Genomic and proteomic characterization of "*Candidatus* Nitrosopelagicus brevis": An ammonia-oxidizing archaeon from the open ocean

Alyson E. Santoro[a,1], Christopher L. Dupont[b], R. Alex Richter[c], Matthew T. Craig[b,d], Paul Carini[a], Matthew R. McIlvin[e], Youngik Yang[c], William D. Orsi[a], Dawn M. Moran[e], and Mak A. Saito[e]

[a]Horn Point Laboratory, University of Maryland Center for Environmental Science, Cambridge, MD 21613; [b]Microbial and Environmental Genomics and [c]Informatics Group, J. Craig Venter Institute, San Diego, CA 92037; [d]Department of Environmental and Ocean Sciences, University of San Diego, San Diego, CA 92110; and [e]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA 02543

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved December 12, 2014 (received for review August 27, 2014)

Thaumarchaeota are among the most abundant microbial cells in the ocean, but difficulty in cultivating marine Thaumarchaeota has hindered investigation into the physiological and evolutionary basis of their success. We report here a closed genome assembled from a highly enriched culture of the ammonia-oxidizing pelagic thaumarchaeon CN25, originating from the open ocean. The CN25 genome exhibits strong evidence of genome streamlining, including a 1.23-Mbp genome, a high coding density, and a low number of paralogous genes. Proteomic analysis recovered nearly 70% of the predicted proteins encoded by the genome, demonstrating that a high fraction of the genome is translated. In contrast to other minimal marine microbes that acquire, rather than synthesize, cofactors, CN25 encodes and expresses near-complete biosynthetic pathways for multiple vitamins. Metagenomic fragment recruitment indicated the presence of DNA sequences >90% identical to the CN25 genome throughout the oligotrophic ocean. We propose the provisional name "*Candidatus* Nitrosopelagicus brevis" str. CN25 for this minimalist marine thaumarchaeon and suggest it as a potential model system for understanding archaeal adaptation to the open ocean.

nitrification | marine metagenomics | genome streamlining | archaea

Planktonic archaea are widespread in the marine environment. Below the photic zone, archaea can constitute greater than 30% of total bacterioplankton (1), making them among the most abundant cells in the ocean. The majority of pelagic archaea belong to the recently described phylum Thaumarchaeota (2, 3), also known as the Marine Group I archaea (4). In addition to representing large fractions of marine metagenomic datasets (5), metatranscriptomic data suggest that thaumarchaeal cells are metabolically active, with thaumarchaeal transcripts ranking as the most abundant in diverse marine environments (6–8). The metabolic activity of marine Thaumarchaeota has important implications for global biogeochemical cycles via their role in nitrogen remineralization, carbon fixation (9), and production of the greenhouse gas nitrous oxide (N_2O) (10).

At present there are six pure cultures of Thaumarchaeota: one from a marine aquarium [*Nitrosopumilus maritimus* SCM1 (11, 12)], two from an estuary in the northeast Pacific [PS0 and HCA1 (13)], and three from soil [*Nitrosphaera viennensis* (14) and *Nitrosotalea devanaterra* strains Nd1 and Nd2 (15)]. Of these isolates, *N. maritimus*, *N. viennensis*, and *N. devanaterra* are able to grow as chemolithoautotrophic ammonia oxidizers. Beyond these organisms, much of our knowledge of the genomic inventory (16–18), physiology, and biogeochemical activity of Thaumarchaeota has come from the characterization of enriched mixed cultures (19, 20) or uncultivated single cells (21, 22). Common genomic features in all sequenced representatives include a modified 3-hydroxypropioinate/4-hydroxybutyrate pathway for carbon fixation (23), an electron transport chain enriched in copper-centered metalloproteins, and lack of an identifiable homolog to hydroxylamine oxidoreductase (18, 24), an Fe-rich decaheme protein that catalyzes the second step of ammonia oxidation in all ammonia-oxidizing bacteria (25).

Given the tropical aquarium and estuarine origins of existing marine isolates, the extent to which their physiology and genomic features are representative of Thaumarchaeota in the open ocean is uncertain. In terms of physiology, *N. maritimus* grows chemolithoautotrophically, with ammonia as its sole energy source and bicarbonate as its sole carbon source. However, mixotrophy has been proposed for both *N. viennensis* and the marine isolates PS0 and HCA1 on the basis of growth stimulation when organic acids are added to the media (13, 14). In terms of genome content, metagenomic recruitment to *N. maritimus* is poor relative to that of single-cell genomes obtained from the open ocean (21).

Here, we present the closed genome of a marine ammonia-oxidizing Thaumarchaeota assembled from a low-diversity metagenome of an enrichment culture originating from the open ocean and previously described as CN25 (26). We mapped peptides collected from early stationary phase cells to translations of the CN25 genome's predicted ORFs to produce the first global proteome, to our knowledge, from a marine thaumarchaeon. Finally, we used the genome to probe existing marine metagenomic and metatranscriptomic datasets to understand the relative distribution of CN25 and *N. maritimus*-like genomes in the ocean.

Results and Discussion

Cultivation, Genome Sequencing, and Global Proteome. Previous fluorescent in situ hybridization characterization of the CN25

Significance

Thaumarchaeota are among the most abundant microbial cells in the ocean, but to date, complete genome sequences for marine Thaumarchaeota are lacking. Here, we report the 1.23-Mbp genome of the pelagic ammonia-oxidizing thaumarchaeon "*Candidatus* Nitrosopelagicus brevis" str. CN25. We present the first proteomic data, to our knowledge, from this phylum, which show a high proportion of proteins translated in oligotrophic conditions. Metagenomic fragment recruitment using data from the open ocean indicate the ubiquitous presence of *Ca.* N. brevis-like sequences in the surface ocean and suggest *Ca.* N. brevis as a model system for understanding the ecology and evolution of pelagic marine Thaumarchaeota.

Author contributions: A.E.S., C.L.D., and M.A.S. designed research; A.E.S., C.L.D., R.A.R., M.T.C., P.C., M.R.M., Y.Y., W.D.O., D.M.M., and M.A.S. performed research; M.A.S. contributed new reagents/analytic tools; A.E.S., C.L.D., P.C., W.D.O., and M.A.S. analyzed data; and A.E.S., C.L.D., P.C., and M.A.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. CP007026).

[1]To whom correspondence should be addressed. Email: asantoro@umces.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1416223112/-/DCSupplemental.

enrichment culture indicated that in late exponential phase, 90–95% of the cells are archaeal (26), and scanning electron microscopy shows the culture is dominated by rod-shaped cells with a diameter of 0.17–0.26 μm (mean 0.15 ± 0.02 μm; $n = 50$ cells) and length of 0.6–1.0 μm (0.78 ± 0.25 μm; *SI Appendix*, Fig. S1). A growth temperature optimum of ~22 °C (*SI Appendix*, Fig. S2) suggests physiological adaptation to subtropical surface ocean temperatures compared with a temperature optimum of 30 °C for *N. maritimus* (12).

Consistent with earlier fluorescent in situ hybridization data, 93.3% of the 49.6 million Illumina HiSeq reads from this low-diversity metagenome were less than 45% GC (guanine-cytosine) content, with the remaining reads falling into two low-coverage bins of ~50% and 65% GC content. A phylogenetic analysis indicated the archaeal reads were found in the low GC cluster. Assembly (via the Celera Assembler; wgs-assembler.sourceforge.net) of the low GC content bin resulted in five contigs at 40× coverage. Manual inspection of the sequence data, followed by PCR amplification and direct Sanger sequencing, resolved the genome into a single chromosome with a GC content of 33% (*SI Appendix*, Table S1 and Fig. S3).

At 1.23 Mbp, the closed CN25 genome is one of the smallest genomes of any free-living cell (Fig. 1, Table 1, and *SI Appendix*, Fig. S4). It encodes for 1,445 predicted protein-coding genes, one rRNA operon, and 42 tRNA genes. No extrachromosomal elements were identified. We propose the provisional name "*Candidatus* Nitrosopelagicus brevis" str. CN25 (*Ca*. N. brevis). The genus name refers to the organism's water column habitat and its ability to oxidize ammonia to nitrite. The species name refers both to the organism's affiliation with a clade of shallow water Thaumarchaeota (26) and its small genome.

The translated ORFs, predicted from the assembled genome, were used as a reference to identify proteins in a global proteome of early stationary phase cells (*SI Appendix* and Fig. 1). The proteome recovered peptides mapping to 1,012 unique proteins, or roughly 70% of the total predicted proteins (*SI Appendix*, Dataset S1). Relative to previously investigated microbes, *Ca.* N. brevis translates a large fraction of its proteome under oligotrophic conditions (*SI Appendix*, Table S2).

Energy Metabolism. The *Ca.* N. brevis genome encodes genes for all three subunits of ammonia monooxygenase (AMO) with the same order and orientation (*amoACB*; T478_0302, _0300, _0298) found in other marine Thaumarchaeota, and all three subunits were detected in the proteome (Fig. 1 and *SI Appendix*, Dataset S1). Although not among the top 15 most abundant proteins in terms of spectral counts, AmoB was highly abundant (top 5% of expressed proteins), as it is in the proteome of the ammonia-oxidizing bacterium *Nitrosomonas europaea* (27). The *Ca*. N. brevis genome also encodes for the 120-amino acid hypothetical protein previously termed AmoX [(28); T478_0301], located between *amoA* and *amoC*, and the proteome confirmed expression of this protein. As with all previously sequenced Thaumarchaeota, no hydroxylamine oxidoreductase homologs were identified. Five of the 15 most abundant proteins in the proteome were involved in energy production and conversion (Fig. 1 and *SI Appendix*, Dataset S1), and energy production proteins are abundant in the proteomes of other chemolithoautotrophic organisms (29); thus, highly abundant hypothetical proteins are promising candidates for additional proteins involved in energy generation.

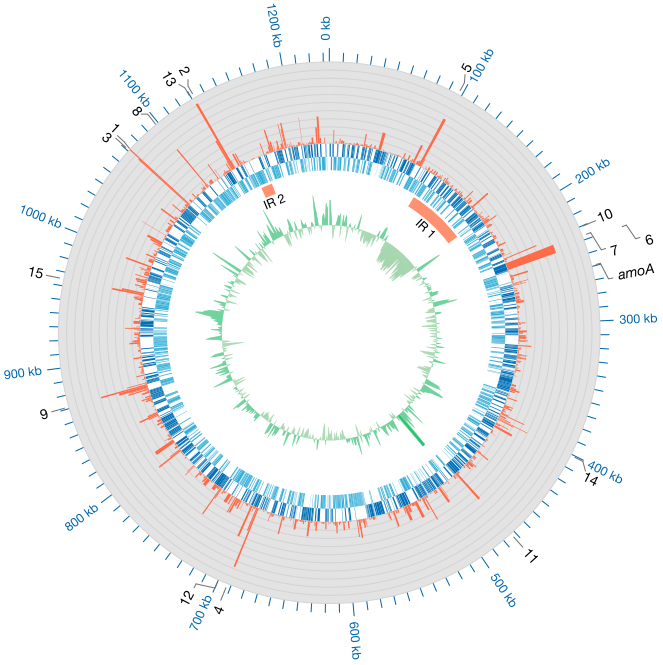


Fig. 1. The 1.23-Mbp genome and proteome of *Ca*. N. brevis str. CN25. The outermost ring is the position along the genome in thousands of nucleotide base pairs and annotations of the 15 most abundant proteins in the proteome, plus ammonia monooxygenase subunit *a* (*amoA*). The second ring (histogram) is the relative abundance of protein spectral counts detected in a global proteome. The third and fourth rings (blue and cyan) indicate predicted ORFs on the plus and minus strands, respectively. The fifth ring (red) indicates the location of putative genomic island regions (IR). The sixth or innermost ring (green) is GC anomaly based on a 2,000-bp moving average. Key to protein annotations: 1. conserved domain protein (T478_1299); 2. ATP synthase (T478_1372); 3. conserved domain protein (T478_1300); 4. translation elongation factor EF-1 (T478_0861); 5. AAA family ATPase (T478_0115); 6. RNA polymerase subunit A (*rpoA*, T478_0275); 7. RNA polymerase subunit B (*rpoB*, T478_0274); 8. alcohol dehydrogenase (T478_1333); 9. putative glutamate dehydrogenase (T478_1059); 10. putative malate dehydrogenase (T478_0268); 11. conserved hypothetical protein (T478_0572); 12. oxidoreductase, short chain dehydrogenase (T478_0869); 13. ATP synthase alpha/beta chain T478_1371); 14. flavodoxin (T478_0486); 15. putative acetyl-CoA carboxylase (T478_1175). Relative abundance of all proteins identified in the global proteome is provided as an *SI Appendix*, Dataset S1.

Metalloenzyme-specific analyses conducted for *Ca.* N. brevis suggest that, similar to *N. maritimus*, there is a reliance on copper-containing electron transport proteins (*SI Appendix*, Dataset S2). The *Ca.* N. brevis genome encodes for 12 cupredoxin domain-containing proteins (Structural Classification of Proteins family 49550), which bind copper in a redox active fashion, compared with 27 proteins for *N. maritimus*. Many of the single-domain cupredoxins

Table 1. Genome sizes and coding densities of select oligotrophic marine bacteria and previously sequenced Thaumarchaeota

	Oligotrophic pelagic marine bacteria			Thaumarchaeota			
Characteristics	*Methylophilales* sp. HTCC2181 (OM43)	*Pelagibacter ubique* HTCC1062	*Prochloroccocus marinus* AS9601	*N. gargensis*	*Ca.* N. limnia SFB1	*N. maritimus*	*Ca.* N. brevis
Size, Mbp	1.304	1.309	1.670	2.834	1.743	1.645	1.232
ORFs	1,377	1,394	1,988	3,599	2,088	1,842	1,501
Percentage coding	95.0	96.1	91.2	81.5	84.8	90.8	94.6
Percentage GC	38	30	31	48	32	34	33

contain long N-terminal extensions lacking annotation, whereas two contain C-terminal PEFG sequences that likely target them to the cell membrane. Multicopper oxidases (Pfam07732) are not typically found in archaeal genomes outside the Thaumarchaeota and have been suggested as potential alternatives to the "missing" hydroxylamine oxidoreductase enzyme (24); *Ca.* N. brevis contains three multicopper oxidases, whereas *N. maritimus* contains six. Of the *Ca.* N. brevis multicopper oxidases, two were detected in the proteome (T478_0212, T478_1026), including the putative copper-containing nitrite (NO_2^-) reductase (*nirK*; T478_1026). *nirK* transcripts are abundant in some marine metatranscriptomes (7) and were abundant in the proteome (*SI Appendix*, Dataset S1).

Reductive N_2O production from NO_2^- has been demonstrated in enrichment cultures of *Ca.* N. brevis (10) and in *N. maritimus* (30, 31), although it is unclear whether reductive N_2O production originates from enzymatic or abiotic reactions. The *Ca.* N. brevis assembly encodes for two putative nitric oxide reductase accessory proteins (*norQ*, T478_0286, and *norD*, T478_0285), both of which were detected in the proteome. NorQ is essential for the activation of NorB, which catalyzes the reduction of NO to N_2O in both nitrifying (32) and denitrifying (33) bacteria. However, no homologs of *norB* were identified in *Ca*. N. brevis or in any other thaumarchaeal genome. Although implicated in reductive N_2O production, *norB* and *norQ* mutants of the bacterial nitrifier *N. europaea* still produce N_2O but have a greatly diminished capability to degrade NO (32). Thus, the genomic data leave the mechanism of reductive N_2O production in *Ca*. N. brevis unresolved.

Central Carbon Metabolism. Candidate genes encoding for a partial 3-hydroxypropionate/4-hydroxybutyrate pathway were identified, suggesting the potential for carbon fixation in *Ca.* N. brevis (*SI Appendix*, Dataset S2). We identified proteins from all eleven enzymes, with a subunit of the acetyl-/propionyl-CoA carboxylase enzyme complex (T478_1175) among the most abundant proteins (Fig. 1), suggestive of active carbon fixation during growth. The thaumarchaeal 3-hydroxypropionate/4-hydroxybutyrate pathway was recently demonstrated to be the most efficient pathway for carbon fixation (23), which is likely an important adaptation for chemolithoautotrophic growth in the oligotrophic ocean.

Putative genes for a complete tricarboxylic acid cycle were also identified, and all were detected in the *Ca.* N. brevis proteome, with malate dehydrogenase among the most abundant proteins (Fig. 2). Glycolysis is apparently incomplete (genes encoding a pyruvate kinase and phosphofructokinase were absent), but a complete gluconeogenic pathway was identified (*SI Appendix*, Dataset S2). However, *Ca*. N. brevis may benefit from the presence of organic compounds. For example, putative transport proteins for the import of lipoproteins, glycerol, and glycine betaine were all identified in the genome, with several present in the proteome, suggestive of potential alternative substrate use. Similarly, the persistence of a small percentage (<10% of total cells) of putatively heterotrophic bacterial cells in the enrichment culture and reports of reliance on organic compounds in other marine Thaumarchaeota (13) leave open the potential that *Ca.* N. brevis may benefit from organic compounds produced by the bacteria or in the natural seawater medium for growth. We found, however, no effect of organic carbon addition on the growth rate or cell yield of *Ca.* N. brevis cultures in tests with 20 different organic compounds (*SI Appendix*, Table S3).

Vitamin and Amino Acid Biosynthesis. Complete biosynthetic pathways for the B vitamin cofactors thiamin (B_1), riboflavin (B_2), pantothenate (B_5), pyridoxine (B_6), and biotin (B_7) are present in the *Ca.* N. brevis genome (*SI Appendix*, Dataset S2). A near-complete pathway for cobalamin (B_{12}) synthesis was also identified in the genome, missing only precorrin-6X reductase (*cbiJ-cobK*), which is also lacking in nearly all known cobalamin-producing archaea (34) except *Methanococcus* (35). Proteins within each of these pathways were detected in the proteome. Distributions of these vitamins in seawater have been suggested to explain the success of various

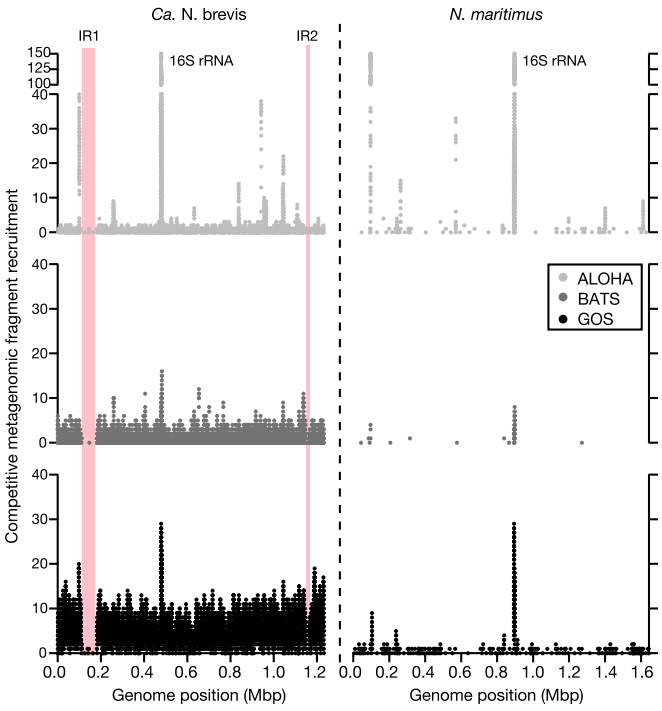


Fig. 2. Sequences highly similar to *Ca*. N. brevis dominate marine metagenomes. Competitive metagenomic fragment recruitment between the *Ca*. N. brevis genome assembly (*Left*) and *N. maritimus* (*Right*) at >90% nucleotide identity in marine metagenomic datasets from the Hawai'i Ocean Time-series (ALOHA), Bermuda Atlantic Time-series Station (BATS), and the Global Ocean Sampling Expedition (GOS). Regions highlighted in red indicate genomic IR in *Ca*. N. brevis.

phytoplankton lineages (36), although little is known about the source of them in seawater, particularly below the euphotic zone.

Consistent with the genetic capacity for B_{12} biosynthesis, *Ca.* N. brevis encodes for three major B_{12}-requiring enzymes: methylmalonyl-CoA mutase (T478_0628), methionine synthase (T478_1032), and ribonucleoside reductase (T478_1341). The genome also encodes for the archaeal-specific cobalt chelatase (*cbiX*) and *cobY-cobU* from the oxygen independent B_{12} biosynthetic pathway, which does not require oxygen to produce the cobalt-binding corrin ring center of the vitamin (34). Because of its small genome size, *Ca.* N. brevis has a relatively large genetic investment in B_{12} synthesis, with 1.7% of the genome encoding B_{12}-related genes compared with 0.7% in *Salmonella* (37). Our findings also support those of a recent metagenomic analysis showing the widespread distribution of thaumarchaeal B_{12} biosynthesis genes in the ocean (38). Six of the seven proteins for B_1 biosynthesis and two of the three proteins in the B_7 pathway were detected in the proteome. Vitamin B_1 is required for several central carbon metabolism enzymes including transketolase (T478_1212, T478_1213) and acetolactate synthase (T478_0886, T478_0887), and vitamin B_7 is a required coenzyme for the acetyl-/propionyl-CoA carboxylase enzyme complex (T478_1174, T478_1175, T478_1176). Other minimal genomes, such as *Pelagibacter* spp., lack the capability for complete B vitamin synthesis, uptake, and use (39, 40). The genomic and proteomic data presented here, together with the abundance of archaea in the mesopelagic (1), suggest Thaumarchaeota such as *Ca.* N. brevis are a potential source of multiple B vitamins required by microorganisms in the upper mesopelagic.

We identified complete or near-complete pathways for the synthesis of 18 amino acids, plus a near-complete pathway for methionine synthesis (*SI Appendix*, Dataset S2). We interpret apparent deficiencies in these pathways as gaps in our understanding of archaeal amino acid biosynthesis, rather than evidence of auxotrophy, as genes for all "missing" enzymes in the *Ca*. N. brevis genome

are also absent in *N. maritimus*, which grows in minimal medium without added amino acids. Proteins in all amino acid biosynthesis pathways except asparagine were detected in the proteome. Although genes coding for known mechanisms of proline biosynthesis were not annotated, the absence of a canonical proline biosynthesis pathway was previously noted in other archaea and may be substituted by synthesis from L-ornithine (41). Again, the presence of several putative amino acid and oligopeptidetransporters lends support to the possibility that amino acids may be acquired exogenously, despite having genomic inventory for their biosynthesis.

Comparative Genomic Analyses Suggest Adaptations to the Surface Ocean. Phylogenetic analysis of an alignment of concatenated ribosomal protein genes unambiguously associates *Ca.* N. brevis within the Thaumarchaeota (*SI Appendix*, Fig. S5), yet a comparative whole-genome analysis highlights the distinction between *Ca.* N. brevis and previously sequenced Thaumarchaeota. The average amino acid identity of aligned proteins between *Ca.* N. brevis and other thaumarchaeal genomes ranged from 34% (against *Candidatus* Nitrosphaera gargensis) to 75% (against *N. maritimus*) (*SI Appendix*, Table S4). Protein sequences from *Ca.* N. brevis and eight other thaumarchaeal genomes were clustered at a range of amino acid identities (*SI Appendix*, Fig. S6). Consistent with each new ammonia-oxidizing archaeal genome sequenced to date (18), *Ca.* N. brevis contains a large number of proteins that are either unique or highly divergent, relative to other thaumarchaea. Using a 50% amino acid identity threshold to define orthologs, the *Ca.* N. brevis genome contains 331 predicted proteins not present in any other thaumarchaeal predicted proteome (*SI Appendix*, Dataset S3).

We investigated the *Ca.* N. brevis proteins with <50% identity to other thaumarchaeal proteins as potentially adaptive to the pelagic environment from which it was enriched; specifically, the lower euphotic zone. UV radiation and reactive oxygen species are two potential physiological stresses present in sunlit waters. We identified two genes encoding putative deoxyribodipyrimidine photolyases (T478_0069 and T478_0078; (*SI Appendix*, Dataset S2), associated with DNA repair resulting from UV damage. Both of these proteins were detected in the dark-grown proteome, suggesting either that these proteins have an alternative function in *Ca.* N. brevis or that they are coregulated as part of a universal stress response, as they are in *Escherichia coli* (42). A unique putative alkyl hydroxy peroxidase (*ahpC*) associated with reactive oxygen and nitrogen stress response was also identified (T478_0940), although homologous sequences were also identified in several deep (4,000 m) ocean fosmids, suggesting this is not a surface ocean-specific gene. In addition to the two "unique" *ahpC*-like genes, five other genes encoding predicted proteins in the same family (Pfam00578) were identified in the *Ca.* N. brevis genome (*SI Appendix*, Dataset S2). Low trace metal concentrations in the surface ocean may also play a role in microbial adaptation to the oligotrophic surface ocean, including marine archaea (43). Consistent with this, several of the "unique" *Ca.* N. brevis genes encode putative metal transport proteins, including a ferrous iron transporter (T478_0963), a putative CorA-like Mg^{2+} or Co^{2+} transporter (T478_0228), and a Zn-binding protein (T478_0238).

The *Ca.* N. brevis genome is also distinguished by the lack of identifiable genes for several features reported in previously sequenced Thaumarchaeota. No genes encoding for flagellar synthesis or chemotaxis proteins were detected, suggesting a nonmotile lifestyle. *Ca.* N. brevis has no apparent capacity for biosynthesis of the osmolyte hydroxyectoine, as is present in the three sequenced *Nitrosopumilus* strains. The *Ca.* N. brevis genome does not encode for the Pst-type high-affinity phosphate transport system present in the *N. maritimus* genome, but it does encode for the transcriptional regulator PhoU (T478_0950) in a putative operon with a low-affinity phosphate transporter (Pit, T478_0951). We speculate that because subtropical North Pacific surface waters often contain residual phosphate, the genetic investment in a high-affinity phosphate transport system may not be necessary (44). Metabolism of methylphosphonic acid (MPn) by phosphate-starved microbes has recently been scrutinized as a possible explanation for the observed methane oversaturation in marine surface waters (45). *N. maritimus* was recently shown to synthesize MPn de novo, suggesting that planktonic marine archaea might be a natural source of MPn, and thus linked to marine methane dynamics (46). Surprisingly, the *Ca.* N. brevis genome does not encode for a complete MPn biosynthesis pathway. In particular, *Ca*. N. brevis does not encode for the key enzyme MpnS (46), suggesting it does not synthesize MPn. It remains to be seen whether other open ocean Thaumarchaeota also lack the capacity to synthesize MPn, but these findings show that MPn synthesis may not be universally conserved in planktonic Thaumarchaeota, and that changes in thaumarchaeal population structure may influence marine methane dynamics.

Evidence for Genome Streamlining in Marine Thaumarchaeota. The genome streamlining hypothesis argues that species with large effective population sizes are under selective pressures that favor small genomes, reducing the material or energetic cost of cellular replication in nutrient-poor environments (47, 48). Streamlined genomes are found in diverse, uncultivated bacteria in the oligotrophic ocean (49, 50), and it has been suggested that evolution of the *Archaea*, in particular, has been dominated by reductive selection (51). As exemplified by *Prochlorococcus* (52), the uncultivated bacterial clade SAR86 (50), and *Pelagibacter* (39), reductive selection can result in a loss of metabolic versatility or nutritional dependencies (53), such as the loss of pathways for assimilation of oxidized forms of nitrogen or essential vitamin cofactors. The *Ca.* N. brevis genome has no apparent loss of cofactor or amino acid metabolism and the concomitant inclusion of a complete pathway for carbon fixation. The genome has the highest coding density of any Thaumarchaeota (94.6%; Table 1), although the coding density is lower than for streamlined bacterial genomes such as *Pelagibacter* (Table 1). We did not find evidence of selection for shorter proteins, as average protein length is not correlated with genome size within the *Archaea*, according to an analysis of all finished archaeal genomes in the Integrated Microbial Genomes (IMG) database ($R^2 < 0.01$; $n = 164$).

Consistent with other streamlined genomes (39), the abundance of paralogous proteins is small, even when normalizing for genome size (*SI Appendix*, Table S5). In particular, the *Ca.* N. brevis genome contains a reduced number of genes involved in environmental sensing and regulation relative to other Thaumarchaeota. Transcriptional regulation in archaea is controlled by two families of basal transcription factors: transcription factor B (TFB) and TATA-binding proteins (54), with orthologous proteins present in eukaryotes. TFBs and TATA-binding proteins combine as TFB-TATA-binding protein pairs, with different regulons according to the pairing, allowing for a complex regulatory scheme with few proteins (54, 55). It has been hypothesized that organisms containing more TFBs may be better suited to changing environmental conditions (56). The *N. maritimus* genome has eight annotated TFBs, which is among the highest in the archaeal domain (56), suggesting a large network of potential regulatory complexes to respond to a changing environment. The *Ca.* N. brevis genome contains only four TFB, in contrast to eight to twelve for other sequenced Thaumarchaeota. Whether transcriptional regulation by factor swapping analogous to sigma factor switching in bacteria occurs within the Thaumarchaeota remains to be demonstrated (55).

Somewhat surprisingly for an oligotrophic microbe, there are fewer predicted transport proteins (77 predicted in IMG and 50 predicted using TransAAP; *SI Appendix*, Table S6 and *SI Appendix*, Dataset S2) compared with other aquatic Thaumarchaeota. This reduction in transport proteins is particularly manifest for ATP-binding cassette (ABC)-type transporters (there are 18 in the *Ca.* N. brevis genome vs. 31 in *N. maritimus*) and holds even when these estimates are normalized to genome size (14.6 vs. 18.9 ABC

transporters per millions of base pairs genome). The identification and expression of two Amt-type ammonium transporters (T478_1378, T478_1350) gives further support for the hypothesis that ammonia-oxidizing archaea actively transport ammonium (NH_4^+) into the cell (13, 20) and is consistent with detection of these genes in environmental metatranscriptomes (8, 57).

Comparison with Marine Metagenomic and Metatranscriptomic Data. We used competitive fragment recruitment (*SI Appendix*, *Materials and Methods*) to determine the relative recruitment to the *Ca.* N. brevis and *N. maritimus* genomes in more than 360 metagenomic and metatranscriptomic datasets from marine environments, including the Global Ocean Sampling (GOS) Expedition data (58) (*SI Appendix*, Dataset S4). At 90% nucleotide identity or higher within the GOS data, read recruitment to *Ca.* N. brevis was 30 times greater than to *N. maritimus* (Fig. 2), and *Ca.* N. brevis had higher recruitment in nearly twice as many samples (*SI Appendix*, Dataset S4). Two regions of the *Ca.* N. brevis genome were rarely observed in the metagenomic datasets, a characteristic associated with genomic islands, regions that are highly variable within a population of otherwise identical organisms (Figs. 1 and 2). Although *N. maritimus* also contains several genomic islands (5, 8), *Ca*. N. brevis's island gene contents are distinct from those in *N. maritimus*. The first *Ca.* N. brevis island, associated with a negative deviation in GC content (Fig. 1), encodes for 75 predicted proteins, of which nearly all appear to be involved in cell surface modifications through glycosylation (*SI Appendix*, Table S7). These enzymes may act to modify the cell surface, changing the palatability to grazers (59) or reducing susceptibility to phage infection. The second, much smaller, island contains mostly genes with unknown function. Although nothing is known about thaumarchaeal phage or thaumarchaeal defenses against them, genomic islands in many marine microbes are dominated by genes involved in cell wall and polysaccharide biosynthesis and modification, as they are in *Ca*. N. brevis, suggesting an important role of phage in thaumarchaeal population dynamics (60). To this end, we did not identify lysogenic phage or phage integrases in the *Ca*. N. brevis genome. Further, we found no clustered regularly interspaced short palindromic repeats (CRISPR) or CRISPR-associated protein systems of phage defense, although a putative abortive infection phage resistance protein was identified (T478_0343).

Two final observations from the competitive fragment recruitment analysis are the ubiquitous presence of thaumarchaeal genomes in the marine environment and the extent to which present cultivated strains do not represent this diversity. Even when recruitment to ribosomal RNA genes is excluded, genome fragments with >50% identity to either *Ca.* N. brevis or *N. maritimus* were found in all but one of the 366 datasets examined. Sequences >90% identity to *Ca.* N. brevis were abundant in the oligotrophic surface ocean (the GOS data), and the two oligotrophic time series datasets (Hawai'i Ocean Time-series and Bermuda Atlantic Time-series Station), implicating *Ca*. N. brevis as a globally abundant contributor to nitrification in ocean surface waters. The majority of the competitive fragment recruitment, however, was at nucleotide identities less than 90% (Fig. 3 and *SI Appendix*, Table S8). This indicates that a vast diversity of Thaumarchaeota distinct from *Ca.* N. brevis and *N. maritimus* exists in marine environments, and that further genomic and metabolic capability within this group remains to be explored.

Multiple cultures and corresponding reference genomes from marine bacterial clades such as the cyanobacterium *Prochlorococcus* and heterotrophic bacterium *Pelagibacter* have provided important insights into the forces driving genome evolution and diversification in the oligotrophic ocean (61, 62). Hindered by a lack of relevant cultures, we know far less about the open ocean Thaumarchaeota, although they play similarly important roles in marine biogeochemical cycling. The genome and proteome presented here for *Ca*. N. brevis, originating from the largest contiguous biome on Earth (63), are the first step to uncovering similar ecological and evolutionary insights into a significant component of the microbial community in the expanding oligotrophic ocean (64).

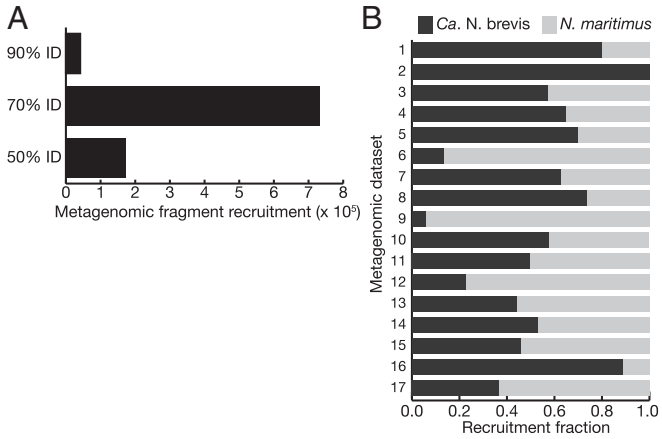


Fig. 3. (*A*) Combined metagenomic fragment recruitment to the *Ca*. N. brevis and *N. maritimus* genomes at three different nucleotide identity cutoffs. Bins are exclusive; that is, once a read is recruited at 90% identity, it is removed from the analysis and is thus not counted twice. Recruitment to ribosomal RNA genes has been excluded. (*B*) Detailed results of competitive fragment recruitment to *Ca*. N. brevis and *N. maritimus* in the 70–89% identity band from *A* indicating the fraction of total reads recruited to each genome. Metagenomic dataset numbers refer to the following accession numbers (preceeded by CAM_) in the CAMERA database: 1, PROJ_AntarcticaAquatic; 2, PROJ_BATS; 3, PROJ_Bacterioplankton; 4, PROJ_BotanyBay; 5, PROJ_HOT; 6, PROJ_LineIsland; 7, PROJ_MontereyBay; 8, PROJ_PML; 9, PROJ_PeruMarginSediment; 10, PROJ_SapeloIsland; 11, PROJ_SargassoSea; 12, PROJ_WesternChannelOMM; 13, P0000712; 14, P0000715; 15, P0000719; 16, P0000828; and 17, P0001028. Details of each metagenomic dataset are provided in *SI Appendix*, Table S8.

Materials and Methods

Cultivation and Genome Sequencing. The enrichment culture CN25 was grown under ammonia-oxidizing conditions in natural seawater-based oligotrophic north Pacific (ONP) medium (26) with 100 μM added NH_4Cl and harvested onto 0.2-μm pore size filters. DNA was extracted using a modified phenol-chloroform extraction. Metagenomic sequencing was done on the Illumina HiSeq platform after paired-end library construction with a 2-Kbp insert size at the University of Maryland Institute for Genome Sciences Genomics Resource Center. Complete details can be found in the *SI Appendix, Materials and Methods*. The *Ca*. N. brevis CN25 genome has been deposited in the National Center for Biotechnology Information's GenBank repository under accession number CP007026.

Scanning Electron Microscopy. One hundred microliters CN25 culture was prefiltered through a 0.45-μm pore size syringe filter and then gently vacuum filtered onto 25 mm, 0.2 μm polycarbonate membrane filters, sequentially dehydrated, sputter coated, and prepared for observation with a Zeiss Supra 40VP scanning electron microscopy. Complete details can be found in the *SI Appendix, Materials and Methods*.

Genome Annotation and Metabolic Reconstruction. Gene prediction and annotation were done using both the J. Craig Venter Institute's microbial genome automated annotation pipeline and the Joint Genome Institute's Integrated Microbial Genomes pipeline with subsequent manual investigation. Complete details can be found in the *SI Appendix, Materials and Methods*.

Global Proteome. Early stationary phase CN25 cells grown under ammonia-oxidizing conditions in ONP medium were harvested by vacuum filtration onto single 0.2-μm pore size filters and frozen at –80 °C. Proteins were extracted using SDS extraction buffer, trypsin digested, purified, and concentrated. Proteins were identified by LC-MS of protein extracts using both one-dimensional and two-dimensional fractional chromatography. Mass spectral libraries were searched using SEQUEST HT (v 1.4). Complete details can be found in the *SI Appendix, Materials and Methods*.

Comparative Genomics and Phylogenetic Analysis. Ortholog clustering was conducted using CD-Hit at the indicated alignment cutoffs with subsequent pairwise BLASTP alignments. Phylogenetic analysis was done using a concatenated alignment of 43 ribosomal proteins, and a tree was generated as described in the *SI Appendix, Materials and Methods*.

Metagenomic Fragment Recruitment. Details of the competitive fragment recruitment analysis can be found in the *SI Appendix, Materials and Methods*.

ACKNOWLEDGMENTS. We thank Jason Smith and Marguerite Blum for obtaining seawater for cultivation and John McCutcheon for helpful comments on a previous version of the manuscript. This work was funded by startup funds from the University of Maryland Center for Environmental Science (to A.E.S.); National Science Foundation awards OCE-1260006 (to A.E.S.), OCE-1259994 (to C.L.D.), OCE-1031271, and OCE-1233261 (to M.A.S.); the Life Technologies Foundation and Beyster Fund of the San Diego Foundation to the J. Craig Venter Institute; and support from the Gordon and Betty Moore Foundation under awards GBMF2724, GBMF3782, and GBMF3934 (to M.A.S.) and GBMF3307 (to A.E.S.). A.E.S. is an associate in the Integrated Microbial Biodiversity program of the Canadian Institute for Advanced Research. This is University of Maryland Center for Environmental Science contribution number 4949.

1. Karner MB, DeLong EF, Karl DM (2001) Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* 409(6819):507–510.
2. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6(3):245–252.
3. Spang A, et al. (2010) Distinct gene set in two different lineages of ammonia-oxidizing archaea supports the phylum Thaumarchaeota. *Trends Microbiol* 18(8):331–340.
4. DeLong EF (1992) Archaea in coastal marine environments. *Proc Natl Acad Sci USA* 89(12): 5685–5689.
5. Tully BJ, Nelson WC, Heidelberg JF (2012) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* 14(1):254–267.
6. Stewart FJ, Ulloa O, DeLong EF (2012) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol* 14(1):23–40.
7. Hollibaugh JT, Gifford S, Sharma S, Bano N, Moran MA (2011) Metatranscriptomic analysis of ammonia-oxidizing organisms in an estuarine bacterioplankton assemblage. *ISME J* 5(5):866–878.
8. Baker BJ, Lesniewski RA, Dick GJ (2012) Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. *ISME J* 6(12):2269–2279.
9. Ingalls AE, et al. (2006) Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proc Natl Acad Sci USA* 103(17):6442–6447.
10. Santoro AE, Buchwald C, McIlvin MR, Casciotti KL (2011) Isotopic signature of N(2)O produced by marine ammonia-oxidizing archaea. *Science* 333(6047):1282–1285.
11. Könneke M, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437(7058):543–546.
12. Martens-Habbena W, Berube PM, Urakawa H, de la Torre JR, Stahl DA (2009) Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* 461(7266):976–979.
13. Qin W, et al. (2014) Marine ammonia-oxidizing archaeal isolates displate obligate mixotrophy and wide ecotypic variation. *Proc Natl Acad Sci USA*.
14. Tourna M, et al. (2011) Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. *Proc Natl Acad Sci USA* 108(20):8420–8425.
15. Lehtovirta-Morley LE, et al. (2014) Characterisation of terrestrial acidophilic archaeal ammonia oxidisers and their inhibition and stimulation by organic compounds. *FEMS Microbiol Ecol* 89(3):542–552.
16. Hallam SJ, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. *Proc Natl Acad Sci USA* 103(48):18296–18301.
17. Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR (2011) Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS ONE* 6(2):e16626.
18. Spang A, et al. (2012) The genome of the ammonia-oxidizing *Candidatus* Nitrososphaera gargensis: Insights into metabolic versatility and environmental adaptations. *Environ Microbiol* 14(12):3122–3145.
19. Hatzenpichler R, et al. (2008) A moderately thermophilic ammonia-oxidizing crenarchaeote from a hot spring. *Proc Natl Acad Sci USA* 105(6):2134–2139.
20. Lehtovirta-Morley LE, Stoecker K, Vilcinskas A, Prosser JI, Nicol GW (2011) Cultivation of an obligate acidophilic ammonia oxidizer from a nitrifying acid soil. *Proc Natl Acad Sci USA* 108(38):15892–15897.
21. Swan BK, et al. (2014) Genomic and metabolic diversity of Marine Group I Thaumarchaeota in the mesopelagic of two subtropical gyres. *PLoS ONE* 9(4):e95380.
22. Luo H, et al. (2014) Single-cell genomics shedding light on marine Thaumarchaeota diversification. *ISME J* 8(3):732–736.
23. Könneke M, et al. (2014) Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. *Proc Natl Acad Sci USA* 111(22): 8239–8244.
24. Walker CB, et al. (2010) *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* 107(19):8818–8823.
25. Arp DJ, Chain PSG, Klotz MG (2007) The impact of genome analyses on our understanding of ammonia-oxidizing bacteria. *Annu Rev Microbiol* 61(1):503–528.
26. Santoro AE, Casciotti KL (2011) Enrichment and characterization of ammonia-oxidizing archaea from the open ocean: Phylogeny, physiology and stable isotope fractionation. *ISME J* 5(11):1796–1808.
27. Pellitteri-Hahn MC, Halligan BD, Scalf M, Smith L, Hickey WJ (2011) Quantitative proteomic analysis of the chemolithoautotrophic bacterium *Nitrosomonas europaea*: Comparison of growing- and energy-starved cells. *J Proteomics* 74(4):411–419.
28. Schleper C, Jurgens G, Jonuscheit M (2005) Genomic studies of uncultivated archaea. *Nat Rev Microbiol* 3(6):479–488.
29. Markert S, et al. (2011) Status quo in physiological proteomics of the uncultured *Riftia pachyptila* endosymbiont. *Proteomics* 11(15):3106–3117.
30. Löscher C, et al. (2012) Production of oceanic nitrous oxide by ammonia-oxidizing archaea. *Biogeosciences* 9:2419–2429.
31. Stieglmeier M, et al. (2014) Aerobic nitrous oxide production through N-nitrosating hybrid formation in ammonia-oxidizing archaea. *ISME J* 8(5):1135–1146.
32. Beaumont HJE, Lens SI, Reijnders WNM, Westerhoff HV, van Spanning RJM (2004) Expression of nitrite reductase in *Nitrosomonas europaea* involves NsrR, a novel nitrite-sensitive transcription repressor. *Mol Microbiol* 54(1):148–158.
33. Zumft WG (1997) Cell biology and molecular basis of denitrification. *Microbiol Mol Biol Rev* 61(4):533–616.
34. Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS (2003) Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J Biol Chem* 278(42): 41148–41159.
35. Kim W, Major TA, Whitman WB (2005) Role of the precorrin 6-X reductase gene in cobamide biosynthesis in *Methanococcus maripaludis*. *Archaea* 1(6):375–384.
36. Sañudo-Wilhelmy SA, et al. (2012) Multiple B-vitamin depletion in large areas of the coastal ocean. *Proc Natl Acad Sci USA* 109(35):14041–14045.
37. Roth JR, Lawrence JG, Rubenfield M, Kieffer-Higgins S, Church GM (1993) Characterization of the cobalamin (vitamin B12) biosynthetic genes of *Salmonella typhimurium*. *J Bacteriol* 175(11):3303–3316.
38. Doxey AC, Kurtz DA, Lynch MDJ, Sauder LA, Neufeld JD (2014) Aquatic metagenomes implicate Thaumarchaeota in global cobalamin production. *ISME J*, 10.1038/ismej.2014.1142.
39. Giovannoni SJ, et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309(5738):1242–1245.
40. Carini P, et al. (2014) Discovery of a SAR11 growth requirement for thiamin's pyrimidine precursor and its distribution in the Sargasso Sea. *ISME J* 8(8):1727–1738.
41. Graupner M, White RH (2001) *Methanococcus jannaschii* generates L-proline by cyclization of L-ornithine. *J Bacteriol* 183(17):5203–5205.
42. Rozen Y, Dyk TK, LaRossa RA, Belkin S (2001) Seawater activation of *Escherichia coli* gene promoter elements: Dominance of *rpoS* control. *Microb Ecol* 42(4):635–643.
43. Amin SA, et al. (2013) Copper requirements of the ammonia-oxidizing archaeon *Nitrosopumilus maritimus* SCM1 and implications for nitrification in the marine environment. *Limnol Oceanogr* 58(6):2037–2045.
44. Martiny AC, Huang Y, Li W (2009) Occurrence of phosphate acquisition genes in Prochlorococcus cells from different ocean regions. *Environ Microbiol* 11(6):1340–1347.
45. Karl DM, et al. (2008) Aerobic production of methane in the sea. *Nat Geosci* 1(7): 473–478.
46. Metcalf WW, et al. (2012) Synthesis of methylphosphonic acid by marine microbes: A source for methane in the aerobic ocean. *Science* 337(6098):1104–1107.
47. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17(10):589–596.
48. Lynch M (2006) Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* 60:327–349.
49. Swan BK, et al. (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* 110(28):11463–11468.
50. Dupont CL, et al. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6(6):1186–1199.
51. Wolf YI, Koonin EV (2013) Genome reduction as the dominant mode of evolution. *BioEssays* 35(9):829–837.
52. Dufresne A, et al. (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci USA* 100(17):10020–10025.
53. Morris JJ, Lenski RE, Zinser ER (2012) The Black Queen Hypothesis: Evolution of dependencies through adaptive gene loss. *MBio* 3(2):e00036-12.
54. Facciotti MT, et al. (2007) General transcription factor specified global gene regulation in archaea. *Proc Natl Acad Sci USA* 104(11):4630–4635.
55. Decker KB, Hinton DM (2013) Transcription regulation at the core: Similarities among bacterial, archaeal, and eukaryotic RNA polymerases. *Annu Rev Microbiol* 67(67): 113–139.
56. Turkarslan S, et al. (2011) Niche adaptation by expansion and reprogramming of general transcription factors. *Mol Syst Biol* 7:554.
57. Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA (2011) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *ISME J* 5(3):461–472.
58. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3):e77.
59. Palenik B, et al. (2003) The genome of a motile marine *Synechococcus*. *Nature* 424(6952):1037–1042.
60. Rodriguez-Valera F, et al. (2009) Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* 7(11):828–836.
61. Coleman ML, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311(5768):1768–1770.
62. Grote J, et al. (2012) Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* 3(5):e00252-12.
63. Karl DM (1999) A sea of change: Biogeochemical variability in the North Pacific Subtropical Gyre. *Ecosystems (N Y)* 2(3):181–214.
64. Polovina JJ, Howell EA, Abecassis M (2008) Ocean's least productive waters are expanding. *Geophys Res Lett* 35(3):L03618.

Supporting Information Appendix for:

Genomic and proteomic characterization of '*Candidatus* Nitrosopelagicus brevis': an ammonia-oxidizing archaeon from the open ocean

Alyson E. Santoro[a], Chris L. Dupont[b], R. Alexander Richter[c], Matthew T. Craig[b,d], Paul Carini[a], Matthew R. McIlvin[e], Youngik Yang[c], William Orsi[a], Dawn Moran[e], Mak A. Saito[e]

This file includes:

SI Datasets provided under separate cover as Excel files:

Supporting Information: Materials and Methods

Cultivation, nucleic acid extraction, and genome sequencing

The enrichment culture CN25 was grown under ammonia-oxidizing conditions May-June 2012 in 250 mL polycarbonate bottles in natural seawater-based ONP medium with 100 µM added NH_4Cl at 22ºC as previously described (1). Cells from 1 L of culture were filtered onto 25 mm 0.2 µm pore size Supor filters (Pall) and DNA was extracted using a modified phenol-chloroform extraction. DNA was further purified and concentrated using Amicon Ultra spin filter units (Millipore) with a 30 KDal molecular weight cutoff and quantified using Quanti-T reagents and a Q-Bit fluorometer (Invitrogen). Approximately 500 ng of DNA was used for library preparation and sequencing.

DNA sequencing was done on the Illumina HiSeq platform following paired end library construction with a 2 Kbp insert size at the University of Maryland Institute for Genome Sciences Genomics Resource Center. An initial analysis of the reads revealed a bimodal %GC distribution with a large peak centered at 32 %GC and a smaller peak between approximately 50 and 65% GC, consistent with the relative percentage of bacterial contaminants in the culture (1). A phylogenetic analysis of the reads indicated the archaeal reads were found in the low GC cluster. An assembly using the Celera assembler using just the reads < 45% GC resulted in five initial contigs. Manual examination reconciled one gap between the contigs due to assembly error, while PCR reactions followed by direct Sanger sequencing reconciled a second. One contig with much lower coverage than the other contigs was found to be absent from genomic DNA from CN25 and subsequently excluded. This resulted in two contigs and two gaps. Manual examination of these contigs revealed matching but reverse orientation sequences linking the ends of each contig. That is, two ends of separate contigs shared inverse repeats of 850 bp (at 99% nt identity) with each other. The other two ends shared separate inverse repeats of 1300 bp (at 99% nt identity) with each other. Theorizing that these may be assembly errors, PCR reactions were performed to confirm the orientation and presence of each inverted repeat half on each contig. However, such inverse repeats are nearly impossible to amplify across and are unamenable to cloning. Therefore we are assuming that these repeats match to each other with no insert. Both inserts are present in single copy within the *N. maritimus* genome, which likely reflects the cloning host recombining out one half of the repeat during bacterial artificial chromosome generation, as is typical.

Electron microscopy

Scanning electron microscopy (SEM) imaging followed the method described in (2). The CN25 culture (100 mL) was gently filtered through a 0.45 µm syringe filter to reduce the abundance of larger bacterial cells, then vacuum filtered onto 25 mm, 0.2 µm polycarbonate membrane filters (Millipore GTTP). The filter was rinsed with 0.2 µm filtered seawater, and passed through a sequential dehydration series of 30, 50, 75, 90, and 100% ethanol before a final dehydration in hexamethyldisilazane (Sigma) and air-drying. For SEM observation, filters were attached to a carbon adhesive tab and mounted on a SEM specimen holder. Mounted specimens were then sputter coated with 10–15 nm of gold and palladium (60:40) using a Tousimis Samsputter 2A and visualized with a Zeiss Supra 40VP scanning electron microscope at the Marine Biological Laboratory, Woods Hole, Massachusetts. The most abundant cell type in the preparations were rods with a diameter of 0.17-0.26 µm and length of 0.6 -1.0 µm. Slightly larger, less abundant cells in the enrichment with evidence of flagella were also present. We assume here that the smaller, more abundant cells are *Ca.* N. brevis.

Temperature optimum determination and organic amendment experiments

Ca. N. brevis was grown in ONP medium as described above with 50 µM NH_4Cl, streptomycin (100 µg L^{-1}) and ampicillin (50 µg L^{-1}). For temperature optimum determination, triplicate 50 mL cultures were initiated by transferring 5 mL of exponential phase culture into 45 mL of medium and grown in the dark in 60 mL acid-cleaned polycarbonate bottles at 9, 16, 22, 28, and 34ºC without shaking. To test the effect of organic amendments on the growth of *Ca.* N. brevis, the organic compounds shown in Table S3 were added to 50 mL cultures to a final concentration of 5 µM each. Growth in all experiments was monitored using the concentration of nitrite (NO_2^-) determined colorimetrically (3). CN25 growth rates determined using changes in [NO_2^-] are indistinguishable from growth rates calculated using cell counts (1).

Annotation and metabolic reconstruction

Gene prediction and annotation were done using both the J. Craig Venter Institute's microbial genome automated annotation pipeline and the Joint Genome Institute's Integrated Microbial Genomes (JGI IMG) pipeline with subsequent manual investigation using IMG Expert Review (IMG/ER, (4)). KEGG annotations were conducted using KASS, with subsequent manual annotation. COG annotations were made in IMG (5). In addition to IMG, putative transport proteins were identified using TransAAP (6). The genome was searched for putative CRISPR regions using CRISPRFinder (7). The presence of integrative elements was investigated using BLASTP queries of putative integrases identified in other thaumarchaeal genomes against the *Ca.* N. brevis genome assembly. Additional manual curation of select pathways was done using KEGG pathway mapping and reciprocal best BLAST searches against available microbial genomes in IMG, and HMMR searches against the NCBI nr database (8).

Comparative genomics and phylogenetic analysis

Ortholog clustering was conducted using CD-Hit at the indicated alignment cutoffs with subsequent pairwise BLASTP alignments to determine ortholog identity of the *Ca.* N. brevis proteins. In parallel, all peptides from the query genome were blasted against all other peptides in the subject genome (all vs. all BLAST), requiring 90% alignment length to the query sequence.

Using the archaeal ribosomal protein alignments from Yutin and coworkers (9) we generated HMMER-3 profiles. We then searched the predicted proteomes against the profiles with hmmsearch at an e-value cutoff of 1e-10 and took the top hit against the profile for each genome as the predicted homolog. Using hmmalign, these predicted homologues were then aligned against the profile and reconciled where possible against each other. The ribosomal alignments for which all members had a representative were then concatenated, and a tree was generated using FastTree (10, 11) with the parameter -wag.

The proteins used, in order of concatenation, were: L2p, L3p, L4p, L5p, L6p, L13p, L14p, L15p, L22p, L23p, L24p, L29p, L30p, S2p, S3p, S4p, S5p, S7p, S8p, S9p, S10p, S11p, S12p, S13p, S14p, S15p, S17p, S19p, L7ae, L15e, L10e, L18e, L24e, L37ae, L44e, S17e, S19e, S24e, S27e, S28e, S4e, S6e, S8e. The total length of the concatenated alignment was 8,794 positions. The longest member of the alignment had 7,168 aa among those positions. The additional reference genomes added to the analysis of Yutin and coworkers were *Candidatus* Ca. N. limnia SFB1 (gb|AEGP00000000), *Candidatus* Nitrosopumilus salaria BD31 (gb|AEXL02000000), *Candidatus* Nitrososphaera gargensis Ga9.2 (ref|NC018719), *Candidatus* Nitrosopumilus koreensis AR1 (gb|CP003842), *Candidatus* Nitrosoarchaeum koreensis MY1 (gb|AFPU01000001), and *Candidatus* Ca. N. limnia BG20 (gb|AHJG00000000). The alignment and tree are available on request (C. L. D).

Metagenomic fragment recruitment

Competitive fragment recruitment against the *Ca.* N. brevis and *N. maritimus* SCM1 genomes was conducted as described in (12). Briefly, alignments via blastn to an in-house genome database (including

the nr database from NCBI and recent single cell genomes obtained from JGI) identified metagenomic reads with highest affinity to Thaumarchaeota. This subset of metagenomic reads was then aligned to the *Ca*. N. brevis and *N. maritimus* genomes, with only the best hits counted, that is, a sequence recruited with higher identity to *N. maritimus* was not recruited to *Ca.* N. brevis, making the recruitment competitive. Recruitment was parsed according to the percent identity (%ID) to the best hit genome, with reads only being counted once according the %ID bandwidth described. For example, once recruited to the > 90%ID bandwidth, the read was excluded from the analysis at the 70%ID bandwidth.

Protein extraction and digestion

CN25 was grown in natural seawater-based ONP medium (1) under ammonia-oxidizing conditions. Early stationary phase CN25 cells were harvested by vacuum filtration onto single 25 mm, 0.2 µm pore size Supor membrane filters (Pall) and frozen at -80ºC. Sample #1 used 5 x ~500 mL of cells grown with 100 µM NH_4Cl (approximately 1.4 x 10^7 cells), Sample #2 used 3 x 250 mL of cells grown with 50 µM NH_4Cl (approximately 2.7 x 10^6 cells). SDS extraction buffer (1% SDS, 0.1 M Tris/HCl pH 7.5, 10 mM EDTA) was added to each filter and incubated at room temperature for 15 min, heated at 95ºC for 10 min and shaken at room temperature (RT) at 350 rpm for 1 h. Protein extract was removed from filter into a new tube and centrifuged for 30 min at 14,100 x *g* at RT. Supernatant was removed and concentrated in a 5000 MWCO filter (Sartorius Stedim Biotech Vivaspin) to ~300 µL. The sample was precipitated with cold 50% MeOH/50% acetone/0.5 mM HCl for 1 week at –20ºC, and centrifuged for 30 min at 4ºC and 14,100 x *g*. Supernatants were removed and pellets dried by vacuum centrifugation (Thermo Savant Waltham, MA) on low setting for 10 min or until completely dry. Pellets were resuspended in 40 µL of 1% SDS extraction buffer and quantified using a DC protein assay kit (Bio-Rad, Hercules, CA) with bovine serum albumin (BSA) as a standard.

Extracted proteins were purified from SDS detergent and digested while embedded within a polyacrylamide tube gel, modified from (13), followed by reduction and alkylation, and trypsin digestion overnight. The tube gel approach allowed all proteins including membrane proteins to be solubilized by detergent and purified while immobilized in the gel matrix. A gel premix was made by combining 1 M Tris HCL (pH 7.5) and 40% Bis-acrylimide L 29:1 (Acros Organics) at a ratio of 1:3. The premix (103 µL) was combined with an extracted protein sample (usually 25 µg-200 µg), TE, 7 µL 1% APS and 3 µL of TEMED (Acros Organics) to a final volume of 200 µL. After 1 h of polymerization at room temperature (RT), 200 µL of gel fix solution (50% ETOH, 10% acetic acid in LC/MS grade water) was added to the top of the gel and incubated at RT for 20 min. Liquid was then removed and the tube gel was transferred into a new 1.5 mL microtube containing 1.2 mL of gel fix solution, then incubated at RT with gentle mixing (350 rpm in a Thermomixer R (Eppendorf)) for 1 h. Gel fix solution was then removed and replaced with 1.2 mL destain solution (50% MeOH, 10% acetic acid in water) and incubated again at RT with gentle mixing at 350 rpm for 2 h. Liquid was then removed, the gel was cut up into 1 mm cubes, then added back to tubes containing 1 mL of 50:50 acetonitrile:25 mM ammonium bicarbonate (ambic) incubated for 1 h at 350 rpm at RT. Liquid was removed and gel pieces were washed with 1ml of 25 mM ambic at 16°C 350 rpm for 1h. Gel pieces were then dehydrated twice in 800 µL of acetonitrile for 10 min at RT and dried for 10 min by vacuum centrifugation after removing solvent. 600 µL of 10 mM dithiothreitol (DTT) in 25 mM ambic was added to reduce proteins incubating at 56°C, 350 rpm for 1 h. Unabsorbed DTT solution was then removed with volume measured. Gel pieces were washed with 25 mM ambic and 600µl of 55 mM iodacetamide was added to alkylate proteins at RT, 350 rpm for 1h. Gel cubes were then washed with 1 mL ambic for 20 min, 350 rpm at RT. Acetonitrile dehydrations and vacuum centrifugation drying were repeated as above.

Trypsin (Promega) was added in appropriate volume of 25 mM ambic to rehydrate and submerse gel pieces at a concentration of 1:20 µg trypsin:protein. Proteins were digested overnight at 37°C while mixing at 350 rpm. Unabsorbed solution was removed and transferred to a new tube. 50 µL of peptide extraction buffer (50% acetonitrile, 5% formic acid in water) was added to gels, incubated for 20 min at RT then centrifuged at 14,100 x g for 2 min. Supernatant was collected and combined with unabsorbed

solution. The above peptide extraction step was repeated combining all supernatants. Combined protein extracts were centrifuged at 14,100 x *g* for 20 min, supernatants transferred into a new tube and dehydrated down to approximately 10 µL-20 µL by vacuum centrifugation. Concentrated peptides were then diluted in 2% acetonitrile 0.1% formic acid in water for storage until analysis. All water used in the tube gel digestion protocol was LC/MS grade, and all plastic microtubes were ethanol rinsed and dried prior to use.

Global proteome analyses

Proteins were identified by liquid chromatography/mass spectrometry (LC/MS) of protein extracts using both 1-dimensional (1-D) and 2-dimensional (2-D) fractional chromatography. For 1-D chromatography, each sample (2 mg protein measured before tryptic digestion) was concentrated onto a trap column (0.3 x 10 mm ID, 3 µm particle size, 200 Å pore size, SGE Protecol C18G) and rinsed with 150 mL 0.1% formic acid, 5% acetonitrile (ACN), 94.9% water before gradient elution through a reverse phase C18 column (0.15 x 150 mm ID, 3 µm particle size, 200 Å pore size, SGE Protecol C18G) on an Advance high performance liquid chromatography (HPLC) system (Michrom Bioresources Inc.) at a flow rate of 1 µL/min. The chromatography consisted of a nonlinear gradient from 5% Buffer A to 95% Buffer B for 230 min, where A was 0.1% formic acid in water and B was 0.1% formic acid in ACN. A Q-Exactive Orbitrap trap mass spectrometer (Thermo Scientific Inc.) was used with an ADVANCE CaptiveSpray source (Michrom Bioresources Inc.). Each mass spectrometer was set to perform MS/MS on the top *n* ions using data-dependent settings ($n = 15$), and ions were monitored over a range of 380-2000 m/z.

2-D chromatography was performed by an initial off-line separation of tryptic digested protein (20 µg protein sample adjusted to pH 10 with ammonium hydroxide) injected onto a reverse phase PLRP-S column (0.2 x 150 mm, 3 µm particle size, 300 Å pore size, Michrom Bioresources Inc.) on a Paradigm MD4 HPLC at a flow rate of 2 mL/min. Peptides were eluted with a nonlinear gradient of 5% to 90% acetonitrile in 20 mM ammonium formate at pH 10. Fractions were collected every minute for 60 minutes and the first 30 fractions were combined with 56 µL of 0.1% formic acid, 2% ACN, 97.9% water, then combined with the following 30 fractions (fraction 1 with 31, 2 with 32, etc.). The 30 combined fractions were then analyzed following similar 1-D LCMS procedures described above, except with a shorter 60 min LC gradient.

Mass spectral libraries were searched using SEQUEST HT within Proteome Discoverer (version 1.4). SEQUEST HT mass tolerance parameters were set at +/- 10 ppm for parent ions and 0.02 Da for fragment ions on the Q-Exactive mass spectrometer. Minimum parent ion size was set at 380 m/z and fragment ion size was set at 100 m/z. Cysteine modification of 57.021 Da and potential modification of +15.995 Da for methionine and cysteine oxidation were incorporated. Protein identifications were made using LFDR scoring in Scaffold 4.0 (Proteome Software, Portland OR USA), with 99.0% peptide confidence level and a <1% False Discovery Rate.

1012 proteins were identified with a 0.19% FDR (99% confidence level) on the peptide level and a 4.8% FDR (98% confidence level) on the protein level, with 52640 spectra matching peptides out of 518826 total spectra from 63 LC/MS runs.

Table S1. Primers used for PCR confirmation of bioinformatically assembled (in silico) scaffolds. 5' and 3' ends refer to initial orientation in CLC Workbench.

	Primer Name	Sequence (5'-3')	Scaffold/Region	Expected Fragment Size (bp)	Result
1	SCF440site1RevB	GCAAAAACTTCCACAAACACAA	Scaffold 440 5' End	n/a	
2	SCF440site1ForB1	CTATTTCCACTTCCAAGAATTGGT	Scaffold 440 5' End	503	Success
3	SCF440site1ForB2	TTTGAATTTGAAAGGTCTGCAC	Scaffold 440 5' End	1006	Success
4	SCF440site1ForB3	GATCTAATCCTGAAAGATTCGCG	Scaffold 440 5' End	1278	Success
5	SCF440site3ForB	CATTTTGTGCAAGTTTTTCAATAT	Scaffold 440 3' End	n/a	
6	SCF440site3RevB1	CACACGAGTTGGACGTCAGTTAT	Scaffold 440 3' End	992	Success
7	SCF440site3RevB2	TCCTAGAAGCACCAATTGGTG	Scaffold 440 3' End	2054	Success
8	SCF440site3RevB3	CGTATCAATTGCAGACTTGAAAG	Scaffold 440 3' End	2605	Success
9	SCF441Site1For	GTTGCAGAGGCGTGCTTC	Scaffold 441 Whole	n/a	
10	SCF441Site1Rev1	GCTGGAGCCTTGATAGGTGTC	Scaffold 441 Whole	540	Fail
11	SCF441Site1Rev2	GCTGCACAACCAAGTTCCAC	Scaffold 441 Whole	1050	Fail
12	SCF441Site1Rev3	CATTTTGGTACGCCGCTG	Scaffold 441 Whole	1625	Fail
13	SCF442Site4Rev	CATTCTTCAATTGCAGTAGTTGG	Scaffold 442 5' End	n/a	
14	SCF442Site4For1	CGTCATTGTAGTCAACATATGCC	Scaffold 442 5' End	515	Success
15	SCF442Site4For2	CGTTCAAGACCAATACCACAACC	Scaffold 442 5' End	1000	Success
16	SCF442Site4For3	CTGGAGCGTATTTTGGAAATGC	Scaffold 442 5' End	1518	Success
17	SCF442Site4For4	GAGGGATTTGTCTTACGCG	Scaffold 442 5' End	2061	Success
18	SCF442site5For	CCAGTATCAATTATAGCAATCGTG	Scaffold 442 3' End	n/a	
19	SCF442site5Rev1	CCGATTGTTGCATCAATCGC	Scaffold 442 3' End	586	Success
20	SCF442site5Rev2	CAATTGGTATTTGCTCCTGGTG	Scaffold 442 3' End	1399	Success
21	SCF442site5rev3	ATACACAGATTGGGCCCCA	Scaffold 442 3' End	2850	Success
22	SCF443site4Rev	TGATGCAACAGAACGTGCAC	Scaffold 443 5' End		
23	SCF443site4For1	ATTGCTGCCCATTCATCAC	Scaffold 443 5' End	574	Success
24	SCF443site4For2	CGCCGTATGTGTCATCTTCGT	Scaffold 443 5' End	995	Success
25	SCF443site4For3	TCTACATCAGATGCGATACTTGAT	Scaffold 443 5' End	1567	Success
26	SCF443site5For	GCAGAAAATGCAGGTATGGATCC	Scaffold 443 3' End	n/a	
27	SCF443site5Rev1	ATGGACAATGGATAAGTCCTCAG	Scaffold 443 3' End	440	Success
28	SCF443site5Rev2	GCCATCAGCAATGTATGCATAC	Scaffold 443 3' End	979	Success
29	SCF443site5Rev3	CTCCGCCTCTTTCGTAAACTAAG	Scaffold 443 3' End	1583	Success
30	SCF444site2ForB	TTAATTACACCATCGGTTGGTCCT	Scaffold 444 3' End	n/a	
31	SCF444site2RevB1	CGATCTTGAATACACAGATTGGGC	Scaffold 444 3' End	445	Success
32	SCF444site1RevB	AACATGAATAAAGAATTAGGACG	Scaffold 444 5' End	n/a	Success

	Primer Name	Sequence (5'-3')	Scaffold/Region	Expected Fragment Size (bp)	Result
33	SCF444site1ForB1	CACCTCTTGATTCTGAAGGAATC	Scaffold 444 5' End	468	Success
34	SCF444site1ForB2	CTCCGCCTCTTTCGTAAACTAAG	Scaffold 444 5' End	921	Success

Table S2. A high fraction of the predicted *Ca.* N. brevis proteome is translated during stationary phase.

Organism	No. of samples or growth conditions	% coverage of predicted proteome	Reference
Nanoarchaeum equitans	2	85	(14)
Ignicococcus hospitalis	2	73	(14)
Ca. N. brevis	2	70	present study
Saccharomyces cerevisiae	2	67	(15)
Deinococcus radiodurans	15	61	(16)
Methylobacterium extorquens AM1	1	58	(17)
Methanococcus jannaschii	1	54	(18)
Prochlorococcus marinus CCMP1986 (MED4)	14	51	(19)
Rhodobacter sphaeroides	2	35	(20)
Rhodopseudomonas palustris	6	34	(21)
Nitrosomonas europaea	2	34	(22)
Prochlorococcus marinus CCMP1986 (MED4)	7	29	(19)
Nitrosomonas eutropha C91	1	24	(23)
Shewanella oneidensis MR-1	26	17	(24)
Pelagibacter ubique HTCC1062	4	16	(25)

Table S3. Growth of *Ca*. N. brevis in ONP medium with 5 μM additions of the indicated organic carbon compounds to medium with 50 μM added ammonium (NH_4Cl). No growth enhancement was observed relative to the ammonium-only control.

Compound	Final [NO_2^-] (μM)	Specific growth rate (d^{-1})
acetate	53.8	0.11
acetone	53.3	0.11
alanine	53.1	0.11
aspartate	53.5	0.11
citrate	52.4	0.11
ethanol	52.5	0.11
fumarate	53.6	0.11
glutamate	52.6	0.11
glycerol	52.8	0.11
glycolic acid	52.6	0.11
β-hydroxybutyrate	53.0	0.11
isocitrate	52.2	0.11
α-ketoglutarate	52.3	0.11
malic acid	52.3	0.11
methanol	52.8	0.11
methionine	53.6	0.11
oxaloacetate	51.4	0.11
pyruvate	51.9	0.11
sulfite	52.8	0.11
succinate	52.2	0.11
ammonium only control	53.0	0.11

Table S4. Average ortholog identity from BLAST queries between pairs of orthologous genes for select archaeal genomes. In parallel, all peptides from the query genome were blasted against all other peptides in the subject genome (all vs. all BLAST), requiring 90% alignment length to the query sequence, resulting in slightly different average identities depending on the direction of the comparison due to differing peptide lengths for orthologs in the genomes being compared.

	C. symbiosum	*Ca.* N. limnia SFB1	*Ca.* N. salaria	*Ca.* N. limnia BG20	*Ca.* N. koreensis AR1	*Ca.* N. koreensis MY1	*N. gargensis*	*N. maritimus*	*Ca.* N. brevis
C. symbiosum	100	62	58	58	64	66	35	72	78
Ca. N. limnia SFB1	59	99	74	84	82	84	39	85	86
Ca. N. salaria	57	77	100	73	80	79	36	83	81
Ca. N. limnia BG20	59	90	76	100	83	87	39	86	88
Ca. N. koreensis AR1	58	78	74	74	99	81	38	88	86
Ca. N. koreensis MY1	60	84	76	81	84	100	39	87	87
N. gargensis	53	63	56	58	62	64	99	69	72
N. maritimus	64	76	72	72	82	79	38	100	87
Ca. N. brevis	57	66	61	63	69	69	34	75	100

Table S5. Comparison of paralog abundance in select archaeal genomes using two different amino acid identity thresholds to define paralogs.

	70% ID threshold		50% ID threshold	
Organism	No.	No. per Mbp genome	No.	No. per Mbp genome
N. gargensis	107	38	198	70
Ca. N. salaria	61	39	98	62
C. symbiosum	41	20	73	36
Ca. N. limnia SFB1	31	18	59	34
N. maritimus	20	12	44	27
Ca. N. koreensis AR1	16	10	40	24
Methanococcus maripaludis S2	14	8	43	26
Sulfolobus acidocaldarius 639	9	4	47	21
Ca. N. brevis	5	4	15	12

Table S6. Abundance of putative transporters in thaumarchaeal genomes as classified in the IMG database. The final two columns indicate abundance of each transporter class normalized to genome size for *N. maritimus* and *Ca*. N. brevis. A complete list of putative transporters and the corresponding NCBI locus is given in the metabolic reconstruction *SI Dataset*.

Function ID	Name	*N. gargensis*	*C. symbiosum* A	*Ca*. N. limnia SFB1	*Ca*. N. koreensis MY1	*N. maritimus*	*Ca*. N. brevis	*N. maritimus* (per Mbp)	*Ca*. N. brevis (per Mbp)
TC:1.A.1	The Voltage-gated Ion Channel (VIC) Superfamily	1	0	0	0	1	0	0.6	0.0
TC:1.A.11	The Ammonia Transporter Channel (Amt) Family	3	2	2	2	2	2	1.2	1.6
TC:1.A.22	The Large Conductance Mechanosensitive Ion Channel (MscL) Family	1	0	1	1	0	0	0.0	0.0
TC:1.A.23	The Small Conductance Mechanosensitive Ion Channel (MscS) Family	6	1	3	2	5	1	3.0	0.8
TC:1.A.28	The Urea Transporter (UT) Family	1	0	0	0	0	0	0.0	0.0
TC:1.A.33	The Cation Channel-forming Heat Shock Protein-70 (Hsp70) Family	1	1	1	1	1	1	0.6	0.8
TC:1.A.35	The CorA Metal Ion Transporter (MIT) Family	2	1	2	1	2	1	1.2	0.8
TC:1.A.62	The Homotrimeric Cation Channel (TRIC) Family	1	0	1	1	1	1	0.6	0.8
TC:1.A.8	The Major Intrinsic Protein (MIP) Family	2	2	2	2	2	2	1.2	1.6
TC:2.A.1	The Major Facilitator Superfamily (MFS)	10	2	6	5	2	2	1.2	1.6
TC:2.A.19	The Ca2+:Cation Antiporter (CaCA) Family	2	1	1	0	1	1	0.6	0.8
TC:2.A.20	The Inorganic Phosphate Transporter (PiT) Family	1	0	1	1	0	1	0.0	0.8
TC:2.A.21	The Solute:Sodium Symporter (SSS) Family	1	1	0	0	0	1	0.0	0.8
TC:2.A.23	The Dicarboxylate/Amino Acid:Cation (Na+ or H+) Symporter (DAACS) Family	0	0	0	0	1	0	0.6	0.0
TC:2.A.37	The Monovalent Cation:Proton Antiporter-2 (CPA2) Family	7	2	3	4	2	2	1.2	1.6
TC:2.A.38	The K+ Transporter (Trk) Family	2	1	4	2	1	0	0.6	0.0
TC:2.A.39	The Nucleobase:Cation Symporter-1 (NCS1) Family	1	0	0	0	0	0	0.0	0.0
TC:2.A.4	The Cation Diffusion Facilitator (CDF) Family	4	0	2	2	3	0	1.8	0.0
TC:2.A.44	The Formate-Nitrite Transporter (FNT) Family	1	0	0	0	0	0	0.0	0.0
TC:2.A.5	The Zinc (Zn2+)-Iron (Fe2+) Permease (ZIP) Family	0	0	2	0	0	0	0.0	0.0
TC:2.A.50	The Glycerol Uptake (GUP) Family	0	0	0	0	0	0	0.0	0.0
TC:2.A.52	The Ni2+-Co2+ Transporter (NiCoT) Family	1	0	1	1	1	0	0.6	0.0
TC:2.A.55	The Metal Ion (Mn2+-iron) Transporter (Nramp) Family	0	1	0	1	1	1	0.6	0.8
TC:2.A.64	The Twin Arginine Targeting (Tat) Family	3	2	3	3	1	3	0.6	2.4
TC:2.A.7	The Drug/Metabolite Transporter (DMT) Superfamily	2	0	2	1	2	0	1.2	0.0
TC:2.A.76	The Resistance to Homoserine/Threonine (RhtB)	1	0	1	1	1	1	0.6	0.8

Function ID	Name	*N. gargensis*	*C. symbiosum* A	*Ca*. N. limnia SFB1	*Ca*. N. koreensis MY1	*N. maritimus*	*Ca*. N. brevis	*N. maritimus* (per Mbp)	*Ca*. N. brevis (per Mbp)
	Family								
TC:2.A.83	The Na+-dependent Bicarbonate Transporter (SBT) Family	0	0	2	0	1	1	0.6	0.8
TC:2.A.89	The Vacuolar Iron Transporter (VIT) Family	1	0	1	1	0	0	0.0	0.0
TC:2.A.95	The 6TMS Neutral Amino Acid Transporter (NAAT) Family	1	0	0	0	0	0	0.0	0.0
TC:3.A.1	**The ATP-binding Cassette (ABC) Superfamily**	**39**	**32**	**21**	**22**	**31**	**18**	**18.9**	**14.6**
TC:3.A.10	The H+-translocating Pyrophosphatase (H+-PPase) Family	1	1	1	1	1	1	0.6	0.8
TC:3.A.2	The H+- or Na+-translocating F-type, V-type and A-type ATPase (F-ATPase) Superfamily	8	8	8	8	8	8	4.9	6.5
TC:3.A.3	The P-type ATPase (P-ATPase) Superfamily	1	0	1	0	0	0	0.0	0.0
TC:3.A.4	The Arsenite-Antimonite (ArsAB) Efflux Family	1	0	0	0	0	0	0.0	0.0
TC:3.A.5	The General Secretory Pathway (Sec) Family	5	5	7	6	4	5	2.4	4.1
TC:3.C.1	The Na+ Transporting Methyltetrahydromethanopterin: Coenzyme M Methyltransferase (NaT-MMM) Family	1	1	1	1	1	0	0.6	0.0
TC:3.D.1	The H+ or Na+-translocating NADH Dehydrogenase (NDH) Family	7	5	6	5	9	6	5.5	4.9
TC:3.D.9	The H+-translocating F420H2 Dehydrogenase (F420H2DH) Family	0	1	0	0	2	0	1.2	0.0
TC:4.C.1	The Proposed Fatty Acid Transporter (FAT) Family	0	0	1	0	1	0	0.6	0.0
TC:5.A.1	The Disulfide Bond Oxidoreductase D (DsbD) Family	2	2	2	2	2	2	1.2	1.6
TC:5.A.4	The Prokaryotic Succinate Dehydrogenase (SDH) Family	3	3	3	3	2	3	1.2	2.4
TC:5.B.1	The Phagocyte (gp91phox) NADPH Oxidase Family	0	0	0	0	1	0	0.6	0.0
TC:8.A.1	The Membrane Fusion Protein (MFP) Family	0	0	0	0	0	0	0.0	0.0
TC:8.A.21	The Stomatin/Podocin/Band 7/Nephrosis.2/SPFH (Stomatin) Family	2	0	1	1	1	1	0.6	0.8
TC:8.A.7	The Phosphotransferase System Enzyme I (EI) Family	0	0	0	0	1	0	0.6	0.0
TC:9.A.10	The Iron/Lead Transporter (ILT) Superfamily	1	0	2	1	0	2	0.0	1.6
TC:9.A.29	The Putative 4-Toluene Sulfonate Uptake Permease (TSUP) Family	2	1	2	1	1	1	0.6	0.8
TC:9.A.30	The Tellurium Ion Resistance (TerC) Family	2	0	1	1	0	0	0.0	0.0
TC:9.A.40	The HlyC/CorC (HCC) Family	1	1	1	1	2	1	1.2	0.8
TC:9.A.41	The Capsular Polysaccharide Exporter (CPS-E) Family	0	0	1	0	0	1	0.0	0.8
TC:9.A.8	The Ferrous Iron Uptake (FeoB) Family	0	0	0	0	0	1	0.0	0.8
TC:9.B.20	The Putative Mg2+ Transporter-C (MgtC) Family	1	0	1	0	0	0	0.0	0.0

Function ID	Name	*N. gargensis*	*C. symbiosum* A	*Ca*. N. limnia SFB1	*Ca*. N. koreensis MY1	*N. maritimus*	*Ca*. N. brevis	*N. maritimus* (per Mbp)	*Ca*. N. brevis (per Mbp)
TC:9.B.27	The DedA or YdjX-Z (DedA) Family	2	2	2	2	2	2	1.2	1.6
TC:9.B.43	The YedZ (YedZ) Family	0	0	0	0	1	0	0.6	0.0
TC:9.B.62	The Copper Resistance (CopD) Family	2	2	2	2	2	2	1.2	1.6
TC:9.B.69	The Putative Cobalt Transporter (CbtAB) Family	3	2	2	3	2	2	1.2	1.6
TC:9.B.71	The Camphor Resistance (CrcB) Family	1	0	1	1	1	0	0.6	0.0
	Totals	141	83	108	93	106	77	64.6	62.6

Table S7. Gene content of the two *Ca.* N. brevis putative genomic islands, closest blastp match in the NCBI non-redundant (nr) database, percent amino acid identity, and presence/absence in the global proteome. N.D. indicates no significant blastp hits in NCBI nr.

NCBI Locus	JCVI Annotation	Closest match in NCBI nr	%ID	Detected in proteome?
Island 1				
T478_0129	beta-lactamase	*Ca*. N. koreensis AR1	62	
T478_0131	glycosyltransferase	*Ca*. N. koreensis MY1	59	+
T478_0130	aminotransferase	*Archaeoglobus veneficus*	44	+
T478_0132	glycosyltransferase group 1	*Ca*. N. koreensis	65	+
T478_0133	UDP-glucose 4-epimerase	*Thaumarcheota* archaeon N4	68	+
T478_0134	UDP-glucose 6-dehydrogenase	*Caldiarchaeum subterranium*	41	+
T478_0135	nucleotidyl transferase	*Caldiarchaeum subterranium*	45	
T478_0136	nucleotide sugar dehydrogenase	*Ca.* N. limnia BG20	64	+
T478_0137	asparagine synthase	*Ca*. N. koreensis MY1	56	
T478_0138	UDP glucose dehydrogenase	Cenarchaeum symbiosum	55	+
T478_0139	glycosyltransferase group 1	*Ca*. N. koreensis MY1	48	+
T478_0140	sulfotransferase	*Ocillatoria nigro-viridis*	38	+
T478_0141	hypothetical	*Zoellia galactanivorans*	41	+
T478_0142	hypothetical, pyruvate kinase domain	*Coccbyxa subellipsoidea* C-169	33	+
T478_0143	phosphodiesterase	*N. gargensis*	42	+
T478_0144	hypothetical	*Ca*. N. limnia BG20	41	
T478_0145	hypothetical	*Ca*. N. limnia BG20	54	+
T478_0146	arylsulfatase	*N. maritimus* SCM1	38	
T478_0147	3-beta hydroxysteroid dehydrogenase	*N. gargensis*	68	+
T478_0148	methyltransferase	*Singulisphaera acidiphila*	38	+
T478_0149	NAD dependent epimerase	*Dyadobacter beijingensis*	39	+
T478_0150	aminotransferase	*Selenomonas sp.*	30	+
T478_0152	phosphodiesterase	Acidobacteriaceae KBS96	23	+
T478_0151	sulfotransferase	*Ca.* Nitrosopumilus sp. AR	41	+
T478_0153	glycosyltransferase group 1	*Ca*. N. limnia BG20	50	+
T478_0154	mannosyltransferase	*Ca*. N. limnia BG20	37	+
T478_0155	polysaccharide biosynthesis protein	*Ca*. N. koreensis MY1	42	+
T478_0156	Wxcm-like protein	*Ca*. N. limnia BG20	61	+
T478_0157	DTDP-glucose 4,6-dehydratase	*Ca*. N. salaria	61	+
T478_0158	glucose-1-phosphate thymidylyltransferase	*Ca*. N. limnia BG20	69	+
T478_0159	O-methyltransferase	*Ca*. N. limnia BG20	55	
T478_0161	glycosyltransferase	*Ca*. N. limnia BG20	59	+
T478_0160	4-phosphopantetheinyl transferase	*Ca*. N. limnia BG20	48	
T478_0162	methylmalonyl-CoA epimerase	*Ca*. N. limnia BG20	59	+
T478_0163	acyl carrier protein	*Ca*. N. limnia BG20	51	+
T478_0164	FkbH-like	*Ca*. N. limnia BG20	53	+

NCBI Locus	JCVI Annotation	Closest match in NCBI nr	%ID	Detected in proteome?
T478_0165	acetyltransferase	*Ca*. N. limnia BG20	61	
T478_0166	xylanase	*Ca*. N. limnia BG20	65	+
T478_0167	glycosyltransferase group 1	*N. maritimus* SCM1	53	+
T478_0169	polysaccharide biosynthesis protein	*Methanocaldococcus jannaschii*	41	+
T478_0168	oxidoreductase	*Ca*. N. limnia BG20	47	+
T478_0170	NDP-hexose 2,3dehydratase	*Saccharophagus degradans*	48	+
T478_0171	glycosyltransferase group 2	*Ca*. Nitrosopumilus sp. SJ	63	+
T478_0172	UDP-N-acetylglucosamine 2-epimerase	*Ca*. N. koreensis MY1	32	+
T478_0173	carbamoyltransferase	*Nitrosopumilus maritimus* SCM1	81	
T478_0174	GDSL family lipase	*Nitrosopumilus maritimus* SCM1	31	+
T478_0175	DTDP-glucose 4,6-dehydratase	*Marinitoga piezophila*	40	+
T478_0176	GHMP kinase	*Ca*. N. koreensis AR1	42	
T478_0177	SIS domain protein	*Ca*. N. koreensis AR1	51	+
T478_0178	D,D-heptose 1,7-bisphophate phosphatase	*Anaerobaculum hydrogeniformans*	48	
T478_0179	reversibly glycosylated polypeptide	*Natrinema veriforme*	28	+
T478_0180	3-beta hydroxysteroid dehydrogenase	*Nitrosopumilus maritimus* SCM1	36	+
T478_0181	radical SAM/B12 binding domain	*Streptomyces argenteolus*	30	+
T478_0182	glycosyltransferase group 2	*Archaeoglobus sufaticallidus*	44	
T478_0183	dolichyl-phosphate-mannose-protein mannosyltransferase	Thaumarchaeote KM_74_H09	35	+
T478_0184	unknown membrane protein	*Ca.* Nitrosopumilus sp. AR	35	+
T478_0186	polysaccharide biosynthesis protein	*Ca*. N. salaria	64	+
T478_0185	GlcNAc-PI de-N-acetylase	*Ca.* Nitrosopumilus sp. SJ	61	+
T478_0187	formyltransferase	*Ca*. N. koreensis AR1	63	+
T478_0188	acetyltransferase	*Ponticaulus koreensis*	38	+
T478_0190	aceyltransferase	*Clostridium clariflavum*	34	+
T478_0189	NeuB family protein	*Ca*. N. limnia BG20	58	+
T478_0191	polysaccharide biosynthesis protein	*Ca*. N. koreensis MY1	61	+
T478_0192	cytidylyltransferase	*Ca*. N. limnia BG20	51	+
T478_0193	polysaccharide biosynthesis protein	*Ca*. N. limnia SFB1	35	+
T478_0194	MetW	*Ca*. N. limnia BG20	55	+
T478_0195	radical SAM/B12 binding	*Chlorobium ferroxidans*	35	+
T478_0196	YrbI family	*Ca*. N. limnia SFB1	65	+
T478_0197	NeuB family protein	*Ca*. N. limnia BG20	77	+
T478_0199	phosphoheptose isomerase	*Ca*. N. limnia SFB1	69	+
T478_0198	phosophoglucose isomerase	*Ca*. N. koreensis MY1	56	+
T478_0200	methylthioribose-1-phosphate isomerase	*Ca*. N. limnia BG20	83	+
T478_0202	hypothetical	*Ca.* Nitrosopumilus sp. AR	77	
Island 2				
T478_1394	thiouridylase	*Fusobacterium necrophorum*	33	

NCBI Locus	JCVI Annotation	Closest match in NCBI nr	%ID	Detected in proteome?
T478_1395	hypothetical	Thaumarchaeote KM3_85_E11	30	
T478_1396	phosphoribosyltransferase	*Mahella australiensis*	26	
T478_1397	PF09369 domain	*Ca.* N. salaria BD31	23	+
T478_1398	helicase C terminal domain	*Ca.* N. salaria BD31	28	+
T478_1399	glycoside hydrolase	N.D.	N.D.	
T478_1400	hypothetical	N.D.	N.D.	
T478_1401	hypothetical	*Leptospira santarosai*	34	
T478_1402	hypothetical	SCGC AB-629-I23	45	
T478_1403	hypothetical	N.D.	N.D.	
T478_1404	PD-(D/e)XK nuclease	*Prochlorococcus* phage Syn33	37	
T478_1405	hypothetical	N.D.	N.D.	
T478_1406	hypothetical	*Ca*. Nitrosopumilus sp. AR2	26	+
T478_1407	cytosine specific methylase	*Paenibacillus alvei*	41	
T478_1408	hypothetical	BAC HF4000APKG3B16	58	+

Table S8. Competitive metagenomic fragment recruitment to the *Ca.* N. brevis and *N. maritimus* genomes from selected marine metagenomes from the CAMERA database (http://camera.calit2.net). Recruitment to ribosomal RNA genes has been removed from the analysis. Dataset numbers in the first column refer to data labels in Fig. 3B of the main text. Competitive fragment recruitment to the GOS data is provided in Excel format as an *SI Dataset*.

Data set	CAMERA Accession Number	CAMERA Project Name	Data Type	90% ID		70% ID		50% ID	
				Ca. N. brevis	*N. maritimus*	*Ca*. N. brevis	*N. maritimus*	*Ca*. N. brevis	*N. maritimus*
	CAM_P_0000545	Guaymas DEEP study	Combined	1215	4054	34064	91065	9717	3801
	CAM_P_0000766	Bloomer DSW addition experiment	Combined	99	8	32703	1235	15424	1807
1	CAM_P_0000712	Bermuda Oceanic Microbial Observatory Course	Metagenome	2758	2	4352	1108	1133	902
2	CAM_P_0000715	Bloomer DOM addition	Metagenome	0	1	50027	84	21438	76
3	CAM_P_0000719	Monterey Bay transect CN207 sampling sites	Metagenome	1110	46	3701	2794	2197	2386
4	CAM_P_0000828	Moore Marine Phage/Virus Metagenomes	Metagenome	41	0	249	137	115	102
5	CAM_P_0001028	North Pacific metagenomes from. Monterey Bay to Open Ocean (CalCOFI Line 67) October 2007	Metagenome	10	93	1757	765	1764	1119
6	CAM_PROJ_AntarcticaAquatic	Antarctica Aquatic Microbial Metagenome	Metagenome	371	1950	33083	214742	24209	22653
7	CAM_PROJ_Bacterioplankton	Marine Bacterioplankton Metagenomes	Metagenome	104	2	390	234	695	680
8	CAM_PROJ_BATS	Metagenomic Analysis of the North Atlantic Spring Bloom	Metagenome	5907	16	5886	2141	2723	2890
9	CAM_PROJ_BotanyBay	Botany Bay Metagenomes	Metagenome	1892	549	4163	66684	2992	5534
10	CAM_PROJ_HOT	Microbial Community Genomics at the HOT/ALOHA	Metagenome	2068	775	34133	25241	7259	7457
11	CAM_PROJ_LineIsland	Marine Metagenome from Line Islands	Metagenome	12	2	424	429	56	81
12	CAM_PROJ_MontereyBay	Monterey Bay Microbial Study	Metagenome	83	38	699	2424	680	669
13	CAM_PROJ_PeruMarginSediment	Metagenomic signatures of the Peru Margin	Metagenome	0	0	196	249	31	56
14	CAM_PROJ_PML	Marine Metagenome from Coastal Waters project at Plymouth Marine Laboratory	Metagenome	0	0	273	243	452	434
15	CAM_PROJ_SapeloIsland	Sapelo Island Bacterioplankton Metagenome	Metagenome	0	8	82	98	3	11
16	CAM_PROJ_SargassoSea	Sargasso Sea Bacterioplankton Community	Metagenome	5	0	880	114	739	143
17	CAM_PROJ_WesternChannelOMM	Western Channel Observatory Microbial Metagenomic Study	Metagenome	4351	532	23995	41415	2869	3071

Data set	CAMERA Accession Number	CAMERA Project Name	Data Type	90% ID		70% ID		50% ID	
				Ca. N. brevis	*N. maritimus*	*Ca*. N. brevis	*N. maritimus*	*Ca*. N. brevis	*N. maritimus*
	CAM_P_0001026	Lagrangian drifter transcriptomes	Metatranscriptome	201	136	504	2437	1019	17
	CAM_PROJ_AmazonRiverPlume	Microbial community gene expression across a productivity gradient of the Amazon River plume	Metatranscriptome	1	0	6771	322	4022	459
	CAM_PROJ_DICE	Dauphin Island Cubitainer Experiment (DICE)	Metatranscriptome	0	0	430	43	835	37
	CAM_PROJ_GeneExpression	Surface Water Marine Microbial Community Gene Expression	Metatranscriptome	1	0	3845	321	1938	286
	CAM_PROJ_PacificOcean	Influence of nitrogen-fixation on microbial community gene expression in the oligotrophic Southwest Pacific Ocean	Metatranscriptome	1	2	12208	399	8740	277
	CAM_PROJ_Sapelo2008	Sapelo Island Summer 2008 Bacterioplankton Metatranscriptome	Metatranscriptome	101	12622	18825	4452	4412	454

SI Figure Captions

Fig. S1. Scanning electron micrograph of putative *Ca.* N. brevis cells. **A**. Scale bar represents 1 μm. **B**. Scale bar represents 400 nm.

Fig. S2. Growth temperature optimum of *Ca.* N. brevis. Error bars are standard error of triplicate cultures and in some cases are smaller than the symbol.

Fig. S3. PCR confirmation of bioinformatically assembled (*in silico*) scaffolds. Unless otherwise indicated, the molecular size marker is the TrackIt 100 bp ladder (Invitrogen) with major size markers indicated in text. Primer numbers refer to Table S1. **A**. Scaffold 440: Lanes 1-3 contain products from primers 1-4; lanes 4-6 contain products from primers 5-8; lane 7 is a negative control with primer set 1+2. **B**. Scaffold 441: Lanes 1-3 contain products from primers 5-8, lane 4 is a negative control with primer set 5+6. **C**. Scaffold 442: Lanes 1-4 contain products from primers 13-17; Lanes 5-7 contain products from primers 18-21 in Table S1; Lane 8 is a negative control with primer set 13+14. **D**. Scaffold 443: Lanes 1-3 contain products from primers 22-25; Lanes 4-6 contain products from primers 26-29; Lane 7 is a negative control with primer set 22+23. **E.** Scaffold 444: Lanes 1-3 contain products from primers 30-34; Lane 4 is a negative control with primer set 30+31. Ladder is in house made 1 kb ladder with major size markers indicated in text.

Fig. S4. Genome size and gene count for select *Archaea* (n = 198) obtained from the JGI IMG database.

Fig. S5. Maximum likelihood phylogenetic tree including *Ca.* N. brevis based on a concatenated ribosomal protein alignment using WAG model of amino acid evolution and the discrete Gamma20 distribution model implemented using FastTree (11).

Fig. S6. The predicted proteomes of each of the indicated Thaumarchaeota was clustered using CD-Hit (26) at the indicated percent amino acid (AA) identity. Shown is the percent of the *Ca.* N. brevis predicted proteome shared in the other predicted proteomes for each identity cutoff relative to the average ortholog AA identity between the *Ca.* N. brevis and other Thaumarchaeota.

REFERENCES

1. Santoro AE & Casciotti KL (2011) Enrichment and characterization of ammonia-oxidizing archaea from the open ocean: Phylogeny, physiology, and stable isotope fractionation. *ISME J* 5:1796-1808.
2. Orsi W*, et al.* (2012) Class Cariacotrichea, a novel ciliate taxon from the anoxic Cariaco Basin, Venezuela. *Int J Syst Evol Microbiol* 62:1425-1433.
3. Strickland J & Parsons T (1968) A practical handbook of seawater analysis. *Fisheries Research Board of Canada Bulletin* 167:71-75.
4. Markowitz VM*, et al.* (2009) IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25(17):2271-2278.
5. Markowitz VM*, et al.* (2006) The integrated microbial genomes (IMG) system. *Nucleic Acids Res* 34(suppl 1):D344-D348.
6. Ren QH, Chen KX, & Paulsen IT (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic Acids Res* 35:D274-D279.
7. Grissa I, Vergnaud G, & Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35:W52-W57.
8. Finn RD, Clements J, & Eddy SR (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39:W29-W37.
9. Yutin N, Puigbo P, Koonin EV, & Wolf YI (2012) Phylogenomics of prokaryotic ribosomal proteins. *PLoS One* 7(5).
10. Price MN, Dehal PS, & Arkin AP (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26(7):1641-1650.
11. Price MN, Dehal PS, & Arkin AP (2010) FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 5(3).
12. Dupont CL*, et al.* (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6:1186–1199.
13. Lu X & Zhu H (2005) Tube-Gel digestion: A novel proteomic approach for high throughput analysis of membrane proteins. *Mol Cell Proteomics* 4(12):1948-1958.
14. Giannone RJ*, et al.* (2011) Proteomic characterization of cellular and molecular processes that enable the Nanoarchaeum equitans-Ignicoccus hospitalis relationship. *PLoS One* 6(8):e22942.
15. de Godoy LM*, et al.* (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455(7217):1251-1254.
16. Lipton MS*, et al.* (2002) Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proc Natl Acad Sci U S A* 99(17):11049-11054.
17. Bosch G*, et al.* (2008) Comprehensive proteomics of Methylobacterium extorquens AM1 metabolism under single carbon and nonmethylotrophic conditions. *Proteomics* 8(17):3494-3505.
18. Zhu W, Reich CI, Olsen GJ, Giometti CS, & Yates JR (2004) Shotgun proteomics of Methanococcus jannaschii and insights into methanogenesis. *J Proteome Res* 3(3):538-548.

19. Waldbauer JR, Rodrigue S, Coleman ML, & Chisholm SW (2012) Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. *PLoS One* 7(8):e43432.
20. Callister SJ*, et al.* (2006) Application of the accurate mass and time tag approach to the proteome analysis of sub-cellular fractions obtained from Rhodobacter sphaeroides 2.4.1. aerobic and photosynthetic cell cultures. *J Proteome Res* 5(8):1940-1947.
21. VerBerkmoes NC*, et al.* (2006) Determination and comparison of the baseline proteomes of the versatile microbe Rhodopseudomonas palustris under its major metabolic states. *J Proteome Res* 5(2):287-298.
22. Pellitteri-Hahn MC, Halligan BD, Scalf M, Smith L, & Hickey WJ (2011) Quantitative proteomic analysis of the chemolithoautotrophic bacterium Nitrosomonas europaea: Comparison of growing- and energy-starved cells. *Journal of Proteomics* 74(4):411-419.
23. Wessels HJCT, Gloerich J, der Biezen Ev, Jetten MSM, & Kartal B (2011) Liquid Chromatography—Mass Spectrometry-Based Proteomics of Nitrosomonas. *Methods Enzymol* 486:465-482.
24. Elias DA, Monroe ME, Smith RD, Fredrickson JK, & Lipton MS (2006) Confirmation of the expression of a large set of conserved hypothetical proteins in Shewanella oneidensis MR-1. *J Microbiol Methods* 66(2):223-233.
25. Smith DP*, et al.* (2010) Transcriptional and translational regulatory responses to iron limitation in the globally distributed marine bacterium candidatus Pelagibacter ubique. *PLoS One* 5(5):e10487.
26. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.

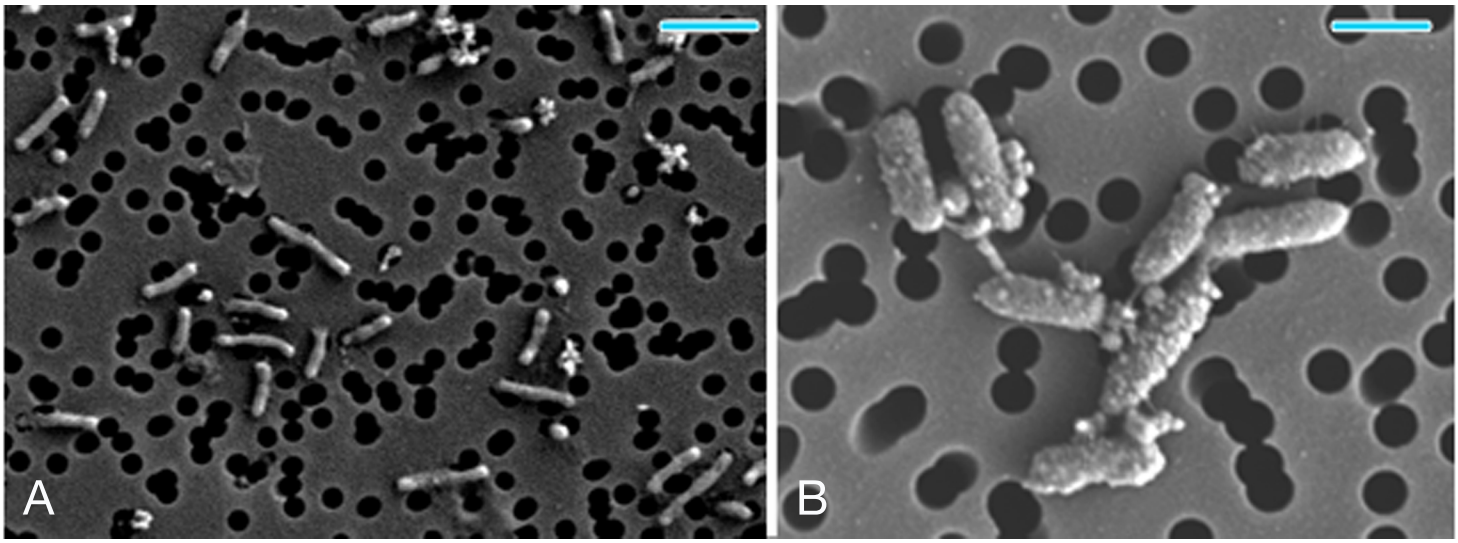


Fig. S1

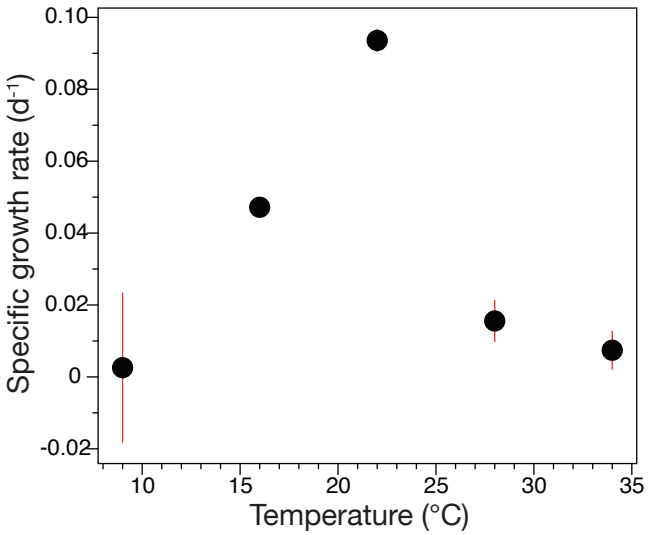
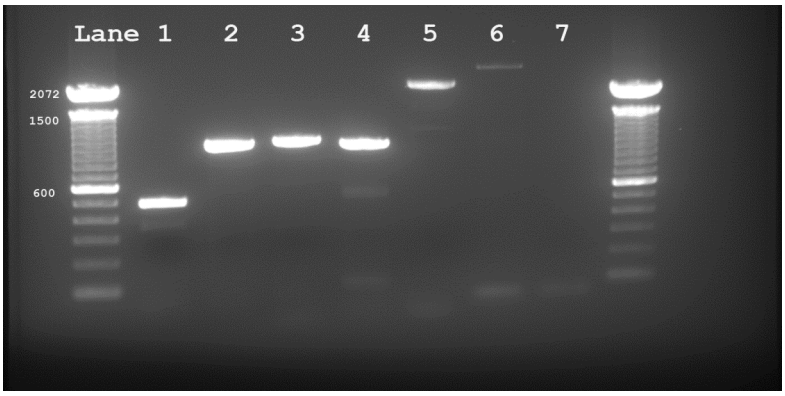

0.10
0.08
0.06
0.04
0.02
0
-0.02
Specific growth rate (d-1)
10
15
20
25
30
35
Temperature (°C)

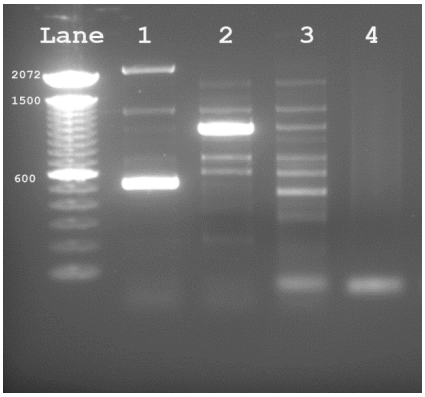
Fig. S2

Fig. S3

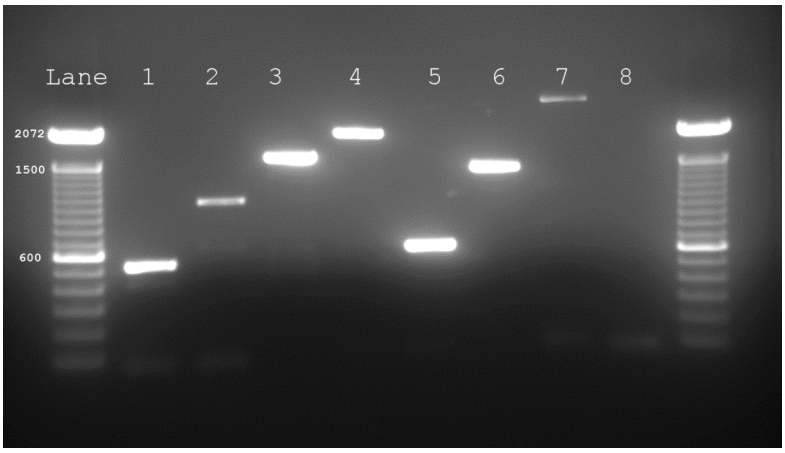
A.


Lane 1 2 3 4 5 6 7
2072
1500
600

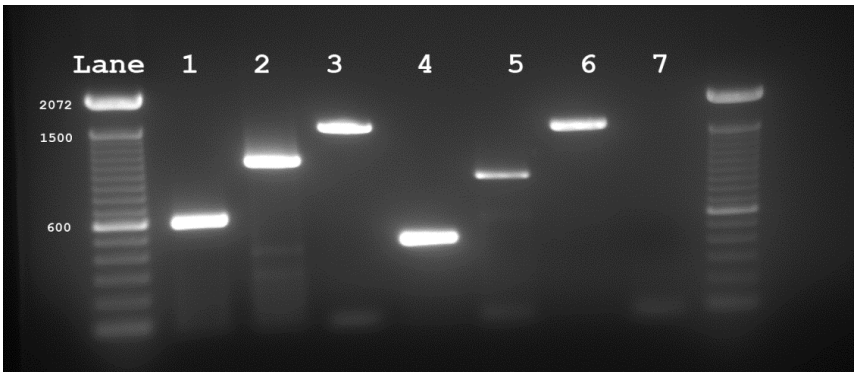
B.


Lane 1 2 3 4
2072
1500
600

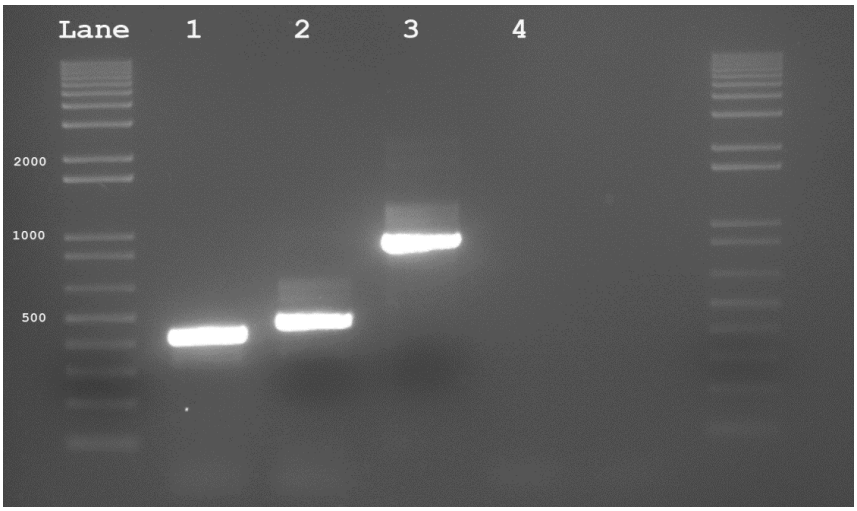
C.


Lane 1 2 3 4 5 6 7 8
2072
1500
600

D.



E.



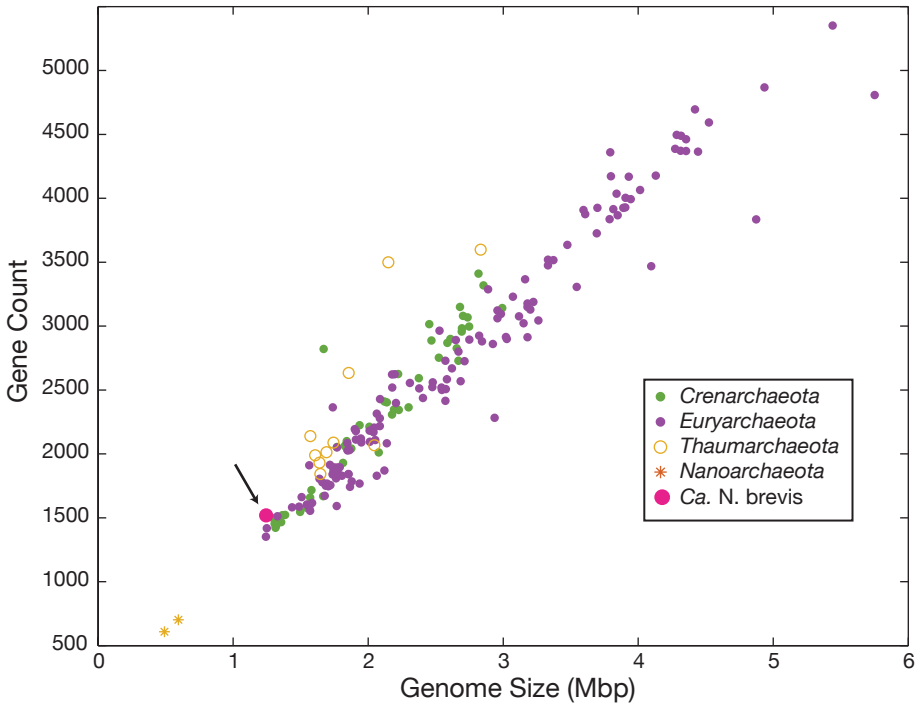

Gene Count
Genome Size (Mbp)
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
Crenarchaeota
Euryarchaeota
Thaumarchaeota
Nanoarchaeota
Ca. N. brevis

Fig. S4

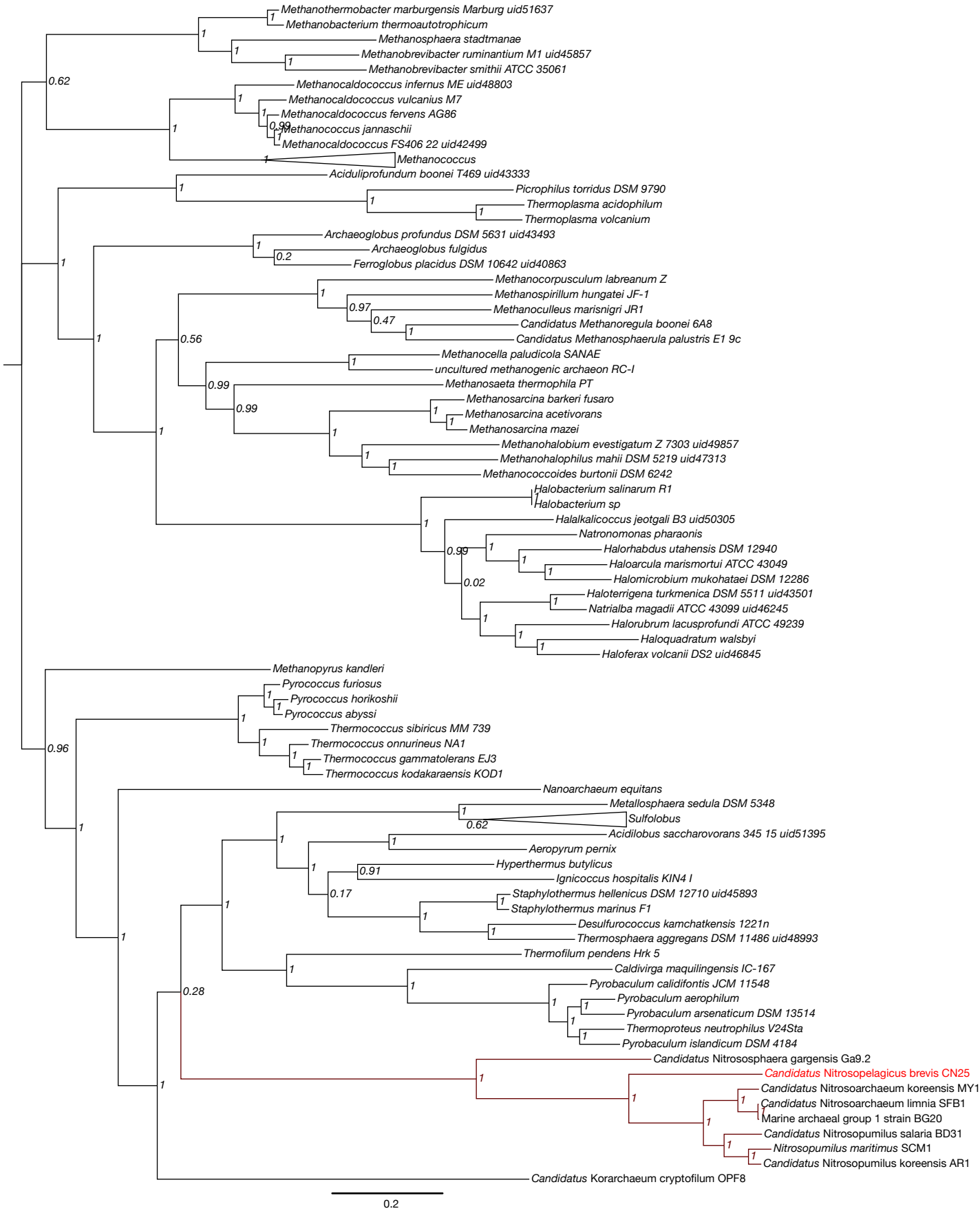

Methanothermobacter marburgensis Marburg uid51637
Methanobacterium thermoautotrophicum
Methanosphaera stadtmanae
Methanobrevibacter ruminantium M1 uid45857
Methanobrevibacter smithii ATCC 35061
Methanocaldococcus infernus ME uid48803
Methanocaldococcus vulcanius M7
Methanocaldococcus fervens AG86
Methanococcus jannaschii
Methanocaldococcus FS406 22 uid42499
Methanococcus
Aciduliprofundum boonei T469 uid43333
Picrophilus torridus DSM 9790
Thermoplasma acidophilum
Thermoplasma volcanium
Archaeoglobus profundus DSM 5631 uid43493
Archaeoglobus fulgidus
Ferroglobus placidus DSM 10642 uid40863
Methanocorpusculum labreanum Z
Methanospirillum hungatei JF-1
Methanoculleus marisnigri JR1
Candidatus Methanoregula boonei 6A8
Candidatus Methanosphaerula palustris E1 9c
Methanocella paludicola SANAE
uncultured methanogenic archaeon RC-I
Methanosaeta thermophila PT
Methanosarcina barkeri fusaro
Methanosarcina acetivorans
Methanosarcina mazei
Methanohalobium evestigatum Z 7303 uid49857
Methanohalophilus mahii DSM 5219 uid47313
Methanococcoides burtonii DSM 6242
Halobacterium salinarum R1
Halobacterium sp
Halalkalicoccus jeotgali B3 uid50305
Natronomonas pharaonis
Halorhabdus utahensis DSM 12940
Haloarcula marismortui ATCC 43049
Halomicrobium mukohataei DSM 12286
Haloterrigena turkmenica DSM 5511 uid43501
Natrialba magadii ATCC 43099 uid46245
Halorubrum lacusprofundi ATCC 49239
Haloquadratum walsbyi
Haloferax volcanii DS2 uid46845
Methanopyrus kandleri
Pyrococcus furiosus
Pyrococcus horikoshii
Pyrococcus abyssi
Thermococcus sibiricus MM 739
Thermococcus onnurineus NA1
Thermococcus gammatolerans EJ3
Thermococcus kodakaraensis KOD1
Nanoarchaeum equitans
Metallosphaera sedula DSM 5348
Sulfolobus
Acidilobus saccharovorans 345 15 uid51395
Aeropyrum pernix
Hyperthermus butylicus
Ignicoccus hospitalis KIN4 I
Staphylothermus hellenicus DSM 12710 uid45893
Staphylothermus marinus F1
Desulfurococcus kamchatkensis 1221n
Thermosphaera aggregans DSM 11486 uid48993
Thermofilum pendens Hrk 5
Caldivirga maquilingensis IC-167
Pyrobaculum calidifontis JCM 11548
Pyrobaculum aerophilum
Pyrobaculum arsenaticum DSM 13514
Thermoproteus neutrophilus V24Sta
Pyrobaculum islandicum DSM 4184
Candidatus Nitrososphaera gargensis Ga9.2
Candidatus Nitrosopelagicus brevis CN25
Candidatus Nitrosoarchaeum koreensis MY1
Candidatus Nitrosoarchaeum limnia SFB1
Marine archaeal group 1 strain BG20
Candidatus Nitrosopumilus salaria BD31
Nitrosopumilus maritimus SCM1
Candidatus Nitrosopumilus koreensis AR1
Candidatus Korarchaeum cryptofilum OPF8
0.2

Fig. S5

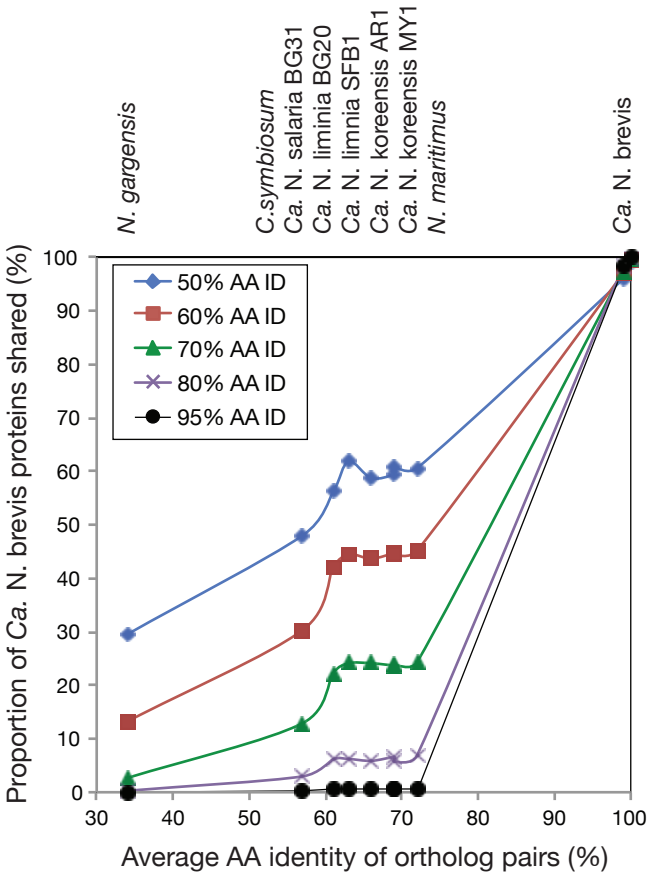

N. gargensis
C.symbiosum
Ca. N. salaria BG31
Ca. N. liminia BG20
Ca. N. limnia SFB1
Ca. N. koreensis AR1
Ca. N. koreensis MY1
N. maritimus
Ca. N. brevis
50% AA ID
60% AA ID
70% AA ID
80% AA ID
95% AA ID
Proportion of Ca. N. brevis proteins shared (%)
Average AA identity of ortholog pairs (%)
0
10
20
30
40
50
60
70
80
90
100
30
40
50
60
70
80
90
100

Fig. S6