MICROBIOLOGY

Postglacial adaptations enabled colonization and quasi-clonal dispersal of ammonia-oxidizing archaea in modern European large lakes

David Kamanda Ngugi[1]*, Michaela M. Salcher[2], Adrian-Stefan Andrei[3], Rohit Ghai[2], Franziska Klotz[4], Maria-Cecilia Chiriac[2], Danny Ionescu[5], Petra Büsing[1], Hans-Peter Grossart[5,6,7], Peng Xing[8], John C. Priscu[9], Salmor Alymkulov[10], Michael Pester[1,11]

Ammonia-oxidizing archaea (AOA) play a key role in the aquatic nitrogen cycle. Their genetic diversity is viewed as the outcome of evolutionary processes that shaped ancestral transition from terrestrial to marine habitats. However, current genome-wide insights into AOA evolution rarely consider brackish and freshwater representatives or provide their divergence timeline in lacustrine systems. An unbiased global assessment of lacustrine AOA diversity is critical for understanding their origins, dispersal mechanisms, and ecosystem roles. Here, we leveraged continental-scale metagenomics to document that AOA species diversity in freshwater systems is remarkably low compared to marine environments. We show that the uncultured freshwater AOA, "*Candidatus* Nitrosopumilus limneticus," is ubiquitous and genotypically static in various large European lakes where it evolved 13 million years ago. We find that extensive proteome remodeling was a key innovation for freshwater colonization of AOA. These findings reveal the genetic diversity and adaptive mechanisms of a keystone species that has survived clonally in lakes for millennia.

INTRODUCTION

Microbes of the phylum Nitrososphaerota (conventionally known as Thaumarchaeota) (*1*, *2*) comprise all ammonia-oxidizing archaea (AOA) that are widely distributed in aquatic and terrestrial environments (*3*, *4*) and include key species that mediate global nitrogen, carbon, and phosphorus cycles (*4*, *5*). Previous studies showed a distinct habitat specificity of AOA in different ecosystems [especially marine and terrestrial systems; see (*6*)], where niche boundaries, diversification, and biogeography are reflected in adaptations to prevailing environmental conditions [e.g., high salinity and pH; reviewed in (*3*, *4*, *7*)].

Recent genome-wide evolutionary studies inferred a timeline of AOA evolution that extends from a thermophilic autotrophic terrestrial ancestor to the initial colonization of the shallow marine environments, and subsequent expansion into the deep interior of the ocean (*8*–*11*). Within this evolutionary framework, the initial transition from terrestrial to shallow marine habitat took place during the Meso-Proterozoic [~1.02 billion years (Ga) ago] and was accompanied by several functional gene gains and losses, including the adoption of metabolic pathways conferring resistance to salinity (*8*). Subsequent niche expansion of the Shallow Marine Group (SMG) AOA into the interior ocean is estimated to have occurred around 635 to 560 million years (Ma) ago through acquisition of alternative strategies of nutrient acquisition and energy conservation (*8*, *12*). A more recent study estimates a much younger evolutionary timeline for the major AOA clades (*11*). However, both studies focused exclusively on marine and terrestrial Nitrososphaerota, completely neglecting transitions into/from brackish and freshwaters.

To date, only one recent study has examined the genetic diversity and evolutionary history of Nitrososphaerota in freshwater environments (*13*). In this pioneering study, three major clades of AOA (related to the genera *Nitrosoarchaeum*, *Nitrosopumilus*, and *Nitrosotenuis*) were identified, their distribution along surface sediment transects and waters in a selected number of lakes and rivers was determined, and possible genetic mechanisms enabling adaptation to freshwater were identified. However, the limited number of lacustrine ecosystems analyzed, the exclusion of globally important freshwaters such as the Laurentian Great Lakes or European lakes, and the neglect of brackish water as a bridging environment for transition from marine to freshwater, combined with the strong focus on surface sediments (which are distinct from open waters), left several fundamental questions about the origin, diversity, and evolutionary history of lacustrine Nitrososphaerota unanswered. First, is there an evolutionary relationship between AOA in ancient and modern freshwater lakes? Second, are brackish and freshwater AOA clades evolutionarily related, or did freshwater AOA directly descend from marine species? Third, are freshwater AOA products of relatively recent biogeographic events that parallel the Miocene (5 to 24 Ma) diversification of fauna inhabiting modern lakes (*14*, *15*)? Fourth, if biased taxon selection affects

[1] Leibniz Institute DSMZ–German Collection of Cell Microorganisms and Cell Cultures GmbH, D-38124 Braunschweig, Germany. [2] Institute of Hydrobiology, Biology Center CAS, Na Sádkách 7, 37005 České Budejovice, Czech Republic. [3] Microbial Evogenomics Lab, Limnological Station, Department of Plant and Microbial Biology, University of Zurich, Kilchberg, Switzerland. [4] Department of Biology, University of Konstanz, D-78457 Constance, Germany. [5] Department of Experimental Limnology, Leibniz Institute for Freshwater Ecology and Inland Fisheries, D-12587 Stechlin, Germany. [6] Institute of Biochemistry and Biology, Potsdam University, D-14469 Potsdam, Germany. [7] Berlin-Brandenburg Institute of Advanced Biodiversity Research, Free University, D-14195 Berlin, Germany. [8] State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China. [9] Department of Land Resources and Environmental Sciences, Montana State University, 334 Leon Johnson Hall, Bozeman, MT 59717, USA. [10] Institute of Physics, National Academy of Sciences of Kyrgyz Republic, Chui Avenue, 265-a, Bishkek 720071, Kyrgyzstan. [11] Institute of Microbiology, Technical University of Braunschweig, D-38108 Braunschweig, Germany.
*Corresponding author. Email: david.ngugi@dsmz.de

estimates of molecular divergence times (*16*), how does the inclusion of brackish and freshwater AOA species affect estimates of divergence times of major Nitrososphaerota clades in aquatic systems, especially given the conflicting results of previous studies that focused exclusively on marine and terrestrial AOA (*8*, *11*)?

Accordingly, we examined the diversity and evolutionary divergence of lacustrine Nitrososphaerota at an unprecedented scale using an extensive metagenomic dataset of brackish waters and freshwater lakes spanning five continents (including five extant Paleolakes and several large Eurasian lakes). Our results provide an integrated global overview of the diversification of planktonic Nitrososphaerota and reveal the unique evolutionary history of AOA in brackish and freshwater ecosystems.

RESULTS AND DISCUSSION

An expanded global-scale genomic database of lacustrine Nitrososphaerota

We compiled a global collection of 105 lacustrine metagenome-assembled genomes (MAGs) comprising 58 new Nitrososphaerota genomes (this study) and 47 previously reported MAGs (*17*–*21*). Accession numbers and associated metadata are provided in table S1. The MAGs are from independent metagenome samples derived from geographically diverse freshwater ($n = 16$) and brackish ($n = 6$) lakes as well as inland seas (Fig. 1A and table S1). These are from time-series metagenomic collections as well as surface (epilimnion) and bottom (hypolimnion) water samples of thermally stratified lakes (table S1). In addition, the studied lacustrine systems vary considerably in their trophic states, basin types, water residence times (~1 to 1000 years), and lifetimes (<1 to 30 Ma; table S2). The compiled global collection of planktonic lacustrine Nitrososphaerota genomes includes 77 freshwater MAGs and 28 MAGs from brackish lakes and inland seas (fig. S1 and table S1).

Phylogenomic inference based on a concatenated set of 122 protein families conserved in 301 archaea (data S1 and S2) revealed that the genomes of lacustrine Nitrososphaerota ($n = 105$) belong to five lineages (Fig. 1B), namely, *Nitrosopumilus* ($n = 73$), *Nitrosoarchaeum* ($n = 22$), *Nitrosotenuis* ($n = 3$), *Nitrosotalea* ($n = 2$), and a basal non-AOA Nitrososphaerota lineage ($n = 5$). Notably, the basal lineage includes Nitrososphaerota that lack the ammonia oxidation machinery and the AOA-specific carbon fixation pathway (*22*). Our results extend recent findings (*13*) of the recognized diversity of lacustrine AOA communities to environments spanning the continents of Europe, Asia, Antarctica, and North America, including the world's largest brackish (Caspian Sea) and freshwater (Laurentian Great Lakes) systems. Crucially, the reconstructed evolutionary history of Nitrososphaerota was robust to marker gene selection (fig. S2 and data S1 to S4), exclusion of redundant genomes ($n = 83$; details in table S1), and inference with complementary phylogenomic approaches (figs. S3 and S4 and data S5 to S7) or use of the conserved 16*S* ribosomal RNA (rRNA) marker gene (fig. S5).

Half of the MAGs ($n = 52$) formed a well-supported monophyletic clade within the genus *Nitrosopumilus* that was phylogenetically distinct from its closest brackish water relatives from the Black Sea and Baltic Sea (figs. S2 to S5). The clade includes representatives from 11 Eurasian freshwater lakes and, unexpectedly, from the northern Caspian Sea (1.1 to 1.2% salinity), which were sampled to a depth of 150 m (*17*). The *Nitrosopumilus*-like AOA genomes from Eurasian lakes are phylogenetically indistinguishable from

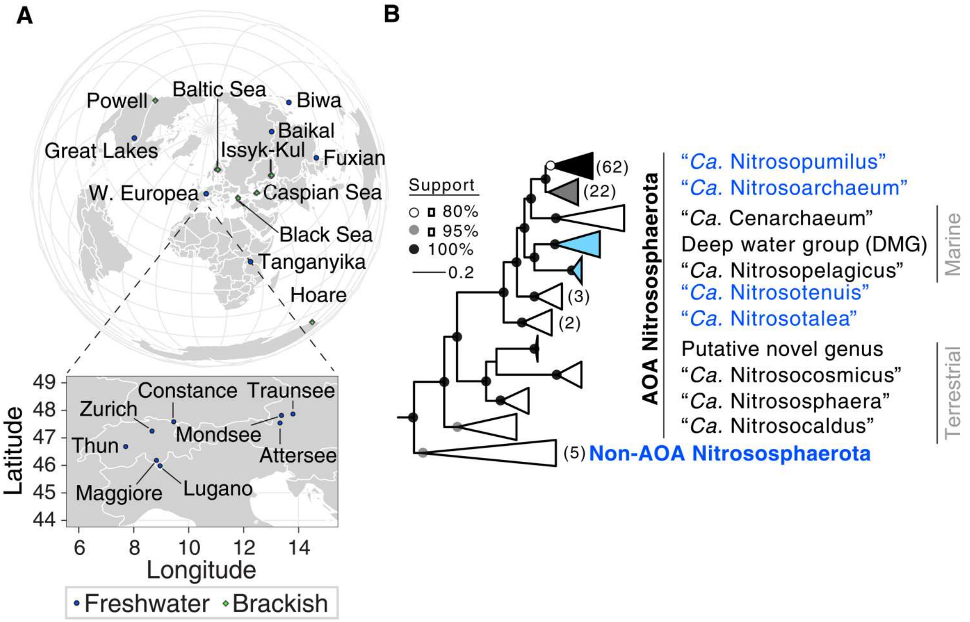

Fig. 1. Evolutionary history of planktonic Nitrososphaerota in lacustrine systems. (**A**) Metagenome sampling locations of freshwater lakes ($n = 14$) and brackish lakes/inland seas ($n = 6$) lakes, from which MAGs were reconstructed or publicly available. (**B**) Phylogenomic maximum-likelihood (ML) tree of Nitrososphaerota based on 122 conserved archaeal single-copy genes. The ML tree was rooted using 38 euryarchaeotal genomes following Ren *et al.* (*8*). Branches are collapsed to genus level for brevity. Values in parentheses indicate the number of genomes in lineages containing lacustrine Nitrososphaerota, with corresponding genus-level taxa highlighted in blue. The closed black circles indicate nodes with bootstrap support of ≥80% [Shimodaira-Hasegawa-like approximate likelihood ratio (SH)] and ≥95% (UFBoot). For a full-size tree, see fig. S2. The scale bar [in (B)] shows the number of amino acid sequence substitutions per variable site.

their Caspian Sea counterparts as reflected by the very short and isochronous branches in all inferred phylogenetic trees (figs. S2 to S5). This is remarkable because the Caspian Sea is the world's largest inland water body and has remained enclosed for nearly 5 Ma (*23*). Overall, this suggests that the ecological speciation of this genotype is relatively close in time, which, in turn, raises the question of whether the 52 geographically widespread lacustrine *Nitrosopumilus* MAGs represent a clonal population or separate subpopulations (i.e., subspecies) of the same species.

The predominant freshwater *Nitrosopumilus* species comprise four geographically distinct subpopulations of "*Candidatus* Nitrosopumilus limneticus"

To gain further insight into the evolutionary history of the predominant freshwater *Nitrosopumilus* clade, we performed genome-wide pairwise comparisons of the nucleotide sequences and predicted protein-coding genes from all available *Nitrosopumilus* genomes ($n = 108$) including those from freshwater, brackish inland seas, and marine environments (Fig. 2 and figs. S6 and S7). The results showed that the 52 lacustrine *Nitrosopumilus* MAGs from large Eurasian freshwater lakes and the Caspian Sea were genetically very similar and belonged to the same species, as reflected by the high average nucleotide identity (>99% ANI) and the high degree of gene conservation (Fig. 2 and fig. S6). This was also reflected in the high proportion of shared orthologous genes, which on average (±SD) accounted for 80 ± 11% of the predicted proteomes (Fig. 2), and the corresponding high average amino acid identity (AAI) of 97 to 99% (fig. S7). The lower pairwise AAI values between the 52 lacustrine *Nitrosopumilus* MAGs and their closest *Nitrosopumilus* relatives from brackish water (81 to 85% AAI; fig. S7) suggest that they are separate *Nitrosopumilus* species based on the operational AAI thresholds for genome-based taxonomic ranks (*24*). On the basis of these criteria, we found two additional geographically distinct new *Nitrosopumilus* species from meromictic Lake Hoare (in Antarctica) and monomictic Lake Issyk-Kul (in eastern Kyrgyzstan) that have only ~73 to 84% AAI to other sequenced Nitrososphaerota (figs. S2 to S7). This speciation is likely due to ecosystem isolation, natural selection under the influence of salinity, or both. Hereafter, we refer to the ubiquitous freshwater *Nitrosopumilus* species represented by the 52 highly similar MAGs as "*Ca.* N. limneticus," as recently proposed (*25*) in recognition of its membership in the genus *Nitrosopumilus* and its ubiquity in large freshwater lakes ("limneticus.")

The second largest group of brackish and freshwater Nitrososphaerota MAGs ($n = 22$) comprised four species of the genus "*Ca.* Nitrosoarchaeum" (figs. S2 to S7), with genome-wide intra-divergences averaging (±SD) 89 ± 6% AAI (range, 52 to 99.6%; fig. S7) and lower gene conservation levels (59 ± 16%; Fig. 2). Overall, these results suggest higher species diversity within the freshwater *Nitrosoarchaeum* lineage from different lakes, despite having similar niches to "*Ca.* N. limneticus".

The relatively isochronous phylogenomic branches within the *Ca.* N. limneticus clade (figs. S2 to S5) suggest that geographic location is irrelevant to the reconstructed evolutionary history of this freshwater AOA species. Accordingly, we performed genome-wide single-nucleotide sequence variant (SNV) analyses to dissect strain-level evolutionary processes and understand the population structure and biogeography of *Ca.* N. limneticus. Phylogenetic inferences from genome-wide SNVs found outside putative recombinant regions derived from a multiple genome sequence alignment [~362 kilo–base pairs (kbp)] of 40 medium- to high-quality *Ca.* N. limneticus MAGs with 1888 core genome SNVs (Table 1) yielded four geographically centered *Ca.* N. limneticus subclades (Fig. 3A). These four subclades are from the European perialpine lakes ($n = 29$), the Caspian Sea ($n = 4$), Lake Fuxian ($n = 5$), and Lake Baikal ($n = 2$). This is also supported by the high genetic similarity between natural populations of *Ca.* N. limneticus at different sites, as determined by a standard measure of genetic differentiation between populations, the fixation index (F_{ST}). On the basis of

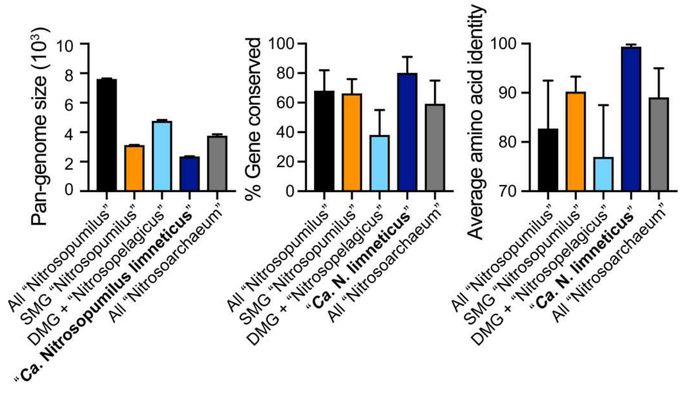

Fig. 2. High degree of genome conservation in *Ca.* N. limneticus genomes. Gene conservation among aquatic Nitrososphaerota genomes based on pairwise comparison of orthologous genes. Bar graphs show total pan-genome size (left), percentage of conserved proteins (middle), and AAI (right) of orthologous genes in the genomes of all *Nitrosopumilus* species ($n = 108$), separately in *Ca.* N. limneticus ($n = 52$), SMG *Nitrosopumilus* ($n = 12$), combined "*Ca.* Nitrosopelagicus" and Deep Marine Group (DMG) genomes ($n = 30$), and "*Ca.* Nitrosoarchaeum" ($n = 26$). Among the freshwater Nitrososphaerota, *Ca.* N. limneticus has a relatively smaller pan-genome (2413 genes) than *Ca.* Nitrosoarchaeum (3651 genes). The bars show mean values, and the error bars show the SD.

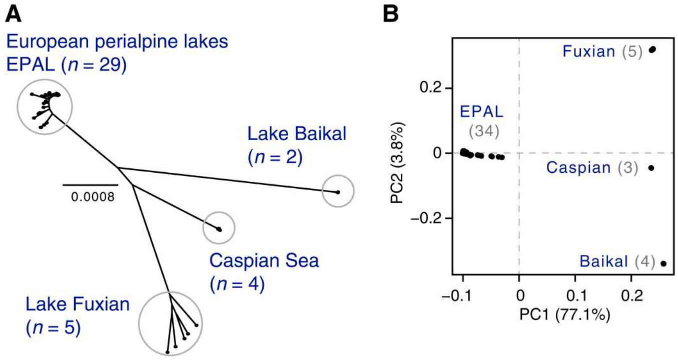

Fig. 3. Four distinct subpopulations of planktonic *Ca.* N. limneticus in freshwater lakes. (**A**) Genome-wide SNV-based ML tree (unrooted) derived from the multiple alignment of 40 *Ca.* N. limneticus MAGs. The scale bar shows the number of nucleotide sequence substitutions per variable site. (**B**) Principal components analysis of the fixation distance (F_{ST}) between lake samples based on pairwise allele frequencies of *Ca.* N. limneticus subpopulations in globally sampled freshwater metagenomes. F_{ST} measures the genetic differentiation between two populations based on their SNVs. The first two axes (PC1 and PC2) account for ~81% of the total variance. The number of metagenomes per cluster is given in parentheses.

Table 1. Mutation rates exceed recombination rates in freshwater *Ca*. N. limneticus. The ratio of recombination and mutation rates (R/θ) and the relative effect of recombination and mutation (r/m) for all *Ca*. N. limneticus MAGs and the different subclades measured using ClonalFrameML. n.d., not determined because the number of genomes is small.

		Subclade (subpopulation)*			
General information	**All genomes**	**European**	**Fuxian**	**Caspian**	**Baikal**
Number of genomes analyzed	40	29	5	4	2
Genome size range (Mbp)	0.83–1.44	0.83–1.17	1.04–1.44	1.04–1.13	1.04–1.16
Completeness (%)†	71.0–100	71.0–100	83.0–99.0	90.3–99.0	94.3–99.0
Contamination (%)†	0–3.9	0–1.0	0–3.9	0	0
% ANI (mean ± SD)‡	99.42 ± 0.48	99.81 ± 0.07	99.22 ± 0.54	99.94 ± 0.02	99.96 ± 0.01
Size of core genome aligned (kbp)§	361.62	211.83	854.27	1008.36	1016.88
Total SNVs	1888	192	3552	3786	1419
ClonalFrameML metric¶					
R/θ	0.48 (0.44–0.52)	0.40 (0.31–0.53)	0.26 (0.22–0.31)	1.12 (0.83–1.66)	n.d.
Mean length of imports (∂)	33 (31–35)	92 (73–111)	29 (26–32)	102 (81–121)	n.d.
Average distance of imports (ν)	0.036 (0.034–0.039)	0.024 (0.020–0.027)	0.091 (0.081–0.095)	0.015 (0.011–0.018)	n.d.
r/m#	0.57 (0.51–0.64)	0.86 (0.59–1.13)	0.67 (0.56–0.81)	1.71 (1.28–2.67)	n.d.

*Subclades defined by the core SNV genome phylogeny of 40 *Ca*. N. limneticus MAGs (see Fig. 1D). † Determined using CheckM (*71*). ‡Pairwise ANI determined using JSpecies (details in the Supplementary Materials). §Core genome alignment determined by progressiveMauve (details in Materials and Methods). ¶Parameter values indicate the estimated mean, with 95% confidence intervals shown in parenthesis. #$r/m = (R/\theta) \times \partial \times \nu$.

genome-wide SNVs obtained from mapping Eurasian lacustrine metagenomes ($n = 46$; table S3) against a high-quality reference *Ca*. N. limneticus MAG (LH-02apr19-284) from Lake Lugano (details in Materials and Methods), we found very low pairwise F_{ST} values between European perialpine lakes (0.001 to 0.307; mean ± SD, 0.095 ± 0.074), but distinctively higher F_{ST} values compared to freshwater lakes outside Europe (0.801 to 0.898; mean ± SD, 0.851 ± 0.027; table S4). Furthermore, samples were tightly clustered according to the geographic origin of lakes (Fig. 3B), strongly suggesting that geographic differentiation of *Ca*. N. limneticus populations in European perialpine lakes is extremely low. The low heterogeneity of *Ca*. N. limneticus populations is likely due to several factors, including common ancestry that is relatively close in time, recent selection, large populations with strong dispersal facilitated by hydrological connectivity between lakes (e.g., via the Rhine River), high gene flow via mobile genetic elements, or additional population constraints (e.g., extreme nutrient deficiency and climate change impacts) that act similarly in the sampled lakes.

Mutation is a major driving force in the diversification of *Ca*. N. limneticus in freshwaters

To gain further insight into the genetic diversification of *Ca*. N. limneticus, we examined the effects of homologous recombination with ClonalFrameML (*26*). First, we evaluated whether recombination rates derived from MAGs, which generally consist of consensus sequences from potentially multiple abundant strains of a species, reflect rates derived from genomes of isolates or single cells. For benchmarking, we used single-cell amplified genomes (SAGs) and MAGs of the ubiquitous freshwater clade LD12 and the marine cyanobacterium *Prochlorococcus* (table S8), and estimated both the relative frequency of recombination to mutation (R/θ) and the effect of recombination relative to mutation (r/m) in MAGs, SAGs, and isolates of the same lineage. As described in Supplementary Notes, we found that R/θ and r/m derived from MAGs were relatively consistent with estimates from SAGs and isolates, although the SAGs had much higher estimated contamination levels (table S9). SAGs also typically have elevated sequencing errors than MAGs (*27*).

The consistency of these results in conjunction with other factors such as the population coherence evidenced at the geographic scale (Fig. 3) coupled with the low number of sequence variants (~2 at 95% global nucleotide identity) of five independent universal maker genes (COG0012, COG0016, COG0215, COG0533, and COG0441) belonging to *Ca*. N. limneticus, detected from the gene catalog of lacustrine microbiomes with ~14 million nonredundant genes (see Materials and Methods), gives us confidence that R/θ and r/m can be estimated from the MAGs. Moreover, the dominance (95% of all reads) of a single sequence type (99% nucleotide identity) of *Nitrosopumilus* 16*S* rRNA genes over an annual cycle in Lake Constance (*28*) combined with the high pairwise nucleotide sequence similarity of MAGs from European perialpine lakes, including Lake Constance (99.5% ANI; Table 1), is also broadly consistent with how clonal complexes (>99% ANI) are designated in environmental and epidemiological studies based on isolates and MAGs (*29*, *30*).

Therefore, we determined recombination rates within the *Ca*. N. limneticus clade ($n = 40$ MAGs) and the geographically distinct subclades (Fig. 3A). The estimated R/θ for the entire *Ca*. N. limneticus clade [mean: 0.48; 95% confidence interval (CI): 0.44 to 0.52] was remarkably close to the estimates for the European (mean: 0.40; 95% CI: 0.31 to 0.53) and Fuxian (mean: 0.26; 95% CI: 0.22 to 0.31) subclades (Table 1), but three to four times lower than the estimate for the Caspian subclade (mean: 1.12; 95% CI: 0.83 to 1.66; Table 1). The lower recombination-to-mutation ratio (R/θ) for the European (0.40) and Fuxian (0.26) freshwater subclades (Table 1) is within the theoretical R/θ threshold (0.25 to 1) at which clonal divergence occurs (*31*, *32*). The measured effect of recombination relative to mutation (r/m) in these three subclades, accounting for the length and divergence of imported fragments (Table 1), was almost twice as high in the Caspian subclade (1.71) as in the European (0.86) and Fuxian (0.67) subclades. Neither evolutionary metric could be determined for the Baikal subclade because of the small number of MAGs ($n = 2$). The twofold lower r/m estimates in the freshwater subclades (Table 1) indicate a lower propensity for homologous recombination than in the brackish Caspian subclade. The higher r/m estimate (>1) indicative of increased homologous recombination in the Caspian subclade might reflect the higher salinity of the Caspian Sea (~1% salinity) (*17*) or its older age (2 to 5 Ma) compared to Eurasian lakes (~10,000 years old) (*23*).

Consistent with these interpretations, R/θ and r/m estimates derived from an alternative approach in which multiple freshwater metagenomes were separately mapped against a single *Ca.* N. limneticus reference MAG (details in Materials and Methods) averaged 0.074 (95% CI: 0.066 to 0.083) and 0.98 (95% CI: 0.86 to 1.05), respectively, similar to results based on MAGs (Table 1). Low r/m values are also reported for the freshwater clade LD12 (r/m, 0.14) (*33*) and *Escherichia coli* ET-1 group (r/m, 0.7) (*34*), with the former being 450-fold lower than their marine SAR11 counterparts (*34*). In a previous metagenome study, a single clonal sweep was also observed for a *Chlorobium* population over a 9-year period in a freshwater lake (*35*). Together, this suggests that recombination rates for these major freshwater taxa are also low or selection is extremely strong. Other studies also suggest that recombination may be less important than clonal diversification and genetic drift for some marine planktonic Nitrososphaerota (R/θ of 0.1 to 0.3) (*29*). Overall, the results suggest that mutation is an important driving force in the diversification of *Ca.* N. limneticus and other major freshwater prokaryotic species.

The exceptionally low genetic diversity of *Ca.* N. limneticus populations in freshwater lakes may be due to the relatively low evolutionary mutation rates in Archaea and the dependence of spontaneous mutation rates on temperature (*36*). This, combined with the fact that deep-water temperatures in most stratified temperate lakes are very low year-round (<5°C) (*37*), and the finding that native *Ca.* N. limneticus populations (at 4°C) have in situ growth rates [~0.012 generations per day (*25*)] that are 30 times lower than those of marine and brackish *Nitrosopumilus* species cultured under physiological environmental conditions (*38*), may contribute to a legacy of spatial homogenization of phylogenetic diversity. In this context, it is interesting that the evolutionary history of Nitrososphaerota shows that they gradually adapted to lower temperatures after the emergence of their thermophilic terrestrial ancestors (*8*–*10*). Low temperatures, such as those found in glacial climates, can constrain the evolution of species and diversity patterns, trapping them in a state of low genetic diversity [see, e.g., (*39*)]. Therefore, the evolution of freshwater AOA may reflect temperature limits on molecular evolution, growth rates, or a recent speciation and colonization of the inhabited lakes.

The divergence timeline of *Ca.* N. limneticus reflects the age of freshwater lakes

To gain more insights, we reconstructed the divergence timeline of the Nitrososphaerota with PhyloBayes (*40*) based on the maximum-likelihood phylogeny of the concatenated alignment of 122 protein families (data S5) and applied previous temporal clocks and root age priors for calibration (see fig. S4). The time-calibrated phylogeny constrains the Nitrososphaerota root of both AOA and basal non-AOA groups at approximately 2.77 Ga [95% highest posterior density (HPD) interval, 3.24 to 2.32 Ga; node 1 in Fig. 4]. The predicted mean age estimate for the root of the AOA (node 2, mean: 2.15 Ga, 95% HPD: 2.55 to 1.77 Ga; Fig. 4) is consistent with several recent molecular clock analyses (*8*, *41*), thus supporting the proposed emergence of AOA after the Great Oxidation Event (GOE; 2.3 to 2.4 Ga) (*8*). However, these results are inconsistent with those of Yang *et al.* (*11*), who used a non-Bayesian approach (i.e., RelTime). Specifically, Yang *et al.* (*11*) estimated that the root of Nitrososphaerota postdates the GOE (2.10 Ga) and that the AOA ancestors diverged from other non-AOA Nitrososphaerota around 1.17 Ga. The disagreement could arise from our extensive taxon sampling of Nitrososphaerota genomes across diverse habitats [as discussed in (*16*)], the additional temporal priors used by Yang *et al.* (*11*), or both. In either case, the accuracy of molecular clocks is constrained by the precision of the temporal priors used to calibrate the clocks (e.g., the root of the Archaea), which is independent of the models used for molecular dating—and remains uncertain in the absence of fossil data. Notably, the evolutionary time scales reported by Ren *et al.* (*8*) for the major nodes (nodes 1 to 4) could be reasonably reproduced with our Bayesian approach using their concatenated alignment of 77 protein families (n = 167 genomes; fig. S8, data S8, and table S5). Therefore, we emphasize the relative rather than the absolute timing of the divergence estimates.

Molecular dating results further suggested that AOA diversified into aquatic systems around 1.40 Ga (node 3, 95% HPD: 1.73 to 1.11 Ga; Fig. 4) before partitioning into distinct clades that were either exclusively marine (node 4, mean: 1.27 Ga, 95% HPD: 1.58 to 0.99 Ma; node 5, mean: 0.97 Ga, 95% HPD: 1.25 to 0.73 Ga; node 6, mean: 0.24 Ga, 95% HPD: 0.36 to 0.16 Ga; Fig. 4) or inhabited shallow nearshore marine, estuarine, or freshwater environments (node 7, mean: 0.67 Ga, 95% HPD: 0.89 to 0.48 Ga; node 8, mean: 0.81 Ga, 95% HPD: 0.11 to 0.61 Ga; Fig. 4). "*Ca.* Nitrosoarchaeum korense" MY1, enriched from rhizospheric soil (*42*), is a rare exception in node 7.

Within the exclusively marine planktonic AOA clades (node 4; Fig. 4), posterior age estimates indicate independent but relatively similar diversification times for the Deep Marine Group (DMG; node 9, mean: 0.93 Ga, 95% HPD: 1.22 to 0.68 Ga) and the SMG (node 10, mean: 0.74 Ga, 95% HPD: 0.97 to 0.55 Ga; Fig. 4) clades. This contrasts with the proposed sequential evolution of the SMG and DMG AOA clades (*11*), which is most likely the result of biased taxon selection that excluded lacustrine AOA species. Note that the SMG clade traditionally includes the "*Nitrosopumilus maritimus* cluster," which typically represents coastal and sedimentary AOA (*6*, *38*, *43*). In addition, we observed a much younger diversification age of the "Ca. Nitrosopelagicus" clade (node 11, mean: 0.48 Ga, 95% HPD: 0.79 to 0.25 Ga; Fig. 4), which is a sister clade of DMG (Fig. 4) and is considered a pelagic model of marine Nitrososphaerota (*44*). The discrete

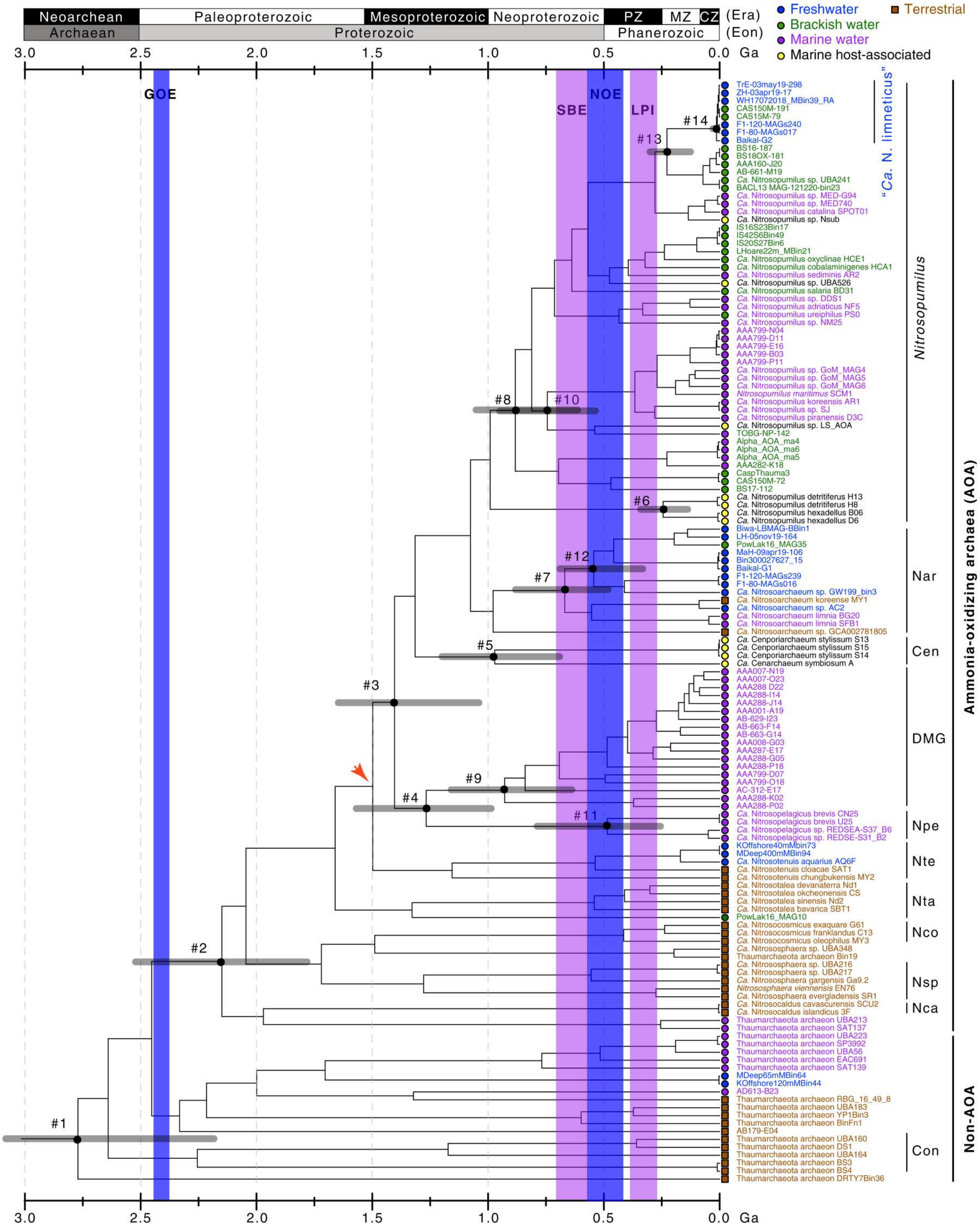

Fig. 4. Timeline of the divergence of *Ca*. N. limneticus in the evolution of Nitrososphaerota. Time-calibrated phylogeny of Nitrososphaerota based on a species tree of 218 high-quality genomes (fig. S4) inferred with four prior temporal constraints [including a gamma-distributed Archaea root prior of 3.95 ± 0.25 Ga; see (*41*, *80*)] and an autocorrelated relaxed clock model (see Materials and Methods for more details). Ages are in billions of years before present (Ga). Horizontal gray bars denote nodes mentioned in the main text and represent 95% CI. Genomes are color-coded based on the species habitat assignment. Vertical blue bars indicate the GOE (~2.33 Ga) (*83*) and the Neoproterozoic Oxidation Event (NOE; 635 to 420 Ma) (*47*), while purple bars indicate glaciation events: the "Snowball Earth" (SBE; ~720 to 635 Ma) (*46*) and the Late Paleozoic Icehouse (LPI; ~335 to 260 Ma) (*48*). Nar, *Nitrosoarchaeum*; Cen, "Cenarchaeum"; Npe, "Nitrosopelagicus"; Nte, *Nitrosotenuis*; Nta, "Nitrosotalea"; Nco, *Nitrosocosmicus*; Nsp, *Nitrososphaera*; Nca, "Nitrosocaldus"; Con, "Conexisphaera." Time scale: PZ, Paleozoic; MZ, Mesozoic; CZ, Cenozoic. The red arrowhead indicates the node with significant rate shifts (i.e., accelerated rates of speciation relative to ancestral lineages) for each of the eight estimated shift configurations (where there is rate heterogeneity) determined using the program BAMM (*88*, *89*) and accounts for 60% of the posterior distribution of rate shift configurations (see fig. S9).

phylogenetic placement and inferred independent diversification of DMG and SMG clades (Fig. 4), combined with their niche preferences (open ocean and nearshore environments, respectively) and unique genetic inventory [e.g., distinct adenosine triphosphatase (ATPase) gene families and osmoregulatory pathways] (*8*, *12*, *45*), emphasize that they comprise ecologically distinct marine Nitrososphaerota clades.

Regarding lacustrine AOA lineages, our dated phylogeny shows that *Ca*. Nitrosoarchaeum colonized freshwater systems much earlier (node 12, mean: 0.54 Ga, 95% HPD: 0.74 to 0.38 Ga) than the most recent common ancestor of *Ca*. N. limneticus and its sister lineage in brackish water (node 13, mean: 0.23 Ga, 95% HPD: 0.33 to 0.15 Ga; Fig. 4). The estimated divergence time (0.54 Ga) of the freshwater *Ca*. Nitrosoarchaeum clade (node 12; Fig. 4) is relatively consistent with recent estimates (363 to 216 Ma) (*13*). This suggests that its members evolved after the "Snowball Earth" (~0.72 to 0.66 Ga) (*46*) in an atmosphere that likely had oxygen levels close to present-day values during the Neoproterozoic Oxygenation Event (~0.64 to 0.42 Ga) (*47*). In contrast, the lacustrine lineage of *Nitrosopumilus* (node 13; Fig. 4) evolved during the Late Paleozoic Icehouse (~0.34 to 0.26 Ga) (*48*), which coincides with the period of the largest species extinction on Earth (Permian-Triassic, ~251 Ma) (*49*). In this context, note that predicted net diversification rates (speciation minus extinction) of Nitrososphaerota based on the time-calibrated tree (data S9) have gradually slowed over the last 750 Ma (Fig. 5), following a burst of diversification between 1.75 and 1 Ga (Fig. 5). Again, the sharp decline in net diversification rates occurred during global glaciations in the Neoproterozoic (730 to 580 Ma) (*46*). The extremely low divergence rate is further evidence of recent speciation that is likely the result of periodic selection. Biogeochemical and evolutionary mechanisms that may favor persistence of *Nitrosopumilus* species in low-temperature waters and during climatic shifts include the observed high biomass and activity of nitrifying communities under ice-covered conditions (<5°C) in freshwater lakes (*50*, *51*) and the ability to produce sufficient oxygen to sustain nitrification under anaerobic conditions (*52*).

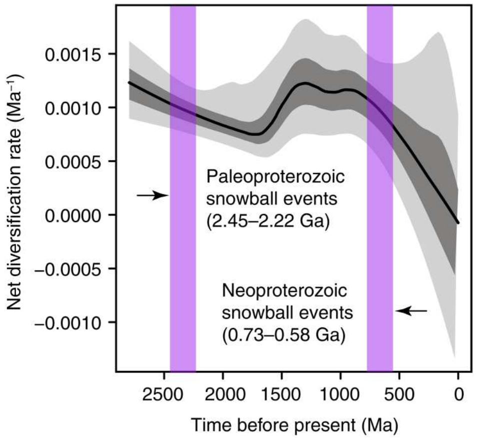

Fig. 5. Net diversification rates of planktonic Nitrososphaerota decreased substantially after glaciation events. The program BAMM (*88*, *89*) was used to determine and quantify the heterogeneity in evolutionary rates based on the time-stamped tree in Fig. 4. The thick black line represents the mean rate through time [in million years (Ma)], and the gray density shading represents (from the bottom) 5, 25, 75, and 95% credible intervals of the posterior distribution of rates through time. Purple vertical bars, delineated by arrows, denote Snowball Earth events separated by a 1.5-Ga gap with no evidence of glaciation at any latitude [see (*46*)].

The root age of the common ancestor of *Ca*. N. limneticus, whose members are widespread in Eurasian lakes (see below), is constrained to approximately 13 Ma old (node 14, 95% HPD: 24 to 8 Ma; Fig. 4). In contrast to recent findings (*13*), the timing of this diversification correlates remarkably well with the life-span range (>1 to 30 Ma) of the inhabited large Eurasian freshwater and brackish lakes (table S2), including ancient lakes such as the Caspian Sea (2 to 5 Ma) and Lake Baikal (>25 Ma) (*23*). However, the estimated divergence time of *Ca*. N. limneticus is an order of magnitude higher than predictions based on evolutionary rates in Archaea (see Supplementary Notes), but is remarkably consistent with the Miocene marine-freshwater transitions (24 to 5 Ma) of various modern freshwater faunas (*14*, *15*).

Extremely low global species diversity of planktonic freshwater Nitrososphaerota

The recovery of the same AOA species in geographically separated Eurasian lakes not only suggests a wide distribution of *Ca*. N. limneticus and its importance in Eurasian freshwater lakes but also raises questions about the global biodiversity of planktonic freshwater Nitrososphaerota. Accordingly, we examined the distribution of planktonic freshwater Nitrososphaerota by assessing the relative abundance of prokaryotic plankton in globally sampled lacustrine metagenomes (*n* = 157; table S3). For this purpose, we used the universally conserved single-copy marker gene COG0012 [a ribosome-associated guanosine triphosphatase (GTPase)], which was recently used in metagenomics to determine operational taxonomic units (i.e., mOTUs) at the species level (*53*). Details are described in the Supplementary Materials. Samples analyzed span the epilimnion (warm, photic layer) and hypolimnion (cold, aphotic layer) of thermally stratified lakes. Taxonomic profiling of the prokaryotic community based on the abundance of 2646 mOTUs (representing all COG0012 gene clusters with 95% global nucleotide sequence identity) in freshwater metagenomes revealed that Nitrososphaerota formed the major nitrifying community relative to the total prokaryotic plankton (Fig. 6A). In the hypolimnion, they were five times more abundant (mean ± SD, 10.7 ± 6.8%; *n* = 64; maximum, ~33%) than ammonia-oxidizing bacteria (mean ± SD, 2.0 ± 3.4%; *n* = 35; maximum, ~16%) and three times more abundant than nitrite-oxidizing bacteria (mean ± SD, 3.8 ± 2.6%; *n* = 69; maximum, ~14%). The abundance of Nitrososphaerota in the hypolimnion was also 10-fold higher than in the epilimnion (mean ± SD, 1.5 ± 1.2%; *n* = 27; maximum, ~5%; Fig. 6A), confirming results from other deep oligotrophic lakes obtained using different techniques [such as Catalysed reporter deposition-fluorescence in situ hybridization (CARD-FISH)] (*25*, *28*, *54*). The low abundance of nitrifying prokaryotes in the sunlit epilimnion may be caused by photoinhibition and sensitivity to reactive oxygen species produced by planktonic phototrophs (*55*, *56*). A similar contribution of active Nitrososphaerota of ~2 to 11% of the total prokaryotic community was found in the hypolimnion (85 m) of Lake Constance (fig. S10), based on recent metatranscriptomes (*n* = 10, two to three replicates) covering a single annual seasonal cycle (*25*). Overall, the results indicate that *Ca*. N. limneticus is an active and abundant component of the hypolimnetic prokaryotic community in European perialpine lakes.

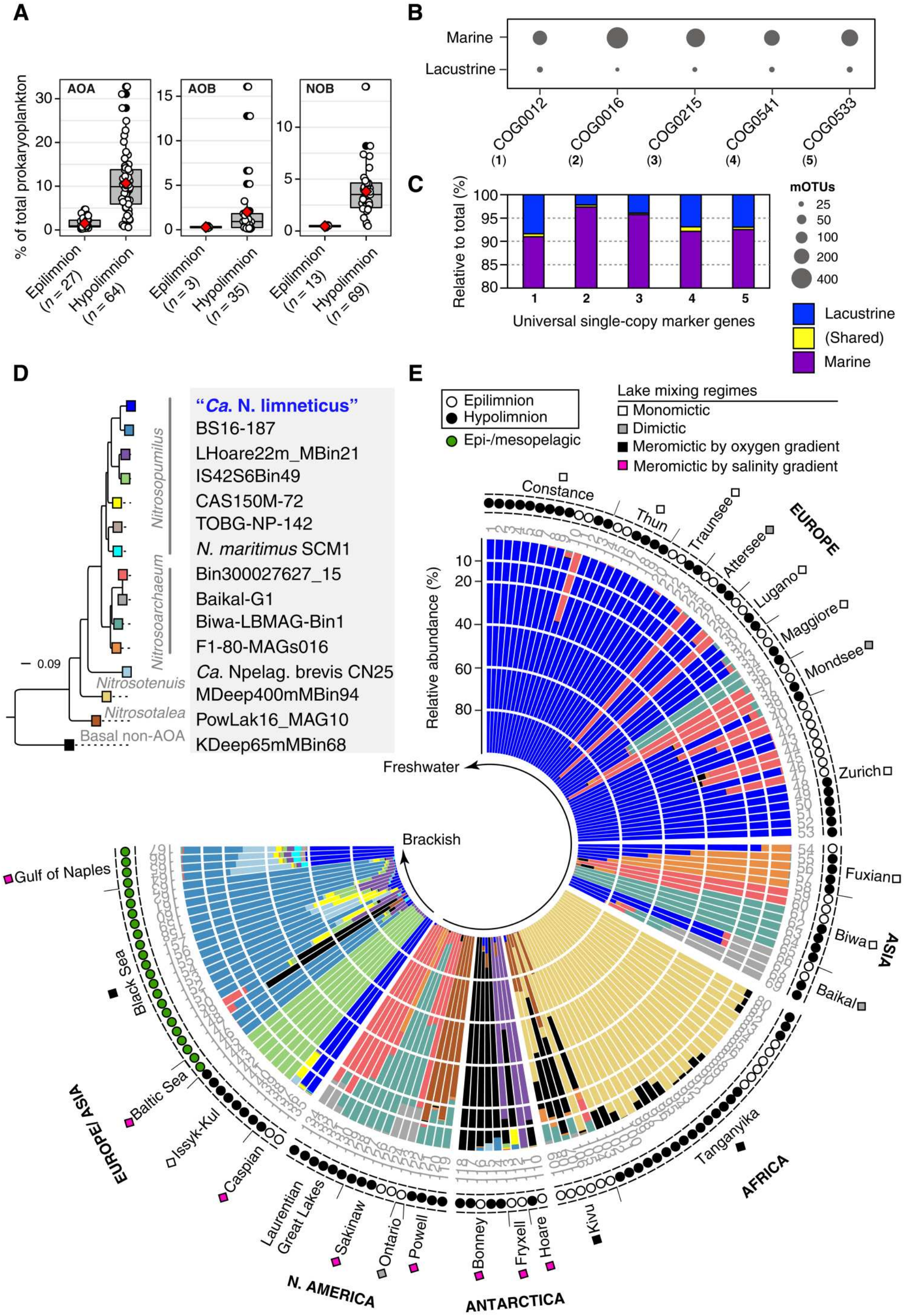

Fig. 6. Global species diversity and biogeography of planktonic Nitrososphaerota in lacustrine systems. (**A**) Relative abundance of AOA, ammonia-oxidizing bacteria (AOB), and nitrite-oxidizing bacteria (NOB) derived from the coverage of near-species counts of the universally conserved single-copy marker gene COG0012 in freshwater metagenomes (table S3). Sample counts are in parentheses; only metagenomes with relative abundances greater than 0.1% ($n = 91$) are shown. Abundances are normalized relative to other prokaryotic plankton in each metagenome. The boxplot shows the median as horizontal lines and the interquartile range as boxes (whiskers extend to 1.5 times the interquartile range). The mean is shown as a red-colored diamond. (**B**) Diversity of planktonic Nitrososphaerota inferred using mOTUs derived from massive integrated microbiome gene catalogs of marine and lacustrine systems. Delimitations were derived from five independent universal marker genes. (**C**) Contribution of lacustrine and marine mOTUs to global planktonic Nitrososphaerota diversity. Depending on the universal marker genes, the number of common Nitrososphaerota mOTUs is 1 to 2; all are conserved only between brackish and marine habitats. (**D** and **E**) Biogeography of 15 Nitrososphaerota genotypes (shown in panel D) in freshwater and brackish lakes based on genome-wide coverage (95% global identity) in globally sampled lacustrine metagenomes ($n = 157$; panel E). The species tree [in (D)] illustrates the phylogenetic relationship and assignment (in gray) of the 15 dereplicated Nitrososphaerota reference genomes with an ANI of 95% (i.e., at the species level). Metagenome IDs are enumerated for brevity; full details are in table S3.

Using four additional universal single-copy marker genes found in Nitrososphaerota genomes (table S6) and other prokaryotes (*53*), we estimated the diversity of planktonic Nitrososphaerota in aquatic systems almost to the species level and constructed a robust profile of prokaryotic mOTUs in massively integrated microbiome gene catalogs of oceanic (*57*) and lacustrine bacterioplankton (this study). Our global census of planktonic Nitrososphaerota diversity showed that species richness in lacustrine systems (mean ± SD, 12 ± 2 mOTUs) was 20 times lower than in the ocean (mean ± SD, 255 ± 100 mOTUs; Fig. 6B). Almost all lacustrine Nitrososphaerota mOTUs occurred exclusively in freshwater systems, with those that phylogenetically overlapped with marine Nitrososphaerota (1 to 2 mOTUs) being mostly brackish (fig. S11). Overall, our findings suggest that the diversity of planktonic Nitrososphaerota in freshwater lakes is exceptionally low compared with that in the ocean, consistent with the notion that oceans are "hotspots" of AOA diversity (*6*).

As described in Supplementary Notes, ancestral state reconstruction also shows that freshwater AOA lineages emerged more frequently from brackish than from marine ancestral lineages and evolved independently only within three AOA clades (*Ca.* N. limneticus, *Nitrosoarchaeum*, and *Nitrosotenuis*) and one basal non-AOA Nitrososphaerota clade (fig. S12). Of these transitions, however, only the genus *Nitrosoarchaeum* is species-rich (about four species; 95% AAI), but its members (and those of the genus *Nitrosotenuis*) emerged much earlier (~0.54 Ga; Fig. 4) than the *Ca.* N. limneticus clade (13 Ma old). The direct transition from marine to freshwater could be metabolically costly, as salinity stress affects the kinetics of ammonia oxidation in *Nitrosopumilus* strains from brackish water grown under freshwater or marine conditions (*38*, *43*). Overall, these results may help explain the low diversity of AOA in freshwater systems and suggest that brackish waters likely serve as bridging habitats for the transition between marine and Eurasian freshwater habitats.

Several factors can influence species diversity differently in freshwater and marine systems. For example, topographic complexity (e.g., reefs and kelp forests), depth, stratification, currents, large spatial extent, and age of the ocean are likely to provide more habitats and conditions conducive to species diversification. Conversely, freshwater lakes favor higher species diversity as they are spatially separated and extremely diverse, for example, in terms of available ions and nutrients and trophic and redox status. However, water masses in lakes have high turnover and are hydrologically connected, which can lead to homogenization of microbial communities. In addition, the effects of climate change are more pronounced in lakes than in the ocean, which has led to an unprecedented loss of biodiversity, particularly in lakes [reviewed in (*58*)].

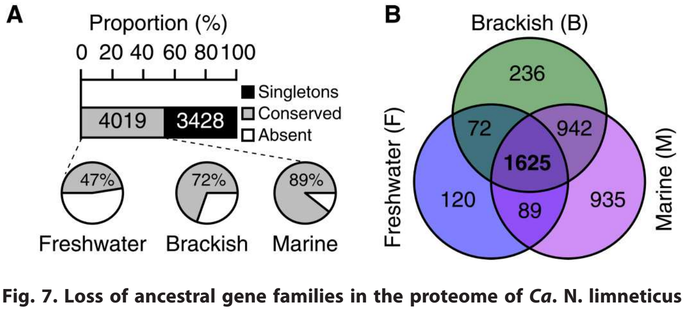

Fig. 7. Loss of ancestral gene families in the proteome of *Ca.* N. limneticus caused by the transition from marine to freshwater. (**A**) Number of gene families (i.e., orthogroups) in *Nitrosopumilus* species from different aquatic systems. A total of 7447 gene orthogroups were predicted, of which 4019 occurred in at least one genome in all three habitats. Pie charts show counts of these 4019 gene families in freshwater (*n* = 48), brackish water (*n* = 25), and marine (*n* = 35) species. (**B**) The Venn diagram shows the conservation of 1625 gene families in *Nitrosopumilus* species from the three habitats. A total of 120 gene families are unique to freshwater genomes (i.e., *Ca.* N. limneticus). Details of the genomes and gene occurrence profiles (including annotations) can be found in tables S1 and S7, respectively.

Ca. N. limneticus predominates over other planktonic Nitrososphaerota in Eurasian freshwater lakes

Next, we assessed the biogeography of *Ca.* N. limneticus in lacustrine ecosystems by comparing the prevalence of 15 high-quality reference Nitrososphaerota genomes (including 12 freshwater genotypes; Fig. 6D; details in the Supplementary Materials) in a global collection of lacustrine metagenomes (*n* = 157; table S3). Overall, we found a dependence of lacustrine Nitrososphaerota distribution on depth layer (epilimnion and hypolimnion) and geographic location at a global scale (Fig. 6E). Of note is the consistently predominant pattern of *Ca.* N. limneticus (Fig. 6E) in large lakes in Europe (relative abundance of 54 to 99%), which are relatively strongly hydrologically connected, and in two Asian lakes (relative abundance of ~14 to 86% in Fuxian and Baikal lakes), which include the world's most voluminous and oldest existing lake (*27*). However, freshwater AOA communities in Asian and North American lakes, including the Laurentian Great Lakes, were more diverse and dynamic. For example, abundances shifted with depth toward different *Nitrosoarchaeum* species in Asian (~15 to 99%) and North American (~10 to 94%) lakes (Fig. 6E). In contrast, the freshwater *Nitrosotenuis* species MDeep400mMBin94 dominated in the two African lakes (relative abundance of >90%; Fig. 6E), whereas the brackish *Nitrosotalea* species PowLak16_MBin10 and a novel brackish *Nitrosopumilus* species LHoare22m_MBin21 dominated in Lake Powell (southwestern USA) at various depths (76 to 89%) and in two Antarctic dry valley lakes (Hoare and Fryxell; 35 to 99%), respectively (Fig. 6E). These genome-wide abundance results also confirm the extremely low global diversity of freshwater planktonic Nitrososphaerota (Fig. 6, B and C) and are consistent with the absence of *Nitrosopumilus*-like 16*S* rRNA gene sequences in the Laurentian Great Lakes (*59*) and Lake Tanganyika genotypes (*60*). Overall, the findings indicate that AOA communities in Eurasian lakes are more similar to each other than in lakes in Africa, Antarctica, and North America, underpinning the influence of geographic separation on Nitrososphaerota evolution.

Proteome sequence alteration as a basis for freshwater colonization

Last, we focused on 108 *Nitrosopumilus* species from fresh (*n* = 48), brackish (*n* = 25), and marine (*n* = 35) waters and inferred orthologous gene families using OrthoFinder2 (*61*) to assess the genomic basis for the apparent clonality of *Ca.* N. limneticus populations in Eurasian freshwater lakes. Comparative genomics was performed as described in the Supplementary Materials. A total of 7447 gene families were predicted in the *Nitrosopumilus* genomes examined, half of which (4019) occurred in at least one genome within the three habitats (Fig. 7A). Of these, only ~47% (or 1906 gene families) were found in freshwater species, but accounted for ~72 and 89% in brackish water and marine species, respectively (including eight symbiotic species; Fig. 7A). Overall, the 1906 gene families found in freshwater species represent ~82% of the 2328 pan-genes of *Ca*. N. limneticus exclusively from freshwater (n = 48 genomes, excluding 6 from the Caspian Sea), suggesting that only 18% were genome-specific genes (or singletons), many of which were not annotated. This result, in turn, reveals a gradual loss of ancestral gene families during the transition of *Nitrosopumilus* species from marine to freshwater, reflected in the incidence of shared gene families (Fig. 7A) and the lower number of habitat-specific gene families (120 versus 935; Fig. 7B) and singleton genes (422 versus 1898) in freshwater compared to marine species.

A total of 1625 gene families were conserved in the species from all three habitats (Fig. 7B), representing 85.3% of the pan-genome of *Ca*. N. limneticus (table S7). These highly conserved gene families encode key carbon and energy metabolic systems of the marine/brackish ancestors (Fig. 7A and table S7). As observed elsewhere (*25*), *Ca*. N. limneticus can oxidize ammonia for energy, fix CO_2 via the conventional autotrophic 3-hydroxypropionate/4-hydroxybuyrate pathway, use alternative nitrogen and carbon sources such as urea, and generate adenosine triphosphate (ATP) using the A-type ATPase complex. *Ca*. N. limneticus is also able to synthesize vitamins (e.g., B12), amino acids, and cofactors and scavenge nutrients (e.g., phosphorus) and has protective mechanisms against ultraviolet damage (e.g., DNA photolyase and UvrABC endonuclease) and oxidative stress (table S7). Most of the gene families unique to *Ca*. N. limneticus (120 in total; Fig. 7B) have no predicted functions (table S7). As expected, *Ca*. N. limneticus cannot synthesize osmolytes such as ectoine and hydroxyectoine (table S7). However, *Ca*. N. limneticus encodes other mechanisms to cope with osmotic stress (e.g., the higher potassium concentration in freshwater), such as the Trk-type potassium efflux pump (table S7), similar to its marine and brackish *Nitrosopumilus* counterparts (Fig. 7A). In addition, *Ca*. N. limneticus has energetically more efficient proton-pumping NADH (reduced form of nicotinamide adenine dinucleotide):quinone dehydrogenase (NDH), which is optimal for low-salinity conditions and is common in freshwater bacteria (*62*). These results suggest that freshwater AOA are maladapted to the higher ionic strength (e.g., of Na^+, Cl^-, and Mg^{2+}) of seawater, as reflected in their proteome signatures (see below).

To address this latter suggestion, we examined the amino acid composition and isoelectric point (pI) of predicted proteomes (details in the Supplementary Materials) that might better reflect evolutionary genome development during the transition from marine to freshwater habitats. On the basis of the same 122 conserved single-copy genes (see above), we analyzed the amino acid composition of the predicted proteomes of the 108 *Nitrosopumilus* comprising species from fresh, brackish, and marine waters (table S1). Principal components (PC) analysis of amino acid frequencies (data S11) distinguished *Nitrosopumilus* genomes by habitat assignment (Fig. 8A). The amino acids isoleucine, aspartic acid, and valine account for 85% of the variability in PC1, whereas leucine, arginine, threonine, and asparagine together contribute 57% in PC2 (Fig. 8B). This is also reflected in the significant (Kruskal-Wallis test, $P < 0.05$) enrichment of charged residues (aspartic acid) in marine/brackish species and residues with hydrophobic side chains (isoleucine) and neutral residues (threonine and asparagine) in *Ca*. N. limneticus (Fig. 8C). Although biases in amino acid composition may be indirectly correlated with GC content, genome-wide GC content (30 to 35%) varied only slightly among *Nitrosopumilus* species (table S1). Together, these results suggest that residues in the core proteome of *Nitrosopumilus* species carry signatures that reflect their ecological lifestyle and underscore the strength of environmental selection in reshaping the proteome across the marine-freshwater boundary.

In addition, we examined the pI of predicted proteomes (data S12) but restricted the analysis to integral membrane proteins with a single transmembrane domain. This is because extracellular proteins evolve much faster than intracellular proteins (*63*) and are thought to interact closely with the extracellular environment (*45*). Moreover, the pI value is the biochemical property of a protein at which pH it has no net charge. The pI value is highly dependent on the solution (buffer) pH, implying that functional and evolutionary links exist between environmental properties (pH) and protein pI values in *Nitrosopumilus* species inhabiting the three distinct aquatic habitats. The proportion of genes encoding integral proteins relative to the total number of predicted proteins varied significantly by species habitat (Kruskal-Wallis test, χ^2 = 18.495, df = 2, $P = 9.636 \times 10^{-5}$). They were lowest in freshwater (mean ± SD, 7.7 ± 0.4%; n = 48), intermediate in brackish water (mean ± SD, 8.4 ± 1.4%; n = 25), and highest in marine species (mean ± SD, 9.1 ± 1.6%; n = 27). In addition, the predicted median pI value of these proteins was significantly shifted in *Nitrosopumilus* species as a function of habitat assignment (Kruskal-Wallis test, χ^2 = 57.328, df = 2, $P = 3.559 \times 10^{-13}$; Fig. 8D). Values were highest in freshwater species (median, 4.77; n = 48), intermediate in brackish water species (median, 4.64; n = 25), and lowest in marine species (median, 4.51; n = 27; Fig. 8D). In addition, the frequency distribution of pI was relatively unimodal in marine species (Hartigan's dip test, P = 0.589), in contrast to lacustrine species with two maxima (Fig. 8D). Overall, these results show that the transition from marine to freshwater in the genus *Nitrosopumilus* was accompanied by substantial amino acid substitutions in the membrane-anchored proteome.

To summarize, we leveraged metagenomes from geographically separated and ecologically diverse freshwater lakes to provide an integrated genome-wide view on the evolution and divergence of planktonic lacustrine Nitrososphaerota on a global scale. We found that the species diversity of planktonic Nitrososphaerota in freshwater lakes (ca. 15 detected "species") was 20-fold lower than in marine ecosystems. However, the abundance of Nitrososphaerota in freshwater lakes [this study and others; e.g., (*25*, *28*, *54*)] is similar to that in the ocean [20 to 40% of prokaryotic plankton; discussed in (*3*)]. This lower diversity is likely the result of a relatively recent habitat transition of planktonic Nitrososphaerota into freshwater systems or due to population bottlenecks associated with the transition from ocean to freshwater, such as adaptation to low salinity (*64*). Thus, the diversity of aquatic Nitrososphaerota could be related to ecosystem age, constrained by environmental conditions, or the interaction of both factors.

Our results demonstrate the emergence of a clonal freshwater AOA species and its propagation in major European perialpine lakes. Theoretically, geographically isolated lakes are expected to harbor locally evolving species—in the absence of disturbance (e.g., human intervention), resulting in restricted gene flow, genotypic variation, and a high proportion of endemic freshwater microorganisms. This is true for freshwater lakes in Africa (Tanganyika

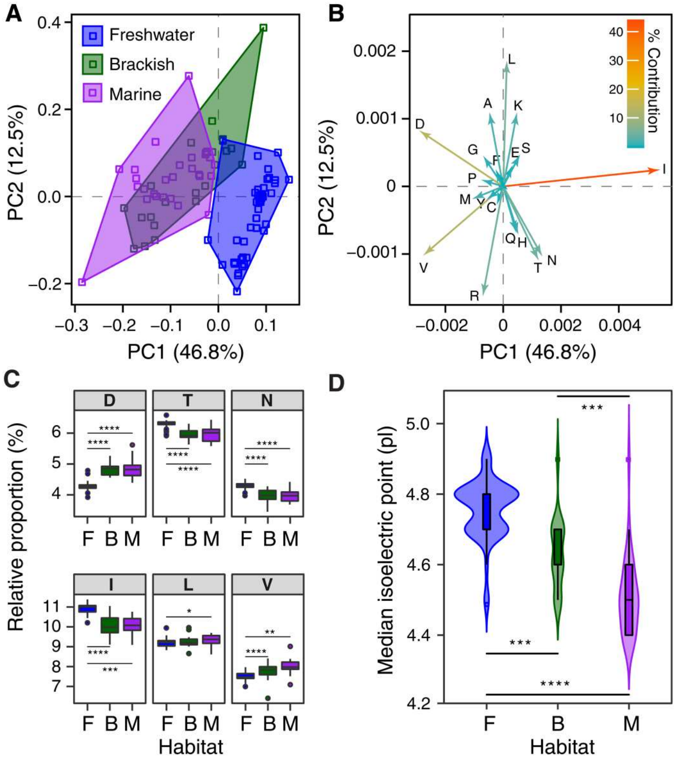

Fig. 8. Marine-freshwater transition prompted extensive amino acid substitutions in the proteome of *Ca.* N. limneticus. (**A**) Amino acid composition profile of 122 core genes in 101 *Nitrosopumilus* genomes (excluding eight host-associated marine species) separates species into discrete ecological clusters using principal components analysis. Polygons show 95% CIs for a set of species in each habitat group. The two principal components explained 59.3% of the total variance. (**B**) The amino acids isoleucine (I), aspartic acid (D), and valine (V) account for 55, 16, and 14% of the variability in PC1, respectively, while leucine (L), arginine (R), threonine (T), and asparagine (N) account for 24, 18, 8, and 7% of the variability in PC2, respectively. (**C**) These residues are also significantly differentially enriched in freshwater and marine species. Boxplots show the median as middle horizontal line and interquartile ranges (IQRs) as boxes (whiskers extend no further than 1.5 times the IQR). (**D**) Distribution of the median pI of putative integral proteins in ecologically assigned *Nitrosopumilus* genomes. Within violins, boxes correspond to the first and third quartiles of the distribution, while a thick horizontal line shows the median, and whiskers extend to extremes no further than 1.5 times the IQR. Asterisks show the adjusted significant *P* values of the unpaired two-sided Wilcoxon test with Benjamini-Hochberg-Yekutieli correction (**P* < 0.05, ***P* < 0.01, ****P* < 10^{-3}, and *****P* < 10^{-6}). F, freshwater; B, brackish water; M, marine water.

and Kivu), Antarctica (e.g., Hoare and Bonney), and exemplary inland brackish lakes (e.g., Issyk-Kul). In the major perialpine lakes of Europe, however, the predominant (and often the only) AOA evolved from a common brackish ancestor and underwent limited diversification over millennia. The oldest lakes in the world such as the Caspian Sea (2 to 5 Ma) and Lake Baikal (>25 Ma) (*23*) were found to harbor the same *Nitrosopumilus* genotype as relatively young Eurasian lakes [10,000 years old (*23*)], despite differences in age, salinity, and spatial separation (3400 to 6500 km). These results contrast with the high microbial species turnover observed at smaller geographic scales in the ocean (*65*). They also challenge the view that very few AOA species can dominate the overall diversity of individual ecosystems (*3*, *6*). Curiously, despite their extraordinarily low genetic diversity, planktonic freshwater AOA exhibit nitrification rates comparable to or many times higher than those of ocean gyres [discussed in (*25*)], converting up to 11% of nitrogen assimilated via primary production (*25*). This, in turn, emphasizes the ecological importance of *Ca.* N. limneticus at the ecosystem level.

Freshwater systems are among the ecosystems undergoing the greatest and most rapid changes because of climate change (*37*), and the pace of these changes has resulted in unprecedented loss of biodiversity (*58*). Recent records show, for example, a decline of up to 80% in reference freshwater vertebrate fauna over the past three decades, many of which are threatened with extinction (*58*). Although low genetic diversity is a hallmark of highly endangered species, extinction of freshwater AOA is less likely because, as shown in our study, they are widespread, accounting for up to 30% of microbial biomass (*25*), and, like other planktonic prokaryotes, have minimal dispersal constraints due to their small size. However, among prokaryotes, ecologically specialized species (such as chemolithoautotrophic AOA) are more likely to face extinction under disturbed conditions than prokaryotes that are generalists [discussed in (*66*)]. Some studies show that AOA are more sensitive to disturbance (e.g., eutrophication) than their bacterial counterparts [see, e.g., (*67*)]. Our findings therefore have implications for the role of Nitrososphaerota in freshwater energy and nutrient fluxes, for the trophic interactions they support, and ultimately for the function of freshwater lakes as sources of drinking water.

MATERIALS AND METHODS

Selection of metagenomes

The metagenomes used in this study are those generated as part of this study and those obtained from the European Nucleotide Archive (ENA). For ENA metagenomes, only those published in a primary publication or for which the principal investigators had given their consent were used. The full list of the studies and samples are provided in table S3, including accession numbers and corresponding metadata (when available). Detailed information on sample collection and metagenome sequencing can be found in the Supplementary Materials for several European perialpine lakes, Lake Issyk-Kul, and Antarctic dry valley lakes (Hoare, the Westlobe of Bonney, and Fryxell). Metagenomes were preprocessed and assembled into contigs using metaSPAdes v3.15.2 (*68*), as described by Duarte *et al.* (*57*), or MEGAHIT v1.1.4-2 (*69*), as described in the Supplementary Materials.

Genome-resolved metagenomic binning and additional reference genomes

The resulting size-filtered metagenomic contigs were mapped to the corresponding error-corrected metagenomic reads using BBMap v38.90 (https://github.com/BioInfoTools/BBMap) with default settings except the "pairedonly=t" option. The resulting mean base coverage per contig was used for unsupervised metagenome-resolved genome binning using MetaBAT2 v2.12.1 (*70*) with default settings and a minimum contig size of 1.5 kbp. The quality of MAGs was assessed using CheckM v1.1.3 (*71*) with the "lineage_wf" command and were taxonomically assigned with GTDB-Tk v1.3.0 (*72*) using the "classify_wf" command with default settings and GTDB database version 95 (accessed 23 July 2020).

Additional reference genome drafts of marine and terrestrial AOA were retrieved from three dedicated databases (accessed May 2018 and June 2020), including the National Center for Biotechnology Information (NCBI; www.ncbi.nlm.nih.gov), the Integrated Microbial Genomes and Microbiome (IMG; https://img.jgi.doe.gov) database, and the National Genomic Data Center (NGDC; https://bigd.big.ac.cn). These reference genomes were also examined using CheckM and GTDB-Tk (as described above) to check their completeness and validate their taxonomic assignment. All genomes used in this study are listed in table S1 (n = 301), where additional information can also be found.

Subsequently, all MAGs and genomes that were assigned to the phylum Nitrososphaerota and had completeness of at least 25% and contamination of <10% were selected for further analyses (see list in table S1). For the initial phylogenomic analyses, this level of completeness was used to obtain divergent clades of Nitrososphaerota before further analyses were performed as described below or as described in the Supplementary Materials, including analysis of genome-wide ANI and AAI, comparative genomics based on orthologous gene families, and pI analysis of predicted proteomes of *Nitrosopumilus* species.

Draft MAGs of Nitrososphaerota from brackish and freshwater lakes have been deposited at NCBI under BioProject number PRJNA820565 and ENA BioProject number PRJEB35640, while metagenomes have been deposited in the Short Reads Archive under the accession numbers provided in table S3.

Nitrososphaerota species tree inference using concatenated marker genes and 16*S* rRNA gene phylogeny

All 301 genomes (table S1) were used to construct an initial maximum-likelihood tree (fig. S2) based on a concatenated protein alignment of 122 conserved archaeal single-copy genes (data S1). Archaeal conserved genes were predicted with GTDB-Tk v1.3.0 (*72*) using the GTDB database (release 95; accessed 23 July 2020) and the "classify_wf" command with default settings. A maximum-likelihood tree (dataset S2) was then constructed from the concatenated protein alignment with IQ-TREE v2.0.6 (*73*) using the best-fitting LG+F+R10 model, and confidence was assessed by ultrafast bootstrapping (1000 iterations).

In parallel, a second complementary phylogenomic tree was inferred from an independent set of 43 conserved single-copy marker genes obtained with CheckM v1.1.3 (*71*) using the "lineage_wf" workflow. Phylogenomic inference was performed from the resulting concatenated protein sequence alignment of 301 genomes (data S3) with IQ-TREE (*73*) using the best-fitting LG+F+R9 model with 1000 ultrafast bootstraps. For visualization, the resulting trees were visualized in FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) and rooted using the 38 Euryarchaeota genomes (see list in table S1) following the protocol of Ren *et al.* (*8*). Crucially, the independent phylogenomic trees inferred from the 122 and 43 single-copy marker genes (data S2 and S4) were highly supported and confidently maintained the topologies of the major lineages (fig. S2), as determined using the "cophylo" command in the R package "PhyTools" v1.0-1 (*74*).

After these tests, we reduced our original genome dataset from 301 to 218 genomes of medium to high quality [following Bowers *et al.* (*75*)] to improve the run times of the computationally intensive molecular dating analyses subsequently performed. The 218 representative genomes were selected by excluding redundant species after we inspected adjacent branches/leaves in the phylogenomic trees (figs. S2 and S3) and considered genome completeness (average ~85%) and estimated contamination levels (maximum 10%), while retaining all archaeal reference genomes (n = 78) from Ren *et al.* (*8*). The 218 genomes are highlighted in table S1. From these 218 genomes, we extracted and constructed a concatenated alignment of 122 protein families (data S5) using GTDB-Tk (*72*), followed by inference of a maximum-likelihood tree (data S6) using IQ-TREE, as described above.

A complementary Bayesian-based phylogenomic tree (data S7) was constructed using the same concatenated alignment with 218 genomes using PhyloBayes v4.1 (*40*) with the CAT-GAT+Γ model, which is considered robust and computationally viable for large phylogenomic datasets (*76*). Two independent Markov chains were run until both reached convergence. The chains were run until the effective sample size for each parameter was >100, and the relative difference between the parameters of each run was <0.3. The burn-in for each chain was 100 points, and each chain was sampled every 10 points up to 30,200 points. Markov chain Monte Carlo (MCMC) convergence was evaluated using the "tracecomp" and "bpcomp" commands in PhyloBayes (*49*). The majority consensus tree resulting from the Bayesian analysis retained the evolutionary relationships already supported by the maximum-likelihood tree (fig. S3), as determined using PhyloTools (*40*). Unless otherwise indicated, the maximum-likelihood tree of 218 genomes (fig. S4) is reported in the main text and used for both divergence time estimation and ancestral state reconstruction analysis (see below for details).

The 16*S* rRNA gene sequences encoded in the 301 genomes were extracted using Barrnap v0.9 (https://github.com/tseemann/barrnap). Of these, 185 genomes encoded at least one gene sequence with lengths ≥ 1 kbp. When a genome carried multiple copies (for 35 non-Nitrososphaerota genomes), only the longest sequence was considered. The predicted 16*S* rRNA gene sequences were aligned using MAFFT-linsi v7.407 (*77*) (option: --adjustdirection) before trimming the resulting alignment using trimAl v1.4.rev15 (option: -gappyout) (*78*) to remove ambiguous positions. The filtered alignment with 1472 nucleotide positions was used to build a maximum-likelihood tree (fig. S5) with IQ-TREE (-m TESTNEW -bb 1000 -alrt 1000 -bnni -safe -mset GTR) under the best-fitting GTR+F+R5 model.

Molecular dating of Nitrososphaerota divergence

The time of divergence of *Ca*. N. limneticus in the evolutionary history of Nitrososphaerota was estimated using the Bayesian dating approach in PhyloBayes v4.1c (*40*) based on the concatenated alignment with 122 protein families (data S5) from 218 genomes of medium to high quality (see full list in table S1). Bayesian analysis of molecular dating was fixed to the corresponding maximum-likelihood tree derived from the concatenated alignment (fig. S4 and data S5), using both the full CAT and reduced CAT20 substitution models separately, each in combination with the following options: the relative exchange rate LG (-lg), a lognormal autocorrelated relaxed clock (-ln), a birth-rate process (-bd), and four gamma categories (-dgam 4). To compare posterior ages with estimations obtained without the information from the concatenated protein sequence alignment, we generated age estimates under a prior process (-prior) in PhyloBayes (*40*). The LG model is based on the best-fitting model derived using IQ-TREE (*73*) after reconstructing the maximum-likelihood tree. On the other hand, the autocorrelated clock model was chosen because the maximum-likelihood tree is significantly autocorrelated (CorrScore = 0.90051, $P < 0.01$) according to CorrTest (*79*). This, in turn, implies that the inferred Nitrososphaerota phylogeny is a function of the temporal lag between ancestors and descendants. In the main text, we report only the CAT20-based results because no significant differences in mean ages of major nodes were observed between the full CAT and the CAT20-based analyses (table S5).

Four nodes in the phylogenomic tree were used for time scaling. These were based on the temporal priors used previously to estimate the divergence of marine Nitrososphaerota (*8*) and major prokaryotic and eukaryotic lineages (including Nitrososphaerota) (*41*). Figure S4 shows labeled nodes with molecular calibration points that were used: (i) The root node of Archaea was set with a gamma-distributed root prior of 3.95 ± 0.25 Ga adopted from (*41*, *80*)—so that the root age is midway between the youngest root of 3.8 to 2.7 Ga (*81*) and the most ancient root of 4.38 to 3.35 Ga used to estimate the evolution of methanogens (*80*); (ii and iii) the inferred independent origin of the Crenarchaea (*Thermoproteales* and *Sulfolobales*) and *Thermoplasma* lineages after the GOE of 2.33 Ga (*82*, *83*) on the basis that members of these discrete clades use oxygen as a terminal electron acceptor (*84*); and (iv) the inferred age of the most recent common ancestor (~0.475 Ga) of the Crenarchaeota *Sulfolobus sulfotaricus* and *Sulfolobus islandicus*, which have chitinases thought to have arisen after the evolution of lignin-producing plants (*81*).

Two independent MCMC were run in parallel until they reached convergence based on a comparison of their posterior distributions with the "tracecomp" program implemented in PhyloBayes (effective sizes of >100 and maximum discrepancy between chains of <0.3; fig. S7). The chronogram of the species tree was generated using the "readdiv" command in PhyloBayes (*40*). MCMC chains were sampled every 10 cycles after a burn-in of 500 points from the first generations. The results of the parallel analyses performed with the CAT20 substitution model were consistent with the full CAT model, as indicated by the similarities in the age estimates of the major Nitrososphaerota groups (table S5). Tree chronograms of species under the CAT20 model and the full CAT model are provided in data S9 and S10, respectively.

A similar Bayesian divergence estimation analysis was performed on the basis of the concatenated alignment of the 77 protein families (data 8) by Ren *et al.* (*8*) (n = 167 genomes) to test the robustness of our molecular estimation approach with respect to the mean age estimates for the nodes of greatest interest compared with the original results derived using MCMCTree (*8*). Accordingly, a maximum-likelihood tree was inferred with IQ-TREE v2.0.6 (*73*) under the best-fitting LG+F+R9 substitution model (1000 ultrafast bootstraps). Next, a molecular dating analysis was performed with PhyloBayes using the above temporal constraints, the reduced CAT20 substitution model, the relative exchange rate LG (-lg), a lognormal autocorrelated relaxed clock (-ln), a birth-rate process (-bd), and four gamma categories (-dgam 4). The resulting time-calibrated tree reasonably reproduced the previous divergence times (fig. S8 and table S5), suggesting that procedural effects in the inferred *Nitrososphaerota* divergence times between PhyloBayes and MCMCTree were negligible and that the differences were mainly due to the expanded dataset in our study.

Genome-wide sequence variants and population structure of *Ca*. N. limneticus

An alignment of the core genome was used to infer genome-wide SNVs and to reconstruct phylogenetic relationships among the *Ca*. N. limneticus genomes (n = 40) at the strain level. This also allowed us to estimate the relative impact of homologous recombination on the genetic diversification of *Ca*. N. limneticus populations in the sampled freshwater systems. Because prediction of core genome size depends on the size and quality of the genomes sampled, we based our analysis on a set of 40 *Ca*. N. limneticus MAGs with a total length of at least 70% of the median length (1.1 Mbp) of *Ca*. N. limneticus MAGs (table S1). The selected MAGs represented 80% of the 52 *Ca*. N. limneticus genomes, originating from Eurasian lakes (including Baikal and Fuxian) and the Caspian Sea.

To measure recombination rates in the *Ca*. N. limneticus population, we created a multiple whole-genome alignment of the 40 MAGs with progressiveMauve v2.4.0 (*85*) using default settings. The core genome alignments longer than 500 bp were then extracted using the "stripSubsetLCB" script provided by Mauve (*85*). The concatenated core genome alignment was then used to build a maximum-likelihood tree using PhyML v3.3.20211231 (*86*) with the options "--datatype nt -p --bootstrap 100 --model GTR -f m --ts/tv e --alpha e --quiet --leave_duplicates." With these datasets, ClonalFrameML v1.12 (*26*) was used to infer the phylogeny of the genomes based on SNVs outside of recombination and to estimate both the relative frequency of recombination to mutation (R/θ) and the relative effect of recombination to mutation (r/m) for the whole *Ca*. N. limneticus population and for each subclade defined in the phylogeny. Note that R/θ and r/m represent different evolutionary metrics because R/θ ignores the length and nucleotide diversity of the imported fragments and therefore does not provide information on the actual effect of recombination on evolutionary change (*34*). The recombination parameter r/m is calculated as the product of $(R/\theta) \times \partial \times \nu$, where ∂ is the average length of the imported sequences and ν is the average nucleotide divergence of the imported sequences [see (*26*)]. In parallel, we investigated whether these recombination rates can be reliably estimated from MAGs by comparing estimates derived from MAGs, single-cell genomes, and isolates from two ubiquitous aquatic bacterial species (table S8). These results are summarized in table S9 and discussed in Supplementary Notes.

Using an alternative approach, we also quantified recombination rates based on the core SNV alignment, which we obtained by mapping independent freshwater metagenomes against the same *Ca*. N. limneticus reference MAG (LH-02apr19-284). The reference has a size of 1.16 Mbp and was selected for its high completeness (100%), lack of contaminants, low number of contigs (26), and high N_{50} (~0.6 Mbp) compared to other genomes (table S1). We inferred R/θ and r/m of native *Ca*. N. limneticus populations based on the metagenomes (n = 60) used to reconstruct medium- to high-quality MAGs by mapping them separately against the reference MAG using Snippy v4.6.0 (https://github.com/tseemann/snippy). Snippy was run with default settings (including the minimum site coverage depth of 10 for allele determination). Twenty-five of the metagenomes provided sufficient coverage (10×) for SNV analysis, and from these, a core genome alignment was created with common variants sites present in at least two samples, representing the core SNV genome of natural *Ca*. N. limneticus populations in each lake. Ambiguous gaps were trimmed from the core SNP genome alignment using trimAl v1.4.rev15 (*78*) with the "-gappyout" option, followed by inference of a maximum-likelihood tree using PhyML and estimation of recombination rates using ClonalFrameML as described above.

Genetic differentiation (population structure) of natural *Ca*. N. limneticus populations in different Eurasian freshwater lakes was also estimated based on the F_{ST}. POGENOM v0.8.3 (*87*) was used to estimate F_{ST} based on pairwise comparisons of allele frequencies in freshwater metagenomes mapped to the same representative genome (LH-02apr19-284) as above. Only metagenomes with at least 10-fold average coverage and a minimum breadth of coverage to the reference genome of 40% were considered (46 of 157).

Statistics

Multivariate statistics were performed in the R environment (www.R-project.org/) using the packages "rcompanion" v2.4.0 and "rstatix" v0.6.0, and results were plotted using "ggplot2" v3.3.3. The Kruskal-Wallis test was performed to examine differences in the median isolectric point (*p*I) of *Nitrosopumilus* species assigned to different habitats. Significant differences were then assessed with the nonparametric Mann-Whitney *U* test for pairwise comparisons using the Benjamini-Hochberg-Yekutieli correction to control for false discovery rate (α = 0.05). All *P* values less than 0.05 were considered significant. Frequency distributions of the median pI that deviated from unimodality were tested with Hartigan's dip test statistic with "diptest" v0.75-7 using simulated *P* values based on 2000 bootstrap replicates. pI distributions with *P* values less than 0.01 were considered significantly unimodal. Principal components analysis biplots were generated using the R packages "reshape2" v1.4.4, "ggbiplot" v0.55, ggplot2, "ggfortify" v0.4.11, and "ggConvexHull" v0.1.0.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S12
References

Other Supplementary Material for this manuscript includes the following:

Tables S1 to S9
Data S1 to S12

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. C. Brochier-Armanet, B. Boussau, S. Gribaldo, P. Forterre, Mesophilic crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* **6**, 245–252 (2008).
2. A. Oren, G. M. Garrity, Valid publication of the names of forty-two phyla of prokaryotes. *Int. J. Syst. Evol. Micr.* **71**, (2021).
3. D. A. Stahl, J. R. de la Torre, Physiology and diversity of ammonia-oxidizing Archaea. *Annu. Rev. Microbiol.* **66**, 83–101 (2012).
4. A. E. Santoro, R. A. Richter, C. L. Dupont, Planktonic marine Archaea. *Annu. Rev. Mar. Sci.* **11**, 131–158 (2019).
5. T. B. Meador, N. Schoffelen, T. G. Ferdelman, O. Rebello, A. Khachikyan, M. Könneke, Carbon recycling efficiency and phosphate turnover by marine nitrifying archaea. *Sci. Adv.* **6**, eaba1799 (2020).
6. R. J. E. Alves, B. Q. Minh, T. Urich, A. von Haeseler, C. Schleper, Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on *amoA* genes. *Nat. Commun.* **9**, 1517 (2018).
7. T. H. Erguder, N. Boon, L. Wittebolle, M. Marzorati, W. Verstraete, Environmental factors shaping the ecological niches of ammonia-oxidizing archaea. *FEMS Microbiol. Rev.* **33**, 855–869 (2009).
8. M. Ren, X. Feng, Y. Huang, H. Wang, Z. Hu, S. Clingenpeel, B. K. Swan, M. M. Fonseca, D. Posada, R. Stepanauskas, J. T. Hollibaugh, P. G. Foster, T. Woyke, H. Luo, Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *ISME J.* **13**, 2150–2161 (2019).
9. P. O. Sheridan, S. Raguideau, C. Quince, J. Holden, L. Zhang, W. H. Gaze, J. Holden, A. Mead, S. Raguideau, C. Quince, A. C. Singer, E. M. H. Wellington, L. Zhang, T. A. Williams, C. Gubry-Rangin, Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. *Nat. Commun.* **11**, 5494 (2020).
10. S. S. Abby, M. Kerou, C. Schleper, Ancestral reconstructions decipher major adaptations of ammonia-oxidizing archaea upon radiation into moderate terrestrial and marine environments. *mBio* **11**, 1–20 (2020).
11. Y. Yang, C. Zhang, T. M. Lenton, X. Yan, M. Zhu, M. Zhou, J. Tao, T. J. Phelps, Z. Cao, The evolution pathway of ammonia-oxidizing archaea shaped by major geological events. *Mol. Biol. Evol.* **38**, 3637–3648 (2021).
12. B. Wang, W. Qin, Y. Ren, X. Zhou, M.-Y. Jung, P. Han, E. A. Eloe-Fadrosh, M. Li, Y. Zheng, L. Lu, X. Yan, J. Ji, Y. Liu, L. Liu, C. Heiner, R. Hall, W. Martens-Habbena, C. W. Herbold, S.-K. Rhee, D. H. Bartlett, L. Huang, A. E. Ingalls, M. Wagner, D. A. Stahl, Z. Jia, Expansion of *Thaumarchaeota* habitat range is correlated with horizontal transfer of ATPase operons. *ISME J.* **13**, 3067–3079 (2019).
13. M. Ren, J. Wang, Phylogenetic divergence and adaptation of *Nitrososphaeria* across lake depths and freshwater ecosystems. *ISME J.* **16**, 1491–1501 (2022).
14. N. R. Lovejoy, E. Bermingham, A. P. Martin, Marine incursion into South America. *Nature* **396**, 421–422 (1998).
15. K. N. Kirchhoff, T. Hauffe, B. Stelbrink, C. Albrecht, T. Wilke, Evolutionary bottlenecks in brackish water habitats drive the colonization of fresh water by stingrays. *J. Evol. Biol.* **30**, 1576–1591 (2017).
16. A. E. R. Soares, C. G. Schrago, The influence of taxon sampling on Bayesian divergence time inference under scenarios of rate heterogeneity among lineages. *J. Theor. Biol.* **364**, 31–39 (2015).
17. M. Mehrshad, M. A. Amoozegar, R. Ghai, S. A. S. Fazeli, F. Rodriguez-Valera, Genome reconstruction from metagenomic data sets reveals novel microbes in the brackish waters of the Caspian Sea. *Appl. Environ. Microbiol.* **82**, 1599–1612 (2016).
18. S. Hiraoka, Y. Okazaki, M. Anda, A. Toyoda, S. I. Nakano, W. Iwasaki, Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community. *Nat. Commun.* **10**, 159 (2019).
19. Y. Okazaki, Y. Nishimura, T. Yoshida, H. Ogata, S. I. Nakano, Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. *Environ. Microbiol.* **21**, 4740–4754 (2019).
20. P. Xing, Y. Tao, J. Luo, L. Wang, B. Li, H. Li, Q. Wu, Stratification of microbiomes during the holomictic period of Lake Fuxian, an alpine monomictic lake. *Limnol. Oceanogr.* **65**, S134–S148 (2020).
21. P. J. Cabello-Yeves, C. Callieri, A. Picazo, M. Mehrshad, J. M. Haro-Moreno, J. J. Roda-Garcia, N. Dzhembekova, V. Slabakova, N. Slabakova, S. Moncheva, F. Rodriguez-Valera, The microbiome of the Black Sea water column analyzed by shotgun and genome centric metagenomics. *Environ. Microbiome* **16**, 5 (2021).
22. L. Reji, C. A. Francis, Metagenome-assembled genomes reveal unique metabolic adaptations of a basal marine Thaumarchaeota lineage. *ISME J.* **14**, 2105–2115 (2020).
23. S. E. Hampton, S. McGowan, T. Ozersky, S. G. P. Virdis, T. T. Vu, T. L. Spanbauer, B. M. Kraemer, G. Swann, A. W. Mackay, S. M. Powers, M. F. Meyer, S. G. Labou, C. M. O'Reilly, M. DiCarlo, A. W. E. Galloway, S. C. Fritz, Recent ecological change in ancient lakes. *Limnol. Oceanogr.* **63**, 2277–2304 (2018).
24. K. T. Konstantinidis, R. Rosselló-Móra, R. Amann, Uncultivated microbes in need of their own taxonomy. *ISME J.* **11**, 2399–2406 (2017).
25. F. Klotz, K. Kitzinger, D. K. Ngugi, P. Büsing, S. Littmann, M. M. M. Kuypers, B. Schink, M. Pester, Quantification of archaea-driven freshwater nitrification from single cell to ecosystem levels. *ISME J.* **16**, 1647–1656 (2022).
26. X. Didelot, D. J. Wilson, ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLOS Comput. Biol.* **11**, e1004041 (2015).
27. J. Alneberg, C. M. G. Karlsson, A.-M. Divne, C. Bergin, F. Homa, M. V. Lindh, L. W. Hugerth, T. J. G. Ettema, S. Bertilsson, A. F. Andersson, J. Pinhassi, Genomes from uncultivated prokaryotes: A comparison of metagenome-assembled and single-amplified genomes. *Microbiome* **6**, 173 (2018).

28. J. Herber, F. Klotz, B. Frommeyer, S. Weis, D. Straile, A. Kolar, J. Sikorski, M. Egert, M. Dannenmann, M. Pester, A single *Thaumarchaeon* drives nitrification in deep oligotrophic Lake Constance. *Environ. Microbiol.* **22**, 212–228 (2020).
29. K. T. Konstantinidis, E. F. DeLong, Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J.* **2**, 1052–1065 (2008).
30. A. Peña-Gonzalez, M. J. Soto-Girón, S. Smith, J. Sistrunk, L. Montero, M. Páez, E. Ortega, J. K. Hatt, W. Cevallos, G. Trueba, K. Levy, K. T. Konstantinidis, Metagenomic signatures of gut infections caused by different *Escherichia coli* pathotypes. *Appl. Environ. Microbiol.* **85**, e01820-19 (2019).
31. W. P. Hanage, B. G. Spratt, K. M. E. Turner, C. Fraser, Modelling bacterial speciation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2039–2044 (2006).
32. C. Fraser, W. P. Hanage, B. G. Spratt, Recombination and the nature of bacterial speciation. *Science* **315**, 476–480 (2007).
33. K. Zaremba-Niedzwiedzka, J. Viklund, W. Zhao, J. Ast, A. Sczyrba, T. Woyke, K. McMahon, S. Bertilsson, R. Stepanauskas, S. G. E. Andersson, Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. *Genome Biol.* **14**, R130 (2013).
34. M. Vos, X. Didelot, A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208 (2009).
35. M. L. Bendall, S. L. Stevens, L.-K. Chan, S. Malfatti, P. Schwientek, J. Tremblay, W. Schackwitz, J. Martin, A. Pati, B. Bushnell, J. Froula, D. Kang, S. G. Tringe, S. Bertilsson, M. A. Moran, A. Shade, R. J. Newton, K. D. McMahon, R. R. Malmstrom, Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).
36. M. Groussin, M. Gouy, Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in Archaea. *Mol. Biol. Evol.* **28**, 2661–2674 (2011).
37. R. I. Woolway, C. J. Merchant, Worldwide alteration of lake mixing regimes in response to climate change. *Nat. Geosci.* **12**, 271–276 (2019).
38. W. Qin, K. R. Heal, R. Ramdasi, J. N. Kobelt, W. Martens-Habbena, A. D. Bertagnolli, S. A. Amin, C. B. Walker, H. Urakawa, M. Könneke, A. H. Devol, J. W. Moffett, E. V. Armbrust, G. J. Jensen, A. E. Ingalls, D. A. Stahl, *Nitrosopumilus maritimus* gen. nov., sp. nov., *Nitrosopumilus cobalaminigenes* sp. nov., *Nitrosopumilus oxyclinae* sp. nov., and *Nitrosopumilus ureiphilus* sp. nov., four marine ammonia-oxidizing archaea of the phylum *Thaumarchaeota*. *Int. J. Syst. Evol. Microbiol.* **67**, 5067–5079 (2017).
39. T. I. Gossmann, A. Shanmugasundram, S. Börno, L. Duvaux, C. Lemaire, H. Kuhl, S. Klages, L. D. Roberts, S. Schade, J. M. Gostner, F. Hildebrand, J. Vowinckel, C. Bichet, M. Mülleder, E. Calvani, A. Zelezniak, J. L. Griffin, P. Bork, D. Allaine, A. Cohas, J. J. Welch, B. Timmermann, M. Ralser, Ice-age climate adaptations trap the Alpine Marmot in a state of low genetic diversity. *Curr. Biol.* **29**, 1712–1720.e7 (2019).
40. N. Lartillot, T. Lepage, S. Blanquart, PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
41. S.-C. Chen, G.-X. Sun, Y. Yan, K. T. Konstantinidis, S.-Y. Zhang, Y. Deng, X.-M. Li, H.-L. Cui, F. Musat, D. Popp, B. P. Rosen, Y.-G. Zhu, The great oxidation event expanded the genetic repertoire of arsenic metabolism and cycling. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10414–10421 (2020).
42. B. K. Kim, M.-Y. Jung, D. S. Yu, S.-J. Park, T. K. Oh, S.-K. Rhee, J. F. Kim, Genome sequence of an ammonia-oxidizing soil archaeon, "*Candidatus* Nitrosoarchaeum koreensis" MY1. *J. Bacteriol.* **193**, 5539–5540 (2011).
43. B. Bayer, J. Vojvoda, T. Reinthaler, C. Reyes, M. Pinto, G. J. Herndl, *Nitrosopumilus adriaticus* sp. nov. and *Nitrosopumilus piranensis* sp. nov., two ammonia-oxidizing archaea from the Adriatic Sea and members of the class *Nitrososphaeria*. *IJSEM* **69**, 1892–1902 (2019).
44. A. E. Santoro, C. L. Dupont, R. A. Richter, M. T. Craig, P. Carini, M. R. McIlvin, Y. Yang, W. D. Orsi, D. M. Moran, M. A. Saito, Genomic and proteomic characterization of "*Candidatus* Nitrosopelagicus brevis": An ammonia-oxidizing archaeon from the open ocean. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1173–1178 (2015).
45. D. K. Ngugi, J. Blom, I. Alam, M. Rashid, W. Ba-Alawi, G. Zhang, T. Hikmawan, Y. Guan, A. Antunes, R. Siam, H. E. Dorry, V. Bajic, U. Stingl, Comparative genomics reveals adaptations of a halotolerant thaumarchaeon in the interfaces of brine pools in the Red Sea. *ISME J.* **9**, 396–411 (2015).
46. P. F. Hoffman, D. S. Abbot, Y. Ashkenazy, D. I. Benn, J. J. Brocks, P. A. Cohen, G. M. Cox, J. R. Creveling, Y. Donnadieu, D. H. Erwin, I. J. Fairchild, D. Ferreira, J. C. Goodman, G. P. Halverson, M. F. Jansen, G. L. Hir, G. D. Love, F. A. Macdonald, A. C. Maloof, C. A. Partin, G. Ramstein, B. E. J. Rose, C. V. Rose, P. M. Sadler, E. Tziperman, A. Voigt, S. G. Warren, Snowball Earth climate dynamics and Cryogenian geology-geobiology. *Sci. Adv.* **3**, e1600983 (2017).
47. T. W. Lyons, C. T. Reinhard, N. J. Planavsky, The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
48. I. P. Montañez, C. J. Poulsen, The late paleozoic ice age: An evolving paradigm. *Annu. Rev. Earth Planet. Sci.* **41**, 629–656 (2013).
49. Z.-Q. Chen, M. J. Benton, The timing and pattern of biotic recovery following the end-Permian mass extinction. *Nat. Geosci.* **5**, 375–383 (2012).
50. E. Cavaliere, H. M. Baulch, Winter nitrification in ice-covered lakes. *PLOS ONE* **14**, e0224864 (2019).
51. T. M. Butler, A.-C. Wilhelm, A. C. Dwyer, P. N. Webb, A. L. Baldwin, S. M. Techtmann, Microbial community dynamics during lake ice freezing. *Sci. Rep.* **9**, 6231 (2019).
52. B. Kraft, N. Jehmlich, M. Larsen, L. A. Bristow, M. Könneke, B. Thamdrup, D. E. Canfield, Oxygen and nitrogen production by an ammonia-oxidizing archaeon. *Science* **375**, 97–100 (2022).
53. A. Milanese, D. R. Mende, L. Paoli, G. Salazar, H.-J. Ruscheweyh, M. Cuenca, P. Hingamp, R. Alves, P. I. Costea, L. P. Coelho, T. S. B. Schmidt, A. Almeida, A. L. Mitchell, R. D. Finn, J. Huerta-Cepas, P. Bork, G. Zeller, S. Sunagawa, Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
54. C. Callieri, S. Hernández-Avilés, M. M. Salcher, D. Fontaneto, R. Bertoni, Distribution patterns and environmental correlates of Thaumarchaeota abundance in six deep subalpine lakes. *Aquat. Sci.* **78**, 215–225 (2015).
55. S. N. Merbt, D. A. Stahl, E. O. Casamayor, E. Martí, G. W. Nicol, J. I. Prosser, Differential photoinhibition of bacterial and archaeal ammonia oxidation. *FEMS Microbiol. Lett.* **327**, 41–46 (2012).
56. J.-G. Kim, S.-J. Park, J. S. S. Damsté, S. Schouten, W. I. C. Rijpstra, M.-Y. Jung, S.-J. Kim, J.-H. Gwak, H. Hong, O.-J. Si, S. Lee, E. L. Madsen, S.-K. Rhee, Hydrogen peroxide detoxification is a key mechanism for growth of ammonia-oxidizing archaea. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7888–7893 (2016).
57. C. M. Duarte, D. K. Ngugi, I. Alam, J. Pearman, A. Kamau, V. M. Eguiluz, T. Gojobori, S. G. Acinas, J. M. Gasol, V. Bajic, X. Irigoien, Sequencing effort dictates gene discovery in marine microbial metagenomes. *Environ. Microbiol.* **22**, 4589–4603 (2020).
58. D. Dudgeon, A. H. Arthington, M. O. Gessner, Z. Kawabata, D. J. Knowler, C. Lévêque, R. J. Naiman, A. Prieur-Richard, D. Soto, M. L. J. Stiassny, C. A. Sullivan, Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biol. Rev.* **81**, 163–182 (2006).
59. S. F. Paver, R. J. Newton, M. L. Coleman, Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environ. Microbiol.* **22**, 433–446 (2020).
60. P. Q. Tran, S. C. Bachand, P. B. McIntyre, B. M. Kraemer, Y. Vadeboncoeur, I. A. Kimirei, R. Tamatamah, K. D. McMahon, K. Anantharaman, Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *ISME J.* **15**, 1971–1986 (2021).
61. D. M. Emms, S. Kelly, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
62. C. L. Dupont, J. Larsson, S. Yooseph, K. Ininbergs, J. Goll, J. Asplund-Samuelsson, J. P. McCrow, N. Celepli, L. Z. Allen, M. Ekman, A. J. Lucas, Å. Hagström, M. Thiagarajan, B. Brindefalk, A. R. Richter, A. F. Andersson, A. Tenney, D. Lundin, A. Tovchigrechko, J. A. A. Nylander, D. Brami, J. H. Badger, A. E. Allen, D. B. Rusch, J. Hoffman, E. Norrby, R. Friedman, J. Pinhassi, J. C. Venter, B. Bergman, Functional tradeoffs underpin salinity-driven divergence in microbial community composition. *PLOS ONE* **9**, e89549 (2014).
63. K. Julenius, A. G. Pedersen, Protein evolution is faster outside the cell. *Mol. Biol. Evol.* **23**, 2039–2048 (2006).
64. R. Logares, J. Bråte, S. Bertilsson, J. L. Clasen, K. Shalchian-Tabrizi, K. Rengefors, Infrequent marine-freshwater transitions in the microbial world. *Trends Microbiol.* **17**, 414–422 (2009).
65. G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, A. C. Gregory, K. Labadie, C. Marec, E. Pelletier, M. Royo-Llonch, S. Roux, P. Sánchez, H. Uehara, A. A. Zayed, G. Zeller, M. Carmichael, C. Dimier, J. Ferland, S. Kandels, M. Picheral, S. Pisarev, J. Poulain, T. O. Coordinators, S. G. Acinas, M. Babin, P. Bork, C. Bowler, C. de Vargas, L. Guidi, P. Hingamp, D. Iudicone, L. Karp-Boss, E. Karsenti, H. Ogata, S. Pesant, S. Speich, M. B. Sullivan, P. Wincker, S. Sunagawa, Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).
66. L. C. Vitorino, L. A. Bessa, Microbial diversity: The gap between the estimated and the known. *Diversity* **10**, 46 (2018).
67. E. French, J. A. Kozlowski, M. Mukherjee, G. Bullerjahn, A. Bollmann, Ecophysiological characterization of ammonia-oxidizing archaea and bacteria from freshwater. *Appl. Environ. Microbiol.* **78**, 5773–5780 (2012).
68. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
69. D. Li, C.-M. Liu, R. Luo, K. Sadakane, T.-W. Lam, MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676 (2015).

70. D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

71. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

72. P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, D. H. Parks, GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2019).

73. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

74. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

75. R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, F. Meyer, R. Knight, R. Finn, A. Lapidus, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, T. Woyke, Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).

76. N. Lartillot, N. Rodrigue, D. Stubbs, J. Richer, PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).

77. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

78. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

79. Q. Tao, K. Tamura, F. U. Battistuzzi, S. Kumar, A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies. *Mol. Biol. Evol.* **36**, 811–824 (2019).

80. J. M. Wolfe, G. P. Fournier, Horizontal gene transfer constrains the timing of methanogen evolution. *Nat. Ecol. Evol.* **2**, 897–903 (2018).

81. C. E. Blank, An expansion of age constraints for microbial clades that lack a conventional fossil record using phylogenomic dating. *J. Mol. Evol.* **73**, 188–208 (2011).

82. A. Bekker, H. D. Holland, P.-L. Wang, D. Rumble, H. J. Stein, J. L. Hannah, L. L. Coetzee, N. J. Beukes, Dating the rise of atmospheric oxygen. *Nature* **427**, 117–120 (2004).

83. G. Luo, S. Ono, N. J. Beukes, D. T. Wang, S. Xie, R. E. Summons, Rapid oxygenation of Earth′s atmosphere 2.33 billion years ago. *Sci. Adv.* **2**, e1600134 (2016).

84. C. E. Blank, Phylogenomic dating—A method of constraining the age of microbial taxa that lack a conventional fossil record. *Astrobiology* **9**, 173–191 (2009).

85. A. E. Darling, B. Mau, N. T. Perna, progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE* **5**, e11147 (2010).

86. S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

87. C. Sjöqvist, L. F. Delgado, J. Alneberg, A. F. Andersson, Ecologically coherent population structure of uncultivated bacterioplankton. *ISME J.* **15**, 3034–3049 (2021).

88. D. L. Rabosky, Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLOS ONE* **9**, e89543 (2014).

89. D. L. Rabosky, M. Grundler, C. Anderson, P. Title, J. J. Shi, J. W. Brown, H. Huang, J. G. Larson, BAMMtools: An R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods Ecol. Evol.* **5**, 701–707 (2014).

90. A.-Ş. Andrei, M. M. Salcher, M. Mehrshad, P. Rychtecký, P. Znachor, R. Ghai, Niche-directed evolution modulates genome architecture in freshwater Planctomycetes. *ISME J.* **13**, 1056–1071 (2019).

91. I. Mukherjee, M. M. Salcher, A.-Ş. Andrei, V. S. Kavagutti, T. Shabarova, V. Grujčić, M. Haber, P. Layoun, Y. Hodoki, S.-I. Nakano, K. Šimek, R. Ghai, A freshwater radiation of diplonemids. *Environ. Microbiol.* **22**, 4658–4668 (2020).

92. K. Rojas-Jimenez, A. Araya-Lobo, F. Quesada-Perez, J. Akerman-Sanchez, B. Delgado-Duran, L. Ganzert, P. O. Zavialov, S. Alymkulov, G. Kirillin, H.-P. Grossart, Variation of bacterial communities along the vertical gradient in Lake Issyk Kul, Kyrgyzstan. *Env. Microbiol. Rep.* **13**, 337–347 (2021).

93. K. Rojas-Jimenez, C. Wurzbacher, E. C. Bourne, A. Chiuchiolo, J. C. Priscu, H.-P. Grossart, Early diverging lineages within Cryptomycota and Chytridiomycota dominate the fungal communities in ice-covered lakes of the McMurdo Dry Valleys, Antarctica. *Sci. Rep.* **7**, 15348 (2017).

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

95. B. Brian, BBMap: A fast, accurate, splice-aware aligner, in *Proceedings of the 9th Annual Genomics of Energy & Environment Meeting* (2014);www.osti.gov/biblio/1241166.

96. S. Andrews, FastQC: A quality control tool for high throughput sequence data (2010); www.bioinformatics.babraham.ac.uk/projects/fastqc.

97. D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

98. M. Steinegger, J. Söding, Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 2542 (2018).

99. S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Doré, S. D. Ehrlich, A. Stamatakis, P. Bork, Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).

100. T. Kahlke, P. J. Ralph, BASTA—Taxonomic classification of sequences and sequence bins using last common ancestor estimations. *Methods Ecol. Evol.* **10**, 100–103 (2018).

101. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2017).

102. R. C. Edgar, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

103. H. Wickham, *ggplot2, Elegant Graphics for Data Analysis* (Springer, 2009), vol. 8.

104. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).

105. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. D. Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).

106. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

107. M. Richter, R. Rosselló-Móra, Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 19126–19131 (2009).

108. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

109. B. Contreras-Moreira, P. Vinuesa, GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **79**, 7696–7701 (2013).

110. F. Boyer, C. Mercier, A. Bonin, Y. L. Bras, P. Taberlet, E. Coissac, obitools: A unix-inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.* **16**, 176–182 (2016).

111. J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, P. Bork, eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).

112. M. L. Borowiec, AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ* **4**, e1660 (2016).

113. J. T. Osvatic, L. G. E. Wilkins, L. Leibrecht, M. Leray, S. Zauner, J. Polzin, Y. Camacho, O. Gros, J. A. van Gils, J. A. Eisen, J. M. Petersen, B. Yuen, Global biogeography of chemosynthetic symbionts reveals both localized and globally distributed symbiont groups. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2104378118 (2021).

114. A. Crits-Christoph, M. R. Olm, S. Diamond, K. Bouma-Gregson, J. F. Banfield, Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME J* **14**, 1834–1846 (2020).

115. Z. Chen, X. Wang, Y. Song, Q. Zeng, Y. Zhang, H. Luo, *Prochlorococcus* have low global mutation rate and small effective population size. *Nat. Ecol. Evol.* **6**, 183–194 (2022).

116. C. Gawad, W. Koh, S. R. Quake, Single-cell genome sequencing: Current state of the science. *Nat. Rev. Microbiol.* **17**, 175–188 (2016).

117. M. L. Reno, N. L. Held, C. J. Fields, P. V. Burke, R. J. Whitaker, Biogeography of the Sulfolobus islandicus pan-genome. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 8605–8610 (2009).

118. N. J. Croucher, A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, S. R. Harris, Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).

119. M. Lynch, Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).

120. J. W. Drake, Avoiding dangerous missense: Thermophiles display especially low mutation rates. *PLOS Genet.* **5**, e1000520 (2009).

121. G. E. A. Swann, V. N. Panizzo, S. Piccolroaz, V. Pashley, M. S. A. Horstwood, S. Roberts, E. Vologina, N. Piotrowska, M. Sturm, A. Zhdanov, N. Granin, C. Norman, S. McGowan,

A. W. Mackay, Changing nutrient cycling in Lake Baikal, the world's oldest lake. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 27211–27217 (2020).

122. J. Quehenberger, L. Shen, S.-V. Albers, B. Siebers, O. Spadiut, *Sulfolobus*—A potential key organism in future biotechnology. *Front. Microbiol.* **8**, 2474 (2017).
123. A.-M. Waldvogel, M. Pfenninger, Temperature dependence of spontaneous mutation rates. *Genome Res.* **31**, 1582–1589 (2021).
124. R. F. Weiss, E. C. C. Carmack, V. M. Koropalov, Deep-water renewal and biological production in Lake Baikal. *Nature* **349**, 665–669 (1991).
125. C. H. Wellman, P. L. Osterloff, U. Mohiuddin, Fragments of the earliest land plants. *Nature* **425**, 282–285 (2003).

Acknowledgments: We thank the U.S. MCM Microbial Observatory and the McMurdo Long Term Ecological Research teams for fieldwork assistance. We thank the Kyrgyz State Agency on Environment Protection and Forestry for providing access to water samples from Lake Issyk Kul under the MTA 24.11.2021 N.01-mb/191 to H.-P.G. (noncommercial use). We thank H. Luo and F. Xiaoyuan for providing the concatenated alignment from their publication. We also thank K. Baumann and H. Bürgmann for assisting with data interpretation. **Funding:** D.K.N. and M.P. were both financially supported by the Leibniz Institute DSMZ and the DFG (NG 183/2-1 to D.K.N. and PE2147/4-1 to M.P.). Data from Central European lakes were provided through funding support from the Grant Agency of the Czech Republic (GACR grant 19-23469S) to M.M.S. R.G. was supported by the Grant Agency of the Czech Republic (GACR grant 20-12496X). M.-C.C. was supported by the Grant Agency of the Czech Republic (GACR grant 22-0332S). A.-S.A. was supported by Ambizione grant PZ00P3_193240 (Swiss National Science Foundation). Data from the Antarctica lakes were provided through funding from the NSF Office of Polar Programs (award nos. 1637708 and 0237335) to J.C.P. and the Leibniz "Mycolink" SAW project (Pakt/SAW-2014-IGB-1) to H.-P.G. **Author contributions:** D.K.N. and M.P. designed the study. D.K.N. integrated the data; performed bioinformatics analyses with support from M.M.S., A.-S.A., and R.G.; and drafted the manuscript. F.K. and M.P. provided metagenomes and metatranscriptomes from Lake Constance, while P.B. performed metagenomic and metatranscriptomic sequencing. M.M.S., M.-C.C., D.I., H.-P.G., A.-S.A., P.X., and R.G. assembled MAGs and provided access to their metagenome datasets of European and Asian freshwater lakes and the Black Sea. H.-P.G. provided genomic DNA from Antarctic lakes. J.C.P. and S.A. organized fieldwork and cruises to Antarctica and Lake Issyk-Kul, respectively, together with H.-P.G. All authors reviewed the results and revised and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Metagenomic data are available in the ENA portal (www.ebi.ac.uk/ena/browser/home), whereas the MAGs are available in NCBI (www.ncbi.nlm.nih.gov/), NGDC (https://ngdc.cncb.ac.cn/), and IMG (https://img.jgi.doe.gov/) portals. Accession numbers and identifiers for the publicly available metagenomes and MAGs are available in tables S1 and S3, respectively. Metagenomes generated in the course of this study are available at NCBI or ENA under BioProject numbers PRJNA820565 and PRJEB35640. In addition, the supporting datasets (data S1 to S12) and the MAG sequences of lacustrine Nitrososphaerota are also made available in FigShare (https://doi.org/10.6084/m9.figshare.19448072.v1). All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials.

Submitted 11 May 2022
Accepted 18 November 2022
Published 1 February 2023
10.1126/sciadv.adc9392