. Spang A, Poehlein A, Offre P, Zumbrägel S, Haider S, Rychlik N, Nowka B, Schmeisser C, Lebedeva EV, Rattei T, Böhm C, Schmid M, Galushko A, Hatzenpichler R, Weinmaier T, Daniel R, Schleper C, Spieck E, Streit W, Wagner M. 2012. The genome of the ammonia-oxidizing *Candidatus* Nitrososphaera gargensis: insights into metabolic versatility and environmental adaptations. Environ Microbiol 14:3122–3145.

126. Vechtomova YuL, Telegina TA, Kritsky MS. 2020. Evolution of proteins of the DNA photolyase/cryptochrome family. Biochemistry Moscow 85:131–153.

127. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534.

128. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2005. GenBank. Nucleic Acids Res 33:D34–D38.

129. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

130. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552.

131. Wick RR, Schultz MB, Zobel J, Holt KE. 2015. Bandage: interactive visualization of de novo genome assemblies. Bioinformatics 31:3350–3352.

132. Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192.

133. Okazaki Y, Nishimura Y, Yoshida T, Ogata H, Nakano S-I. 2019. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. Environ Microbiol 21:4740-4754.

134. Orellana LH, Rodriguez-R LM, Konstantinidis KT. 2017. ROCker: accurate detection and

quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. Nucleic Acids Res 45:e14.

135. Team RC. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing V, Austria. 2016.

136. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

137. Murphy TP. 1980. Ammonia and nitrate uptake in the lower Great Lakes. Can J Fish Aquat Sci 37:1365–1372.

138. Gardner WS, Lavrentyev PJ, Cavaletto JF, McCarthy MJ, Eadie BJ, Johengen TH, Cotner JB. 2004. Distribution and dynamics of nitrogen and microbial plankton in southern Lake Michigan during spring transition 1999-2000. J Geophys Res: Oceans 109:C03007.

Figure 1

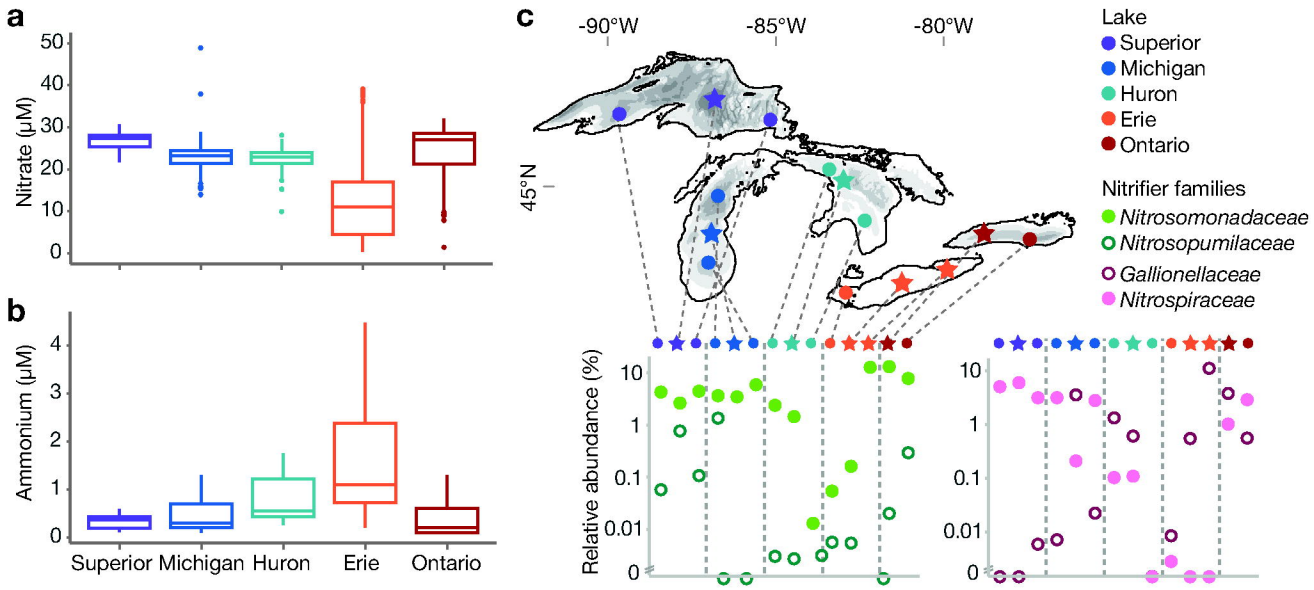

a
Nitrate (µM)
50
40
30
20
10
0
b
Ammonium (µM)
4
3
2
1
0
Superior
Michigan
Huron
Erie
Ontario
c
-90°W
-85°W
-80°W
45°N
Lake
Superior
Michigan
Huron
Erie
Ontario
Nitrifier families
Nitrosomonadaceae
Nitrosopumilaceae
Gallionellaceae
Nitrospiraceae
Relative abundance (%)
10
1
0.1
0.01
0

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 *(n=3)*	31.4	1.24	90	-	+	-	-	+	-	-	-	+	HP/HB
	Nitrosarchaeum 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 *(n=7)*	40.7	1.42	89	G	+	+	+	+	-	-	-	+	CBB (IAc)
	NspGL3 *(n=1)*	50.3	1.66	90	-	-	-	-	-	-	+	-	-	CBB (IAq)
	NspGL2a *(n=2)*	50.1	1.57	90	-	-	-	-	-	-	+	-	-	CBB (IC)
	NspGL2b *(n=5)*	50.0	1.53	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

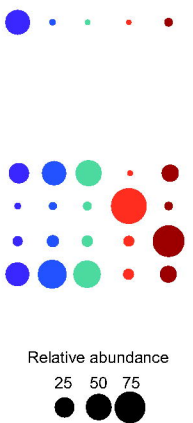


Figure 3

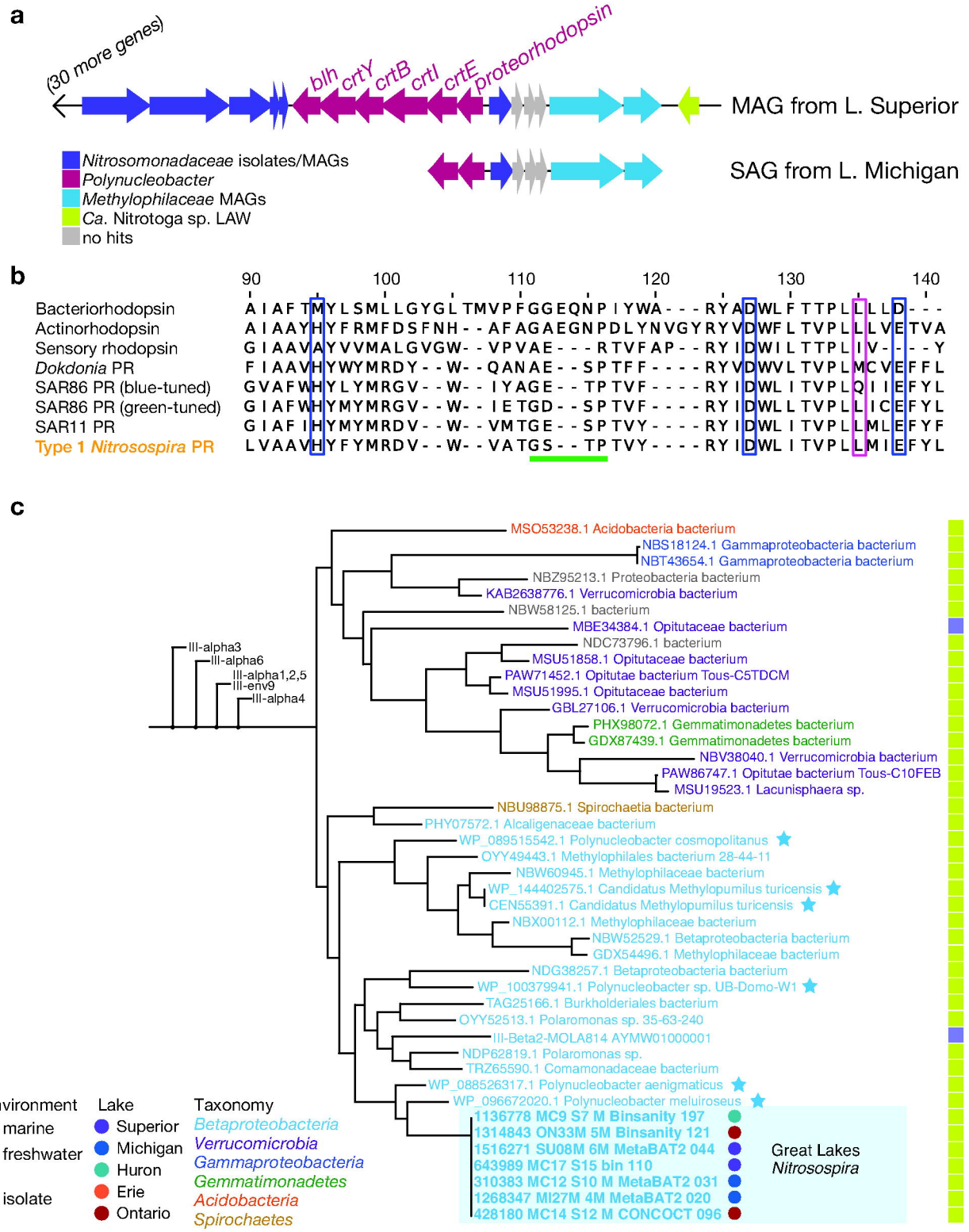

a
(30 more genes)
blh
crtY
crtB
crtI
crtE
proteorhodopsin
MAG from L. Superior
SAG from L. Michigan
Nitrosomonadaceae isolates/MAGs
Polynucleobacter
Methylophilaceae MAGs
Ca. Nitrotoga sp. LAW
no hits
b
90 100 110 120 130 140
Bacteriorhodopsin
Actinorhodopsin
Sensory rhodopsin
Dokdonia PR
SAR86 PR (blue-tuned)
SAR86 PR (green-tuned)
SAR11 PR
Type 1 Nitrosospira PR
c
MSO53238.1 Acidobacteria bacterium
NBS18124.1 Gammaproteobacteria bacterium
NBT43654.1 Gammaproteobacteria bacterium
NBZ95213.1 Proteobacteria bacterium
KAB2638776.1 Verrucomicrobia bacterium
NBW58125.1 bacterium
MBE34384.1 Opitutaceae bacterium
NDC73796.1 bacterium
MSU51858.1 Opitutaceae bacterium
PAW71452.1 Opitutae bacterium Tous-C5TDCM
MSU51995.1 Opitutaceae bacterium
GBL27106.1 Verrucomicrobia bacterium
PHX98072.1 Gemmatimonadetes bacterium
GDX87439.1 Gemmatimonadetes bacterium
NBV38040.1 Verrucomicrobia bacterium
PAW86747.1 Opitutae bacterium Tous-C10FEB
MSU19523.1 Lacunisphaera sp.
III-alpha3
III-alpha6
III-alpha1,2,5
III-env9
III-alpha4
NBU98875.1 Spirochaetia bacterium
PHY07572.1 Alcaligenaceae bacterium
WP_089515542.1 Polynucleobacter cosmopolitanus
OYY49443.1 Methylophilales bacterium 28-44-11
NBW60945.1 Methylophilaceae bacterium
WP_144402575.1 Candidatus Methylopumilus turicensis
CEN55391.1 Candidatus Methylopumilus turicensis
NBX00112.1 Methylophilaceae bacterium
NBW52529.1 Betaproteobacteria bacterium
GDX54496.1 Methylophilaceae bacterium
NDG38257.1 Betaproteobacteria bacterium
WP_100379941.1 Polynucleobacter sp. UB-Domo-W1
TAG25166.1 Burkholderiales bacterium
OYY52513.1 Polaromonas sp. 35-63-240
III-Beta2-MOLA814 AYMW01000001
NDP62819.1 Polaromonas sp.
TRZ65590.1 Comamonadaceae bacterium
WP_088526317.1 Polynucleobacter aenigmaticus
WP_096672020.1 Polynucleobacter meluiroseus
1136778 MC9 S7 M Binsanity 197
1314843 ON33M 5M Binsanity 121
1516271 SU08M 6M MetaBAT2 044
643989 MC17 S15 bin 110
310383 MC12 S10 M MetaBAT2 031
1268347 MI27M 4M MetaBAT2 020
428180 MC14 S12 M CONCOCT 096
Great Lakes Nitrosospira
Environment
marine
freshwater
isolate
Lake
Superior
Michigan
Huron
Erie
Ontario
Taxonomy
Betaproteobacteria
Verrucomicrobia
Gammaproteobacteria
Gemmatimonadetes
Acidobacteria
Spirochaetes
Tree scale: 0.1

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*	# *nxrAB* copies	Cyanase	Urease	
Nitrospira	NspiraGL *(n=6)*	57.8	1.83	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 *(n=4)*	51.7	1.42	91	4	(+)	G	-	-	-	-	-	1*	+	-	CBB (IC)
	NtogaGL1b *(n=2)*	51.8	1.46	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

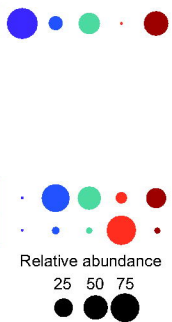
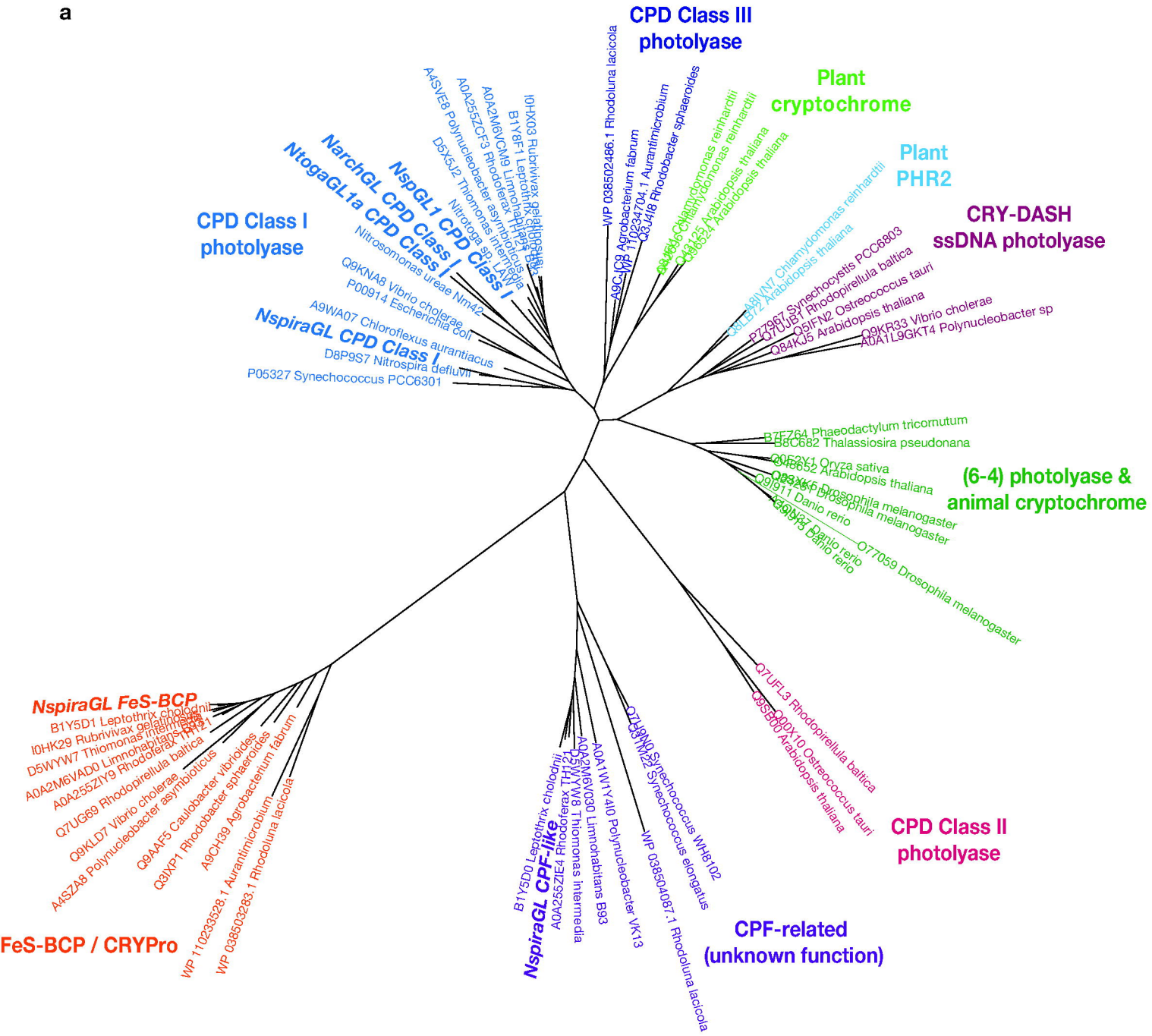


Figure 5

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 0385024 86.1 Rhodoluna lacicola
A9CJC9 Agrobacterium fabrum
WP 110234704.1 Aurantimicrobium
Q3J418 Rhodobacter sphaeroides
Q9KNA8 Vibrio cholerae
P00914 Escherichia coli
A9WA07 Chloroflexus aurantiacus
D8P9S7 Nitrospira defluvii
P05327 Synechococcus PCC6301
B7Z764 Phaeodactylum tricornutum
B8C682 Thalassiosira pseudonana
Q9KR33 Vibrio cholerae
A0A1L9GKT4 Polynucleobacter sp
Q84KJ5 Arabidopsis thaliana
Q5IFN2 Ostreococcus tauri
Q7UUB1 Rhodopirellula baltica
P77967 Synechocystis PCC6803
Q7UFL3 Rhodopirellula baltica
Q9SB00 Arabidopsis thaliana
Q00X10 Ostreococcus tauri
WP 038504087.1 Rhodoluna lacicola
Q7UGN2 Synechococcus elongatus
Q7U3N2 Synechococcus WH8102
A0A1W1Y4I0 Polynucleobacter VK13
A0A2M6V030 Limnohabitans B93
A0A2552IE4 Rhodoferax TH121
B1Y5D0 Leptothrix cholodnii
B1Y5D1 Leptothrix cholodnii
I0HK29 Rubrivivax gelatinosus
D5WYW7 Thiomonas intermedia
A0A2M6VAD0 Limnohabitans B93
A0A255ZIY9 Rhodoferax TH121
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

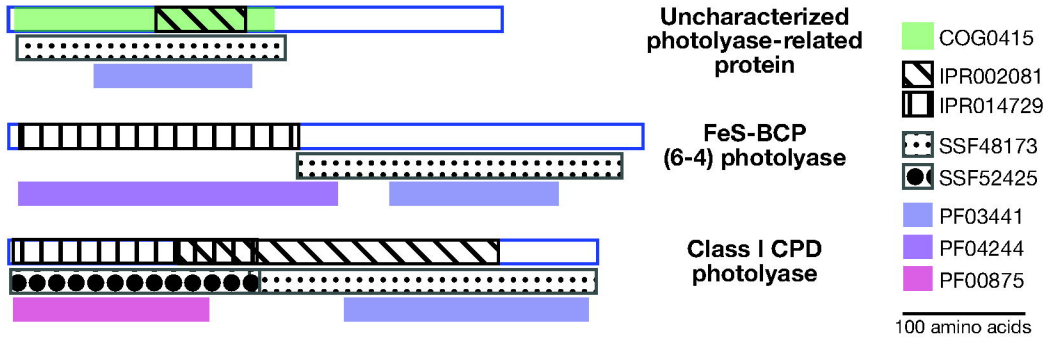

Uncharacterized photolyase-related protein
FeS-BCP (6-4) photolyase
Class I CPD photolyase
COG0415
IPR002081
IPR014729
SSF48173
SSF52425
PF03441
PF04244
PF00875
100 amino acids

Table 1. Evidence for genome streamlining in nitrifiers from the Laurentian Great Lakes. W, two-sided Wilcoxon/Mann-Whitney test statistic. n.s., not significant at 0.05 level.

taxonomic group	genome feature	median (reference)	median (GL)	W	p value	n (ref)	n (GL)
Nitrosomonadaceae	coding density	0.856	0.892	34	5.5E-09	86	15
	est complete size	3210560	1450843	1285	1.0E-09	86	15
	median intergenic	114	80	1254	6.3E-09	86	15
	paralogs	123	29	1189	2.1E-07	86	15
	pseudogenes	101	11	1215	9.0E-10	81	15
	sigma factors	8	4	1275	1.2E-09	86	15
Nitrospira	coding density	0.876	0.894	64	0.0074	64	6
	est complete size	3790956	1828031	373	1.5E-04	64	6
	median intergenic	90	78	299	0.026	64	6
	paralogs	212	49	376	1.2E-04	64	6
	pseudogenes	69	9	182	2.9E-04	31	6
	sigma factors	13	5	379	8.3E-05	64	6
Ca. Nitrotoga	coding density	0.857	0.910	0	0.0080	5	6
	est complete size	2858108	1441179	30	0.0081	5	6
	median intergenic	122	72	30	0.0080	5	6
	paralogs	93	23	30	0.0081	5	6
	pseudogenes	18	8	6	n.s.	1	6
	sigma factors	8	4	30	0.0054	5	6
Thaumarchaeota (Nitrosopumilaceae)	coding density	0.900	0.898	102	n.s.	62	3
	est complete size	1398741	1242579	153	n.s.	62	3
	median intergenic	61	66	60	n.s.	62	3
	paralogs	85	38	175	0.011	62	3
	pseudogenes	22	13	82	n.s.	34	3