**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 Materials and Methods*

Table S1
Table S2
Table S3
Table S4
Table S5
Table S6
Table S7
Table S8

Fig. S1
Fig. S2
Fig. S3 (a-e)
Fig. S4
Fig. S5
Fig. S6

**SI Datasets provided under separate cover as Excel files:**

Dataset S1. Complete proteome
Dataset S2. Metabolic reconstruction
Dataset S3. Genes unique to *Ca.* N. brevis
Dataset S4. Competitive metagenomic fragment recruitment to GOS data

## 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*.

| | | | | 90% ID | | 70% ID | | 50% ID | |
|---|---|---|---|---|---|---|---|---|---|
| Data set | CAMERA Accession Number | CAMERA Project Name | Data Type | *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 |

|  |  |  |  | 90% ID |  | 70% ID |  | 50% ID |  |
|---|---|---|---|---|---|---|---|---|---|
| Data set | CAMERA Accession Number | CAMERA Project Name | Data Type | *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.

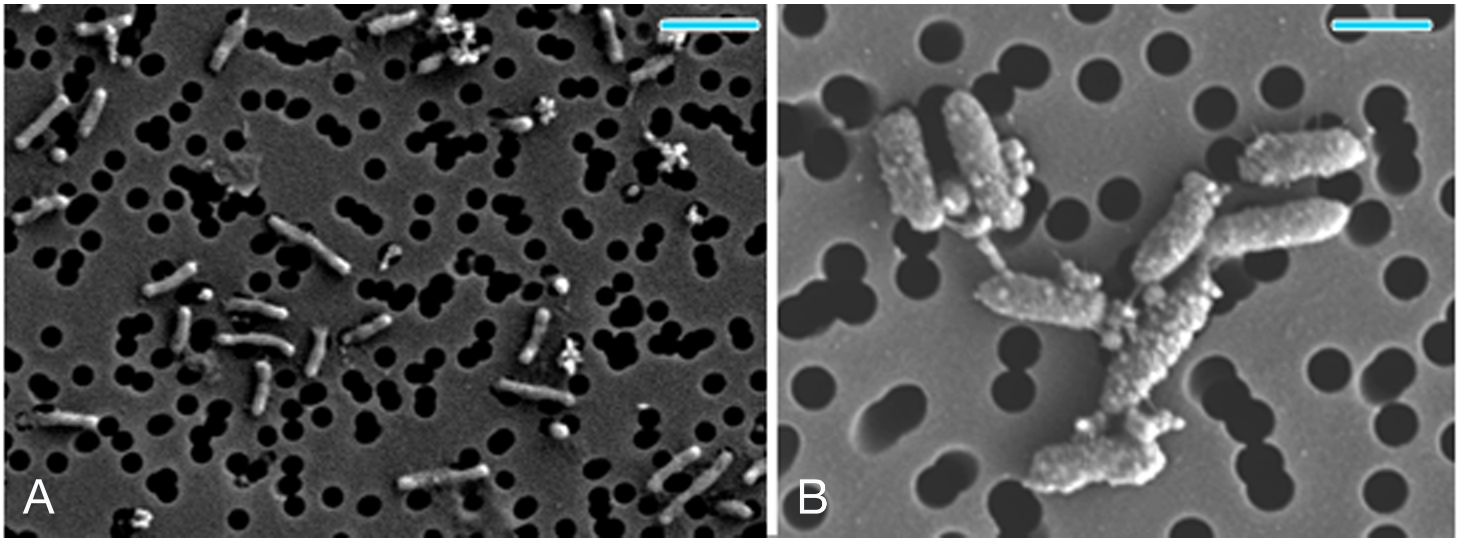


Fig. S1

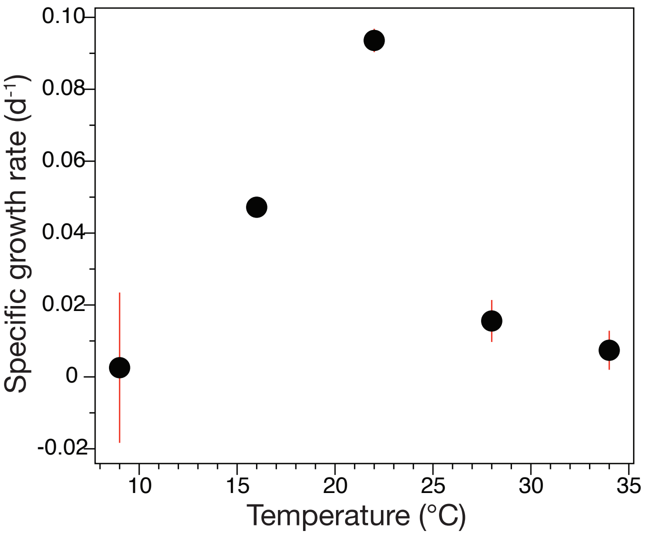
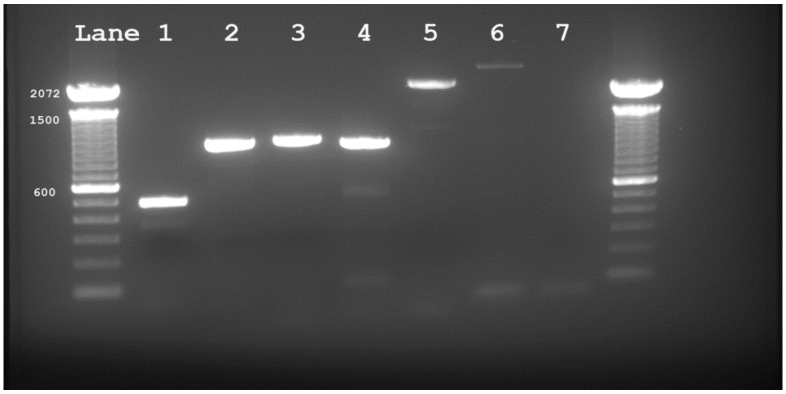

Specific growth rate (d-1)
0.10
0.08
0.06
0.04
0.02
0
-0.02
10
15
20
25
30
35
Temperature (°C)

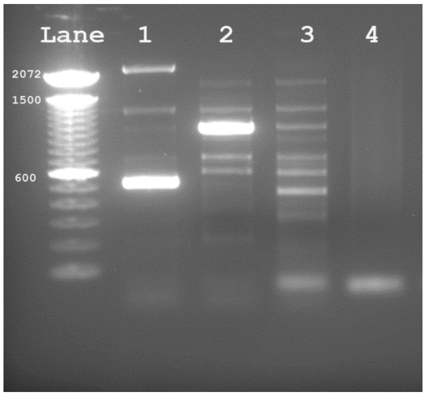
Fig. S2

Fig. S3

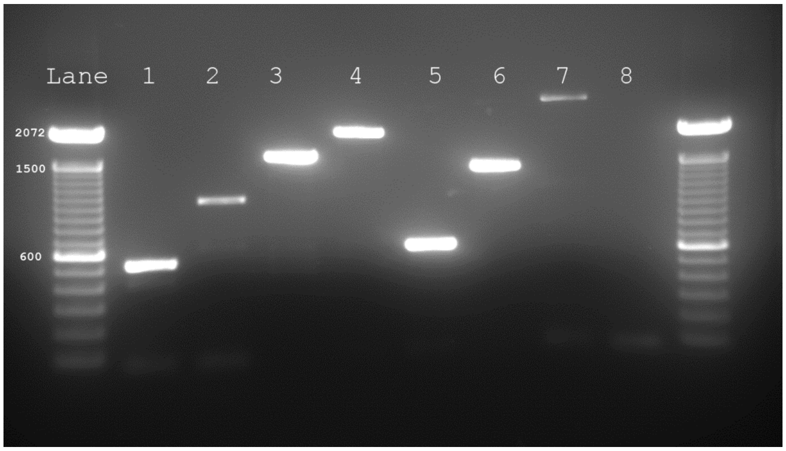
A.


Lane
1
2
3
4
5
6
7
2072
1500
600

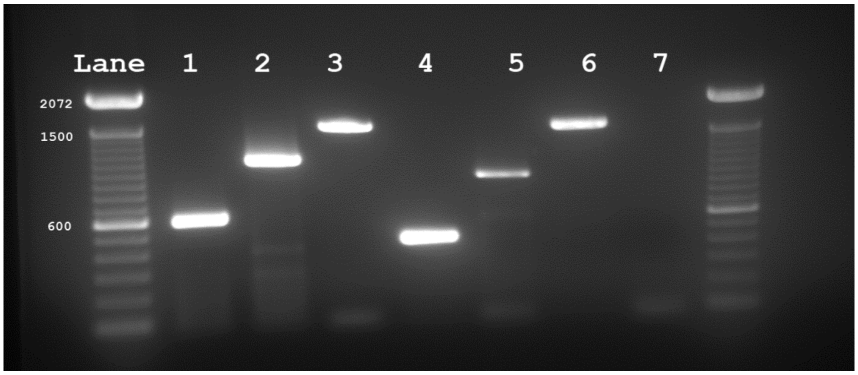
B.


Lane
1
2
3
4
2072
1500
600

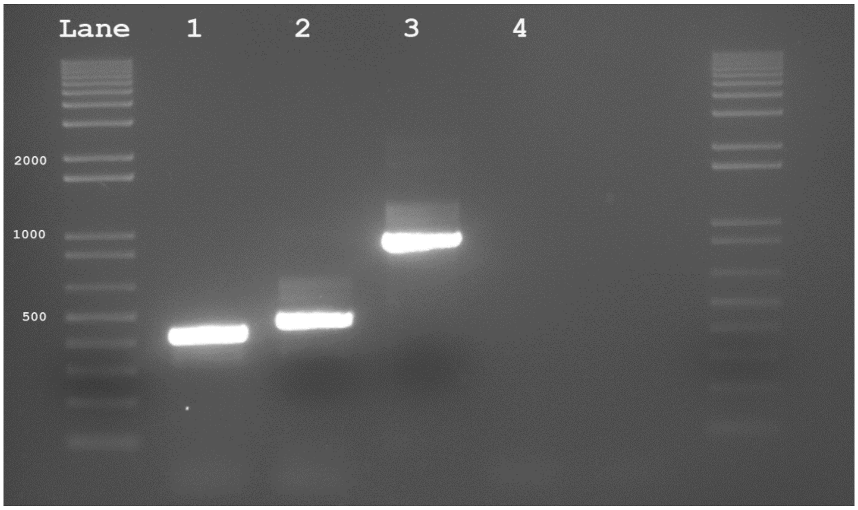
C.


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

D.



E.



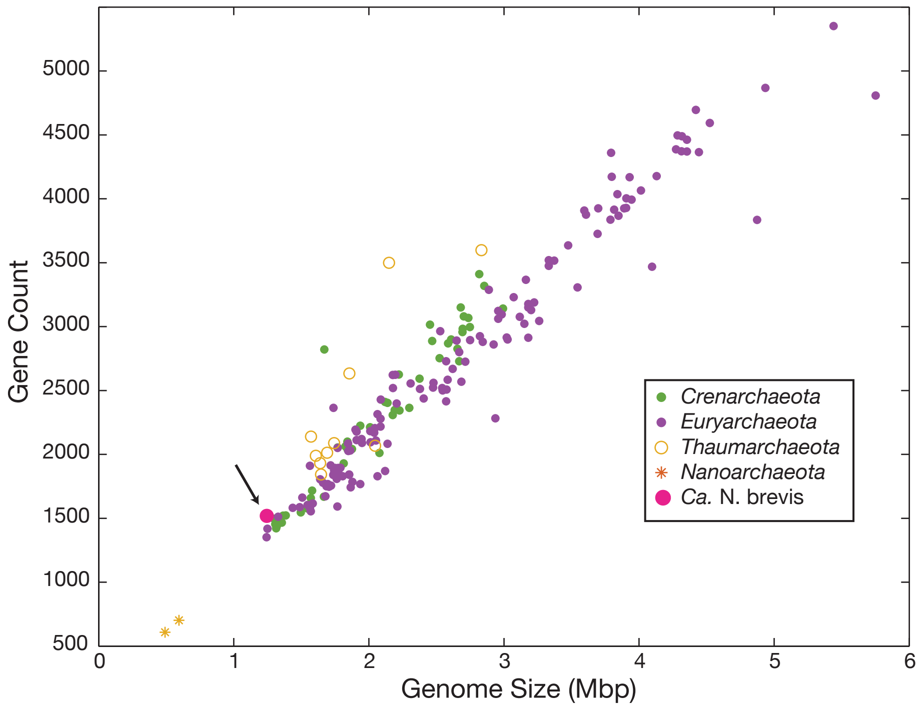

Gene Count
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
1
2
3
4
5
6
Genome Size (Mbp)
Crenarchaeota
Euryarchaeota
Thaumarchaeota
Nanoarchaeota
Ca. N. brevis

Fig. S4

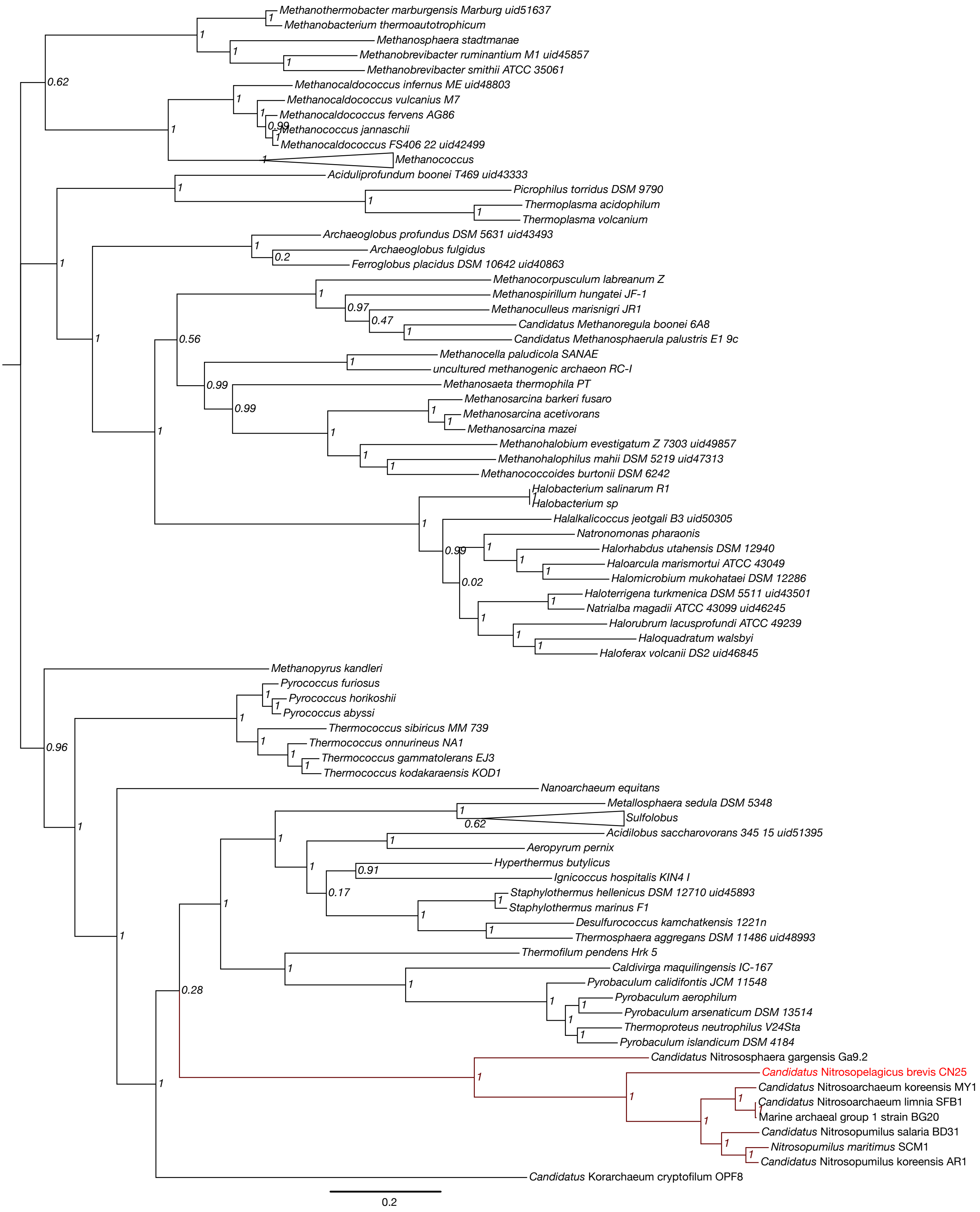

Methanothermobacter marburgensis Marburg uid51637
1
Methanobacterium thermoautotrophicum
1
Methanosphaera stadtmanae
1
Methanobrevibacter ruminantium M1 uid45857
1
Methanobrevibacter smithii ATCC 35061
0.62
Methanocaldococcus infernus ME uid48803
1
Methanocaldococcus vulcanius M7
1
Methanocaldococcus fervens AG86
0.99
Methanococcus jannaschii
1
Methanocaldococcus FS406 22 uid42499
1
1
Methanococcus
Aciduliprofundum boonei T469 uid43333
1
Picrophilus torridus DSM 9790
1
Thermoplasma acidophilum
1
Thermoplasma volcanium
Archaeoglobus profundus DSM 5631 uid43493
1
Archaeoglobus fulgidus
0.2
Ferroglobus placidus DSM 10642 uid40863
1
Methanocorpusculum labreanum Z
1
Methanospirillum hungatei JF-1
0.97
Methanoculleus marisnigri JR1
0.47
Candidatus Methanoregula boonei 6A8
1
Candidatus Methanosphaerula palustris E1 9c
1
0.56
Methanocella paludicola SANAE
1
uncultured methanogenic archaeon RC-I
0.99
Methanosaeta thermophila PT
0.99
Methanosarcina barkeri fusaro
1
Methanosarcina acetivorans
1
Methanosarcina mazei
1
Methanohalobium evestigatum Z 7303 uid49857
1
Methanohalophilus mahii DSM 5219 uid47313
1
Methanococcoides burtonii DSM 6242
1
Halobacterium salinarum R1
1
Halobacterium sp
1
Halalkalicoccus jeotgali B3 uid50305
Natronomonas pharaonis
0.99
1
Halorhabdus utahensis DSM 12940
1
Haloarcula marismortui ATCC 43049
1
Halomicrobium mukohataei DSM 12286
0.02
Haloterrigena turkmenica DSM 5511 uid43501
1
Natrialba magadii ATCC 43099 uid46245
1
Halorubrum lacusprofundi ATCC 49239
1
Haloquadratum walsbyi
1
Haloferax volcanii DS2 uid46845
Methanopyrus kandleri
Pyrococcus furiosus
1
Pyrococcus horikoshii
1
Pyrococcus abyssi
1
Thermococcus sibiricus MM 739
1
Thermococcus onnurineus NA1
1
Thermococcus gammatolerans EJ3
1
Thermococcus kodakaraensis KOD1
0.96
1
Nanoarchaeum equitans
Metallosphaera sedula DSM 5348
1
Sulfolobus
0.62
Acidilobus saccharovorans 345 15 uid51395
1
1
Aeropyrum pernix
1
Hyperthermus butylicus
0.91
Ignicoccus hospitalis KIN4 I
0.17
Staphylothermus hellenicus DSM 12710 uid45893
1
Staphylothermus marinus F1
1
Desulfurococcus kamchatkensis 1221n
1
Thermosphaera aggregans DSM 11486 uid48993
1
Thermofilum pendens Hrk 5
1
Caldivirga maquilingensis IC-167
1
Pyrobaculum calidifontis JCM 11548
1
Pyrobaculum aerophilum
1
Pyrobaculum arsenaticum DSM 13514
1
Thermoproteus neutrophilus V24Sta
1
Pyrobaculum islandicum DSM 4184
1
0.28
Candidatus Nitrososphaera gargensis Ga9.2
1
Candidatus Nitrosopelagicus brevis CN25
1
Candidatus Nitrosoarchaeum koreensis MY1
1
Candidatus Nitrosoarchaeum limnia SFB1
1
Marine archaeal group 1 strain BG20
1
Candidatus Nitrosopumilus salaria BD31
1
Nitrosopumilus maritimus SCM1
1
Candidatus Nitrosopumilus koreensis AR1
1
Candidatus Korarchaeum cryptofilum OPF8
0.2

Fig. S5

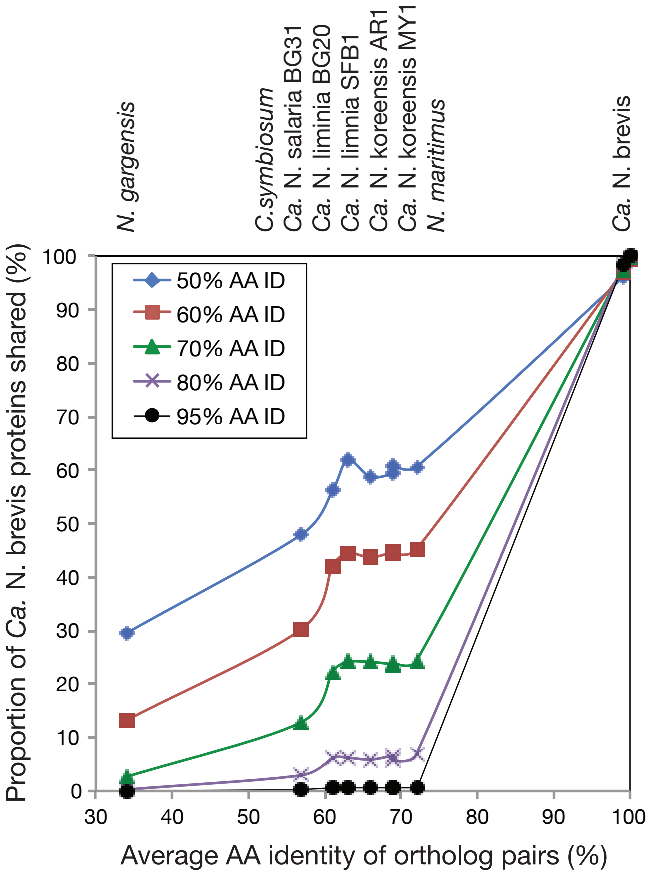

N. gargensis
C.symbiosum
Ca. N. salaria BG31
Ca. N. liminia BG20
Ca. N. limnia SFB1
Ca. N. koreensis AR1
Ca. N. koreensis MY1
N. maritimus
Ca. N. brevis
50% AA ID
60% AA ID
70% AA ID
80% AA ID
95% AA ID
Proportion of Ca. N. brevis proteins shared (%)
Average AA identity of ortholog pairs (%)
0
10
20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

Fig. S6