environmental microbiology

Environmental Microbiology (2017) **19**(6), 2434–2452 doi:10.1111/1462-2920.13768

Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems

Nathan A. Ahlgren ,[1*†] Yangyang Chen,[2,3,4] David M. Needham,[1] Alma E. Parada,[1‡] Rohan Sachdeva,[1] Vickie Trinh,[1] Ting Chen[5] and Jed A. Fuhrman[1]

[1]*Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA.*

[2]*College of Environmental Science and Engineering, Ocean University of China, Qingdao, China.*

[3]*Key Laboratory of Marine Environment and Ecology, Ministry of Education, Qingdao, China.*

[4]*Laboratory for Marine Ecology and Environmental Science, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China.*

[5]*Bioinformatics Division, TNLIST, Department of Computer Science and Technology, Tsinghua University, Beijing, China.*

Summary

Marine Thaumarchaeota are abundant ammonia-oxidizers but have few representative laboratory-cultured strains. We report the cultivation of *Candidatus Nitrosomarinus catalina* SPOT01, a novel strain that is less warm-temperature tolerant than other cultivated Thaumarchaeota. Using metagenomic recruitment, strain SPOT01 comprises a major portion of Thaumarchaeota (4–54%) in temperate Pacific waters. Its complete 1.36 Mbp genome possesses several distinguishing features: putative phosphorothioation (PT) DNA modification genes; a region containing probable viral genes; and putative urea utilization genes. The PT modification genes and an adjacent putative restriction enzyme (RE) operon likely form a restriction modification (RM) system for defence from foreign DNA. PacBio sequencing showed >98% methylation at two motifs, and inferred PT guanine modification of 19% of possible TGCA sites. Metagenomic recruitment also reveals the putative virus region and PT modification and RE genes are present in 18–26%, 9–14% and <1.5% of natural populations at 150 m with ≥85% identity to strain SPOT01. The presence of multiple probable RM systems in a highly streamlined genome suggests a surprising importance for defence from foreign DNA for dilute populations that infrequently encounter viruses or other cells. This new strain provides new insights into the ecology, including viral interactions, of this important group of marine microbes.

Introduction

Discovered about 25 years ago (Fuhrman *et al.*, 1992), planktonic, mesophilic archaea are recognized as abundant and important microbes in the oceans, making up significant portions of marine microbial communities especially in deeper waters below the euphotic zone (Karner *et al.*, 2001; Teira *et al.*, 2004; Teira *et al.*, 2006). In the water column, archaea predominantly belong to two phyla, the Thaumarchaeota, which were formerly known as Marine Group I Crenarchaea (Brochier-Armanet *et al.*, 2008), and the Euryarchaea, which include Marine Group II and III archaea (Delong, 1992; Fuhrman *et al.*, 1993). There are no cultured isolates of Marine Group II or III archaea, and only recently have marine Thaumarchaeota been brought into culture, either as pure isolates or enrichment cultures (Könneke *et al.*, 2005; Santoro and Casciotti, 2011; Qin *et al.*, 2014; Bayer *et al.*, 2016). Additional partial or nearly complete genomes have advanced our knowledge of marine archaea genomics for both Thaumarchaeota and Marine Group II and III archaea using culture-independent single-cell amplified genome (SAG) techniques (Swan *et al.*, 2014), sequencing of fosmids (Deschamps *et al.*, 2014) and assembly from metagenomes (Iverson *et al.*, 2012).

Received 13 December, 2016; revised 8 April, 2017; accepted 11 April, 2017. *For correspondence. E-mail nahlgren@clarku.edu; Tel. +1 (508) 793-7107; Fax +1 (508) 793-7174. Present addresses: †Department of Biology, Lasry Center for Bioscience, Clark University, 950 Main St, Worcester, MA 01610, USA; ‡Department of Earth Systems Sciences, Stanford University, Stanford, CA, USA

Marine Thaumarchaeota in particular represent 20–40% of prokaryotes in waters below the euphotic zone (Fuhrman and Ouverney, 1998; DeLong *et al*., 1999; Karner *et al*., 2001; Teira *et al*., 2006) and comprise important contributors to global C and N cycles (Ingalls *et al*., 2006; Yool *et al*., 2007; Santoro *et al*., 2010). They fix carbon via the modified 3-hydroxypropionate/4-hydroxybutyrate (3HP/4HB) pathway (Könneke *et al*., 2014), and they derive energy from the oxidation of ammonia (Könneke *et al*., 2005; Walker *et al*., 2010). They are estimated to contribute to a major portion of marine nitrification (Wuchter *et al*., 2006; Martens-Habbena *et al*., 2015), such that they contribute to regeneration of nitrate in deeper waters that then in turn is upwelled to fuel phytoplankton primary productivity. Thaumarchaeota may also significantly contribute to regenerated nitrite and nitrate within the euphotic zone (Yool *et al*., 2007). Thaumarchaeota are also highly diverse (Garcia-Martinez and Rodriguez-Valera, 2000; Francis *et al*., 2005; Biller *et al*., 2012), so obtaining additional isolates is valuable for a better assessment of genomic and physiological diversity in this group. Because of their importance in C and N cycles, it is critical to obtain and characterize multiple representative members of this phylum in culture.

It has been recognized that urea, an organic form of N, may play a role in marine nitrification by Thaumarchaeota. A few recently obtained cultured strains, *Candidatus. Nitrosopumilus piranensis* D3C and *Ca. Nitrosopumilus sp.* PS0 have been shown to possess urease operons and/or use urea (Qin *et al*., 2014; Bayer *et al*., 2016). Likewise several single-cell amplified genomes (SAGs) contain urease genes (Luo *et al*., 2014). Urea is an abundant form of organic N in the oceans and can sometimes be found at higher standing stocks than ammonium (Remsen, 1971; Harrison *et al*., 1985; Painter *et al*., 2008). Metabolism of urea by urease to CO_2 and ammonia can subsequently feed both carbon fixation and ammonia oxidation pathways in Thaumarchaeota. Utilization of urea therefore could represent an important source of N for nitrification. Indeed recent reports observe high prevalence of thaumarchaeal urease genes in the Antarctic, Arctic and northeast Pacific Oceans suggesting that perhaps urea utilization by these archaea may represent a major source of nitrification (Alonso-Sáez *et al*., 2012; Smith *et al*., 2016; Tolar *et al*., 2016). While the *ureC* gene in the urease operon may be prevalent among thaumarchaeal genomes, this was rarely expressed (RNA levels) in samples from the Arctic (Pedneault *et al*., 2014). Work off of California likewise failed to detect thaumarchaeal *ureC* transcription despite recovery of diverse *ureC* genes from environmental DNA samples (Smith *et al*., 2016). It is unclear how prevalent urease genes are in other temperate thaumarchaeal populations.

Another important open question is how infection and interactions with viruses impact Thaumarchaeota and their productivity. No viruses have yet been isolated that infect marine Thaumarchaeota or any mesophilic marine archaea for that matter, likely because archaea themselves are difficult to isolate and grow in pure culture. Virus-like particles have been observed in cultures of *Pyrococcus abyssi*, which is a hyperthermophilic, vent-associated euryarchaeon (Geslin *et al*., 2003), representing the only example of a marine archaeal virus. Mesophilic Thaumarchaeota also occur in soil habitats, and the genome of *Nitrososphaera viennensis* EN76 contains a probable provirus (Krupovic *et al*., 2011). Two culture-independent studies have recently identified possible viral sequences associated with marine thaumarchaeal genomes (Chow *et al*., 2015; Labonte *et al*., 2015), suggesting the existence of viruses that infect marine Thaumarchaeota. Another recent study identified a metagenomically assembled contig with probable viral genes and an *amoC* gene, required for ammonia oxidation, with high similarity to thaumarchaeal *amoC* genes (Roux *et al*., 2016).

Our lab recently has enriched in culture a new strain of marine Thaumarchaeota, for which we propose the name *Candidatus Nitrosomarinus catalina* SPOT01, representing a novel, ecologically relevant lineage of marine Thaumarchaeota that is abundant in subsurface Pacific waters off of California. While it shares $\sim$ 80% of its genes with other Thaumarchaeota isolates like *Nitrosopumilus maritimus* and *Ca. Nitrosopelagicus brevis*, it contains unique regions of ecological significance including genes for a newly recognized form of DNA modification, multiple putative host restriction systems, a region containing probable virus genes, and urea utilization genes. This study broadens our view of the ecology of these globally important archaea and presents strong evidence that marine archaea are infected by viruses.

Results

Growth characteristics of strain SPOT01

16S rDNA amplicon sequencing revealed that our enrichment culture of a new thaumarchaeon was primarily ($\geq$ 97%) comprised of a single dominant 16S rDNA genotype distinct from other marine Thaumarchaeota (Fig. 1), but the enrichment also contains two bacteria of the genera *Erythrobacter* and *Sphingomonas* (Supporting Information Fig. S1). In nutrient-amended seawater, the enrichment culture consumes ammonium and produces nitrite, consistent with it being dominated by an ammonia-oxidizing thaumarchaeon (Fig. 2A), and we have named this dominant strain in the enrichment SPOT01. Specific growth rates based on accumulation of nitrite were determined over several temperatures show that the SPOT01 culture is better adapted to growth at cooler temperatures that other strains (Fig. 2B). The lowest temperature tested at which the culture grew was 10°C. The temperature at

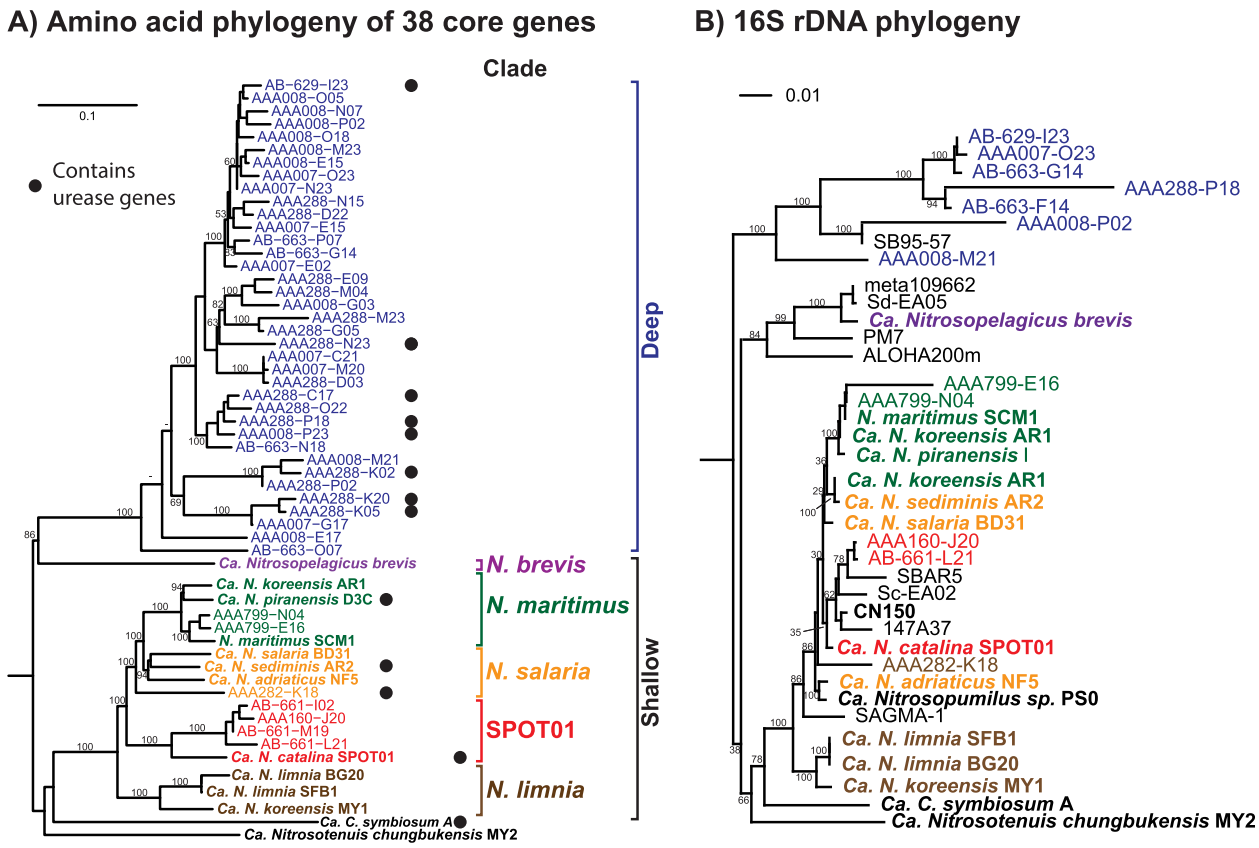

Fig. 1. Phylogeny of Thaumarchaeota isolates and genomes based on (A) concatenated sequences of 38 conserved core genes (protein tree) or (B) the 16S rRNA gene (nucleotide tree).
Numbers at the nodes indicate bootstrap values for 100 replicates. Taxa in bold represent sequences from enrichment or isolate cultures. Taxa names starting with 'AAA-' or 'AB-' are from SAGs. Non-bold taxa names in the 16S rDNA tree are from cloned sequences used previously in (Santoro and Casciotti, 2011). Taxa are coloured according to the clade to which they belong in the core gene tree and the same colours are used for corresponding taxa in the 16S rRNA tree. Genus names abbreviated as '*N.*' are *Nitrosoarchaeum* for taxa in the '*N. limnia*' clade, *Nitrosomarinus* for the SPOT01 strain, and *Nitrosopumilus* for all other cases. Genomes in **A** that contain homologues of the urease *ureA-C* genes are denoted with a black circle. [Color figure can be viewed at wileyonlinelibrary.com]

which it had its maximum growth rate, 23°C, was lower than all other strains tested except for *Ca. N. brevis* (22°C) (Qin *et al.*, 2014; Santoro *et al.*, 2015; Bayer *et al.*, 2016). Below 20°C the culture grew faster than all other strains but was the only strain that could not grow at any temperature at or above 30°C. Plotting the specific growth rates of each strain relative to its peak specific growth rate also shows how strain SPOT01's specific growth rates did not drop as rapidly as other strains at temperatures below 23°C (Supporting Information Fig. S2).

General genome characteristics

Illumina MiSeq (>1300X coverage) and Sanger sequencing were used to obtain the complete 1.36 Mb genome of strain SPOT01 (see Methods). Assembly of PacBio sequences (~770X coverage) produced a nearly identical genome (only ~100 mismatches). The strain SPOT01 genome is just ~10% larger than the minimal genome of *Ca. N. brevis* (1.23 Mb) (Table 1). Strain SPOT01 has the lowest GC content (31.4%) of all Thaumarchaeota isolates to date (Table 1) and its GC content is only slightly higher than the streamlined genomes of free-living *Prochlorococcus* (Rocap *et al.*, 2003) and *Pelagibacter* strains (Giovannoni *et al.*, 2005) (minimum GC's of 30.8% and 29.1% respectively). The strain SPOT01 genome encodes 1,677 predicted proteins and shares a high proportion of genes with other marine Thaumarchaeota genomes (Table 1, Supporting Information Table S1).

Phylogenomic analysis of 38 genes conserved (Supporting Information Table S2) across available thaumarchaeal genomes (Supporting Information Table S3) was used to better resolve the phylogeny of marine Thaumarchaeota than 16S rDNA analysis (Fig. 1). Thaumarchaeal genomes belong to two major groups that are consistent with previous multi-gene (Luo *et al.*, 2014) and single-gene (*amoA*

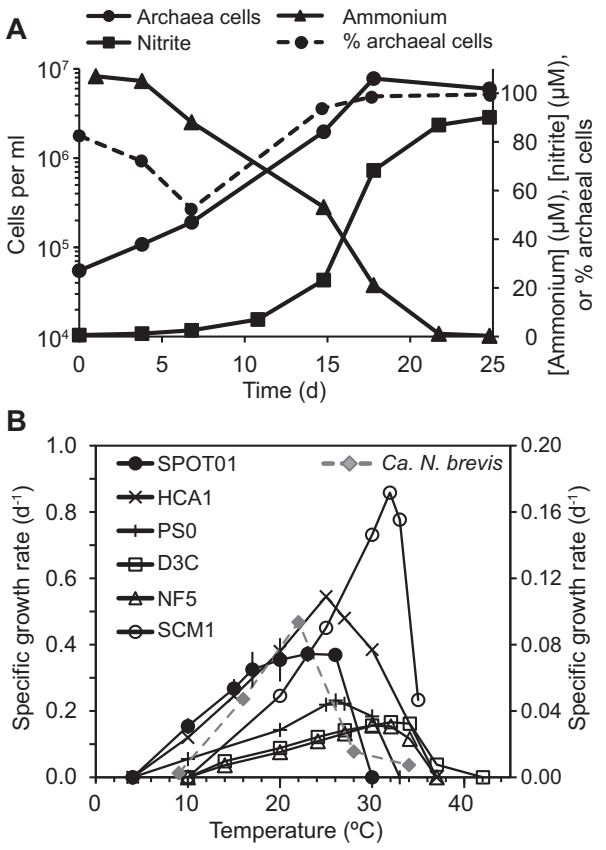

Fig. 2. A. Representative growth curve of the *Ca. Nitrosomarinus catalina* SPOT01 culture demonstrating consumption of ammonium and production of nitrite concomitant with growth of archaeal cells. B. Temperature versus specific growth rates for *Ca. N. catalina* SPOT01 (this study) and previously published data for five other marine Thaumarchaeota strains (Qin *et al*., 2014; Santoro *et al*., 2015; Bayer *et al*., 2016). Points for strain SPOT01 are averages of three to four transfers of triplicate, acclimated batch cultures maintained in exponential phase, and the error bars depict standard deviations. Strain names are provided in the legend as shorthand for the following organisms: 'SPOT01' for *Ca. N. catalina* SPOT01, 'HCA1' for *Ca. N. sp.* HCA1, 'PS0' for *Ca. N. sp.* PS0, 'D3C' for *Ca. N. piranensis* D3C, 'NF5' for *Ca. N. adriaticus* NF5, and 'SCM1' for *N. maritimus* SCM1. Note that *Ca. N. brevis* specific growth rates are plotted using the right vertical axis, and specific growth rates for all other strains are plotted using the left axis.

and *accA*) phylogenies (Francis *et al*., 2005; Hallam *et al*., 2006). The first group contains single-cell amplified genomes (SAGs) from mostly deep, mesopelagic waters ($>$ 770 m), corresponding to the *amoA* water column B (WCB) clade (Francis *et al*., 2005) (Supporting Information Fig. S3). We subsequently refer to this group as the 'Deep' group. The second group contains genomes primarily from shallower waters. We classified these 'shallow' group genomes into five clades, provisionally named by a representative cultured clade member (clade names: *N. brevis, N. limnia, N. maritimus*, *N. salaria*, SPOT01). The SPOT01 strain, representing the newly designated SPOT01 clade, is the first cultured representative of a distinct *amoA* (and urease gene *ureC*) clade only known previously by cloned sequences or SAGs (Supporting Information Figs S3 and S4). The phylogenomic clades in Fig. 1 are distinguished by an average nucleotide identity (ANI) threshold of approximately 85% (Supporting Information Table S4), suggesting they represent different genera or broader-level taxa rather than different species that typically are separated by ~95% ANI values (Konstantinidis and Tiedje, 2005). While the members of the '*N. salaria*' clade have ANI values $>$ 84.5% to each other, they branch deeply to each other in the core gene tree and as such each could represent ecologically distinct lineages.

Based on competitive recruitment of metagenomic reads to marine thaumarchaeal genomes (requiring $\geq$ 85% identity to include all clade members), clade SPOT01 can be an abundant member of natural communities. The SPOT01 clade was the dominant group at the deep chlorophyll maximum (45 m) in September 2012 in the temperate waters of the San Pedro Ocean Time-Series (SPOT) site off of California (Fig. 3A). At 150 m over a full seasonal cycle at SPOT, the Deep group was dominant except for two months, September and October, when instead clade SPOT01 was the most abundant clade (Fig. 3D). In contrast, the *N. brevis* and *N. salaria* clades were the most abundant thaumarchaeal members in the upper column at the Bermuda Atlantic Time Series (BATS) in August 2002 and the Hawaii Ocean Time-series (HOT) in March 2006 respectively (Fig. 3B and C). The SPOT01 clade still comprised 4–8% of thaumarchaeal populations at 25 and 75 m at HOT. Except for surface waters at BATS, the SPOT01 clade was consistently more abundant than the *N. maritimus* clade at all three sites. The *N. salaria* clade was most abundant in near surface waters (0–25 m) at all three sites where total Thaumarchaeota abundance was lower. Genomes from the Deep group were dominant at depths $>$ 200 m at all three sites, congruent with their label.

Similarity to previously sequenced genomes

The predicted biochemical capacity of strain SPOT01 overall was similar to previously described *N. maritimus* and *Ca. N. brevis* strains based on gene content. Strain SPOT01 contains the same complement of genes as *Ca. N. brevis* for ammonia oxidation (*amo* operon) and important carbon pathways including the modified 3-hydroxypropionate/4-hydroxybutyrate (3HP/4HB) pathway for carbon fixation; complete sets of genes for the oxidative tricarboxylic acid cycle, pentose phosphate pathway and gluconeogenesis; and an incomplete glycolysis pathway (Supporting Information Table S5). Likewise, strain SPOT01 has the same complement of predicted proteins

Table 1. Genomes of thaumarchaeal isolates or enrichment culture.

Organism	Source	Genome size (Mb)	GC (%)	No. of proteins	No. of tRNAs	Genes shared with strain SPOT01 (%) [a]	Urease genes	Relevant reference
Ca. N brevis	Water column, near-shore California	1.23	33.2	1445	42	67	–	(Santoro and Casciotti, 2011; Santoro *et al.*, 2015)
Ca. Nitrosomarinus catalina SPOT01	Water column, near-shore California	1.36	31.4	1677	40	100	+	This study
Ca. N. salaria BD31	Marine sediments (San Francisco Bay)	1.57	33.8	2089	41	71	–	(Mosier *et al.*, 2012)
Ca. N. koreensis AR1	Arctic marine sediment	1.64	34.2	1890	38	74	–	(Park *et al.*, 2014)
N. maritimus SCM1	Tropical aquarium gravel	1.65	34.2	1796	44	75	–	(Könneke *et al.*, 2005; Walker *et al.*, 2010)
Ca. N. sediminis AR2	Arctic marine sediment	1.69	33.6	1974	37	77	+	(Park *et al.*, 2014)
Ca.N. piranensis D3C	Surface water, Mediterranean Sea	1.71	33.8	2161	43	79	+	(Bayer *et al.*, 2016)
Ca. N. adriaticus NF5	Surface water, Mediterranean Sea	1.80	33.4	2184	43	78	–	(Bayer *et al.*, 2016)

[a]The fraction of protein encoding genes in strain SPOT01 that are shared with this genome.

for amino acid synthesis as *Ca. N. brevis* and *N. maritimus*, the latter of which can grow on minimal medium and thus synthesizes all essential amino acids (Könneke *et al.*, 2005; Santoro *et al.*, 2015). Strain SPOT01, like *Ca. N. brevis*, possesses the genes for synthesis of B vitamin cofactors thiamin (B_1), riboflavin (B_2), pantothenate (B_5), pyridoxine (B_6) and biotin (B_7) and a thaumarchaeal pathway for B_{12} synthesis (Doxey *et al.*, 2015; Heal *et al.*, 2017) (Supporting Information Table S5). Because of its very similar genomic composition to other thaumarchaeal isolates, we focussed on strain SPOT01 genes that are rare or absent in its close relatives.

Unique or unusual genomic characteristics

Phosphorothioation (PT) genes. Strain SPOT01 possesses genes responsible for phosphorothioation (PT), a newly recognized form of DNA modification that replaces a sulfur atom for an oxygen atom in phosphate groups of the DNA backbone (Wang *et al.*, 2007; Wang *et al.*, 2011). Genes NMSP_1264 to NMSP_1268 in strain SPOT01 had 24–47% amino acid identity to PT modification genes *dndA-E* in *Streptomyces lividans* 1326 (Zhou *et al.*, 2005), and they have been annotated as *dndABCDE*. Strain SPOT01 *dnd* genes are organized the same as in *S. lividans* 1326, except that *dndBCDE* are in the opposite orientation respective to *dndA* (Supporting Information Fig. S5). Strain SPOT01 DNA was verified to contain PT modifications by treatment with iodine, which chemically breaks PT bonds (Cao *et al.*, 2015). Iodine treatment fragmented strain SPOT01 DNA compared with the 'no iodine' control, consistent with positive control DNA from *E. coli* expressing *S. enterica* PT modification genes (Fig. 4).

Epigenetic modifications on the strain SPOT01 genome were identified with PacBio sequencing (Flusberg *et al.*, 2010; Cao *et al.*, 2014). Two canonical methylation motifs were detected: m6A methylation of GATC and m4C methylation of AGCT. Greater than 99% of GATC and AGCT sites were methylated and >98% of these sites were methylated on both strands (Table 2). Strain SPOT01 contains two putative methylase genes, NMSP_0260 and NMSP_0378, that are homologues of methylases identified in *N. maritimus*, Nmar_1499 and Nmar_1319 respectively, the latter of which is predicted to methylate at GATC sites (Roberts *et al.*, 2015). PacBio analysis revealed another modified motif, TGCA, for which the G nucleotide is modified. This modification motif was inferred to be the site of PT modification (see Discussion). This modification was detected at 19% of the nearly 20,000 possible TGCA sites in strain SPOT01, and 83% of the modified sites showed modification on both strands.

Genes NMSP_1261-1263 adjacent to the strain SPOT01 *dnd* operon exhibited homology to histidine-asparagine-histidine (HNH) type nuclease or restriction

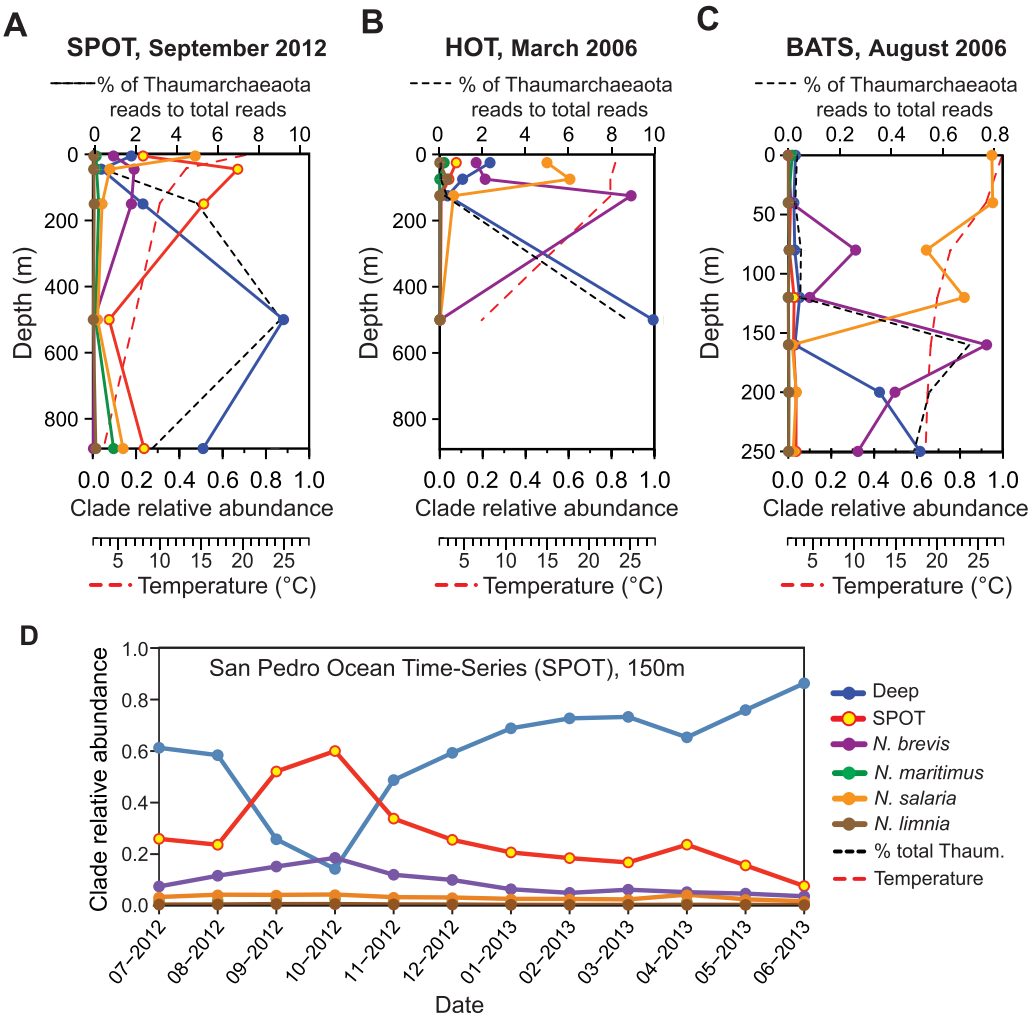

Fig. 3. Competitive recruitment analysis of metagenomic reads to thaumarchaeal genomes.
Metagenomic reads were searched against a database of available thaumarchaeal genomes and only the top hit with ≥85% identity over ≥100 bp were retained. Read counts to belonging to each clade were normalized to total genome length to determine the relative abundance of thaumarchaeal clades (Fig. 1A).
A. Depth profile at the San Pedro Time-Series (SPOT) site off of Los Angeles, CA in September 2012.
B. Depth profile at the Hawaii Ocean Time-series (HOT) in March 2006.
C. Depth profile at the Bermuda Atlantic Time Series (BATS) in August 2002.
D. Abundance of thaumarchaeal clades at 150 m at SPOT over one year.
Relative recruitment to *C. symbiosum* A and *Ca. N. chungbukensis* MY2 were negligible (< 0.01%) in all samples. The fraction of reads matching any Thaumarchaeota genome to the total reads in each sample depicts the estimated relative abundance of Thaumarchaeota in the total microbial community (dashed line). [Color figure can be viewed at wileyonlinelibrary.com]

enzyme (RE) domains (Table 3), suggesting it encodes a putative restriction enzyme. Furthermore NMSP_1262 shows similarity (E-value 10^{-4}, 21% amino acid identity) to an annotated restriction endonuclease in *Salmonella enterica*. NMSP_1262 is also homologous to the TIGR04095 gene family of predicted restriction enzymes linked to PT-modification genes by phylogenetic profiling of microbial genomes (equivalogs) (Haft *et al.*, 2001). The organization of PT modification and putative RE genes parallels that of adjacent PT modification (*dptB-E*) and PT-dependent RE (*dptF-H*) genes in *S. enterica* which forms a host-specific, restriction system (Xu *et al.*, 2010) (Supporting Information Fig. S5).

Putative viral genes. Strain SPOT01 interestingly contains a unique region for which three independent virus prediction programs, PHAST, VirSorter and phiSpy, reported overlapping regions as likely to be viral (Fig. 5). The program PHAST predicted two regions of viral genes based on localized clusters of genes homologous to known viral genes. The former region overlaps with a 'possible' (category level III), 80 gene viral region identified by VirSorter

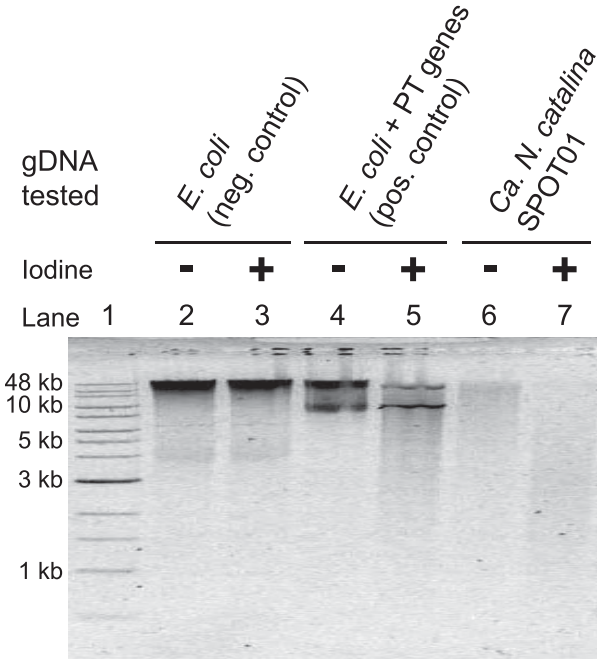

Fig. 4. Agarose gel electrophoresis image of genomic DNA treated with or without iodine.
Treatment of iodine with strain SPOT01 gDNA (lane 7) cleaves the DNA at PT-modified sites resulting in a smaller fragment size distribution than without iodine treatment, consistent with positive control DNA from *E. coli* expressing PT genes from *S. enterica*. *E. coli* without the PT genes (*E. coli* strain DH5 with empty cloning plasmid pCR™2.1-TOPO® was used as a negative control). Ladder fragment sizes (kb) are listed to the left of the gel. The strain SPOT01 DNA tested came from a culture for which strain SPOT01 was 99% of the total cells based on 16S rDNA sequencing (Supporting Information Fig. S1), indicating that iodine degradation was due to PT-modifications of strain SPOT01 but not bacteria in the enrichment.

(Fig. 5A) based on significant depletion of genes with homology to the PFAM (Protein family) database and enrichment of uncharacterized genes. The authors of VirSorter recommend that such 'possible' predictions be carefully inspected and confirmed, which we have done below. The default phiSpy results predicted a region that was centred on an operon of 17 ribosomal proteins (Supporting Information Fig. S6). Given the possibility that this region is a false positive result, we considered an alternative lower threshold for virus prediction whereby the default predicted region was excluded (see Supporting Information Fig. S6). This alternative threshold produced five possible viral regions, one of which overlapped with the VirSorter predicted region (Fig. 5A).

Subsequent sequence similarity analyses using viral sequence databases provided additional evidence that the VirSorter predicted region contains several viral genes. We first noted that NMSP_1215 had its best hit among NCBI nr database sequences to the SegD protein encoded in the genome of the *Vibrio*-infecting T4-like KVP40 phage (39% identity) (Table 4). The SegD protein belongs to a family of phage-specific homing endonucleases in canonical T4 and T-even phages (Miller *et al.*, 2003). Two of the four other significant (defined throughout as E-value $\leq$1e-5) blastp results for NMSP_1215 were to viral sequences: *Caulobacter* phage Cr30 and *Paramecium bursaria* Chlorella virus FR483, the former of which is another T4-like virus.

Although we found one gene with homology to a known virus isolate gene, in general, one would expect very few significant results to known viruses given the paucity of sequenced archaeal virus genomes, and indeed the VirSorter region was enriched in genes with no significant match to nr (Fig. 5A), a signal characteristic of viral regions but also for hypervariable regions more generally. To further assess the probable taxonomy of genes in this region, each was searched against databases of viral isolates and vetted marine viral metagenomes: the Broad Marine Viral database (containing metagenomes and isolates genomes not in nr) and viral protein clusters found in the Pacific Ocean Virome metagenomes. We assessed if each gene was likely viral or cellular based on whether each gene had a higher normalized bit score (bit score divided by the alignment length) to cellular genes in the nr database or genes in viral sequence databases. Most genes with a significant match to any database had a higher normalized bit score to a protein in nr (93%), all of which were cellular sequences except for the SegD hit for NMSP_1215 described above. Nearly all (98%) of those cellular proteins belonged to other marine Thaumarchaeota sequences. Twenty-two genes in the strain SPOT01 genome had stronger similarity to a sequence in one of the viral

Table 2. DNA modifications detected in *Ca. Nitrosomarinus catalina* SPOT01 by PacBio sequencing.

Modification	Motif	No. of motifs in strain SPOT01	Fraction of modified motifs	Of modified motifs, fraction that are modified on both strands	Average incidence of modification[b]
m6A	GATC	7392	99.7%	99.9%	1 in every 185 bp
m4C	AGCT	5720	99.0%	98.4%	1 in every 240 bp
PT (G)[a]	TGCA	19956	19.4%	83.3%	1 in every 351 bp

[a]The guanine base is modified and is inferred to be the PT-modified motif.
[b]Calculated as the number of modified sites divided by the total genome length.

Table 3. Top CD-Search results of putative restriction enzyme operon to conserved domains databases (PFAM, COG, cd and TIGR).

Locus tag	Domain	Domain description	E-value
NMSP_1261	pfam13395	HNH endonuclease	5×10^{-04}
	cd00085	HNH nucleases	2×10^{-03}
NMSP_1262	TIGR04095	DNA phosphorothioation system RE	2×10^{-102}
	COG1061	Superfamily II DNA or RNA helicase	2×10^{-40}
	pfam04851	Type III RE, res subunit;	2×10^{-13}
NMSP_1263	COG3183	Predicted RE, HNH family	5×10^{-04}
	pfam13391	HNH endonuclease	2×10^{-03}

databases (Table 4), and six of these genes were notably concentrated in the VirSorter predicted viral region (Fig. 5, Table 4). A moving window analysis shows that the incidence of genes with stronger similarity to viral database sequences, is highest in the front half of the VirSorter predicted region (Fig. 5).

We also noted that while many genes in the VirSorter region were most similar to cellular proteins, their best hit to the viral databases were often nearly as similar as the best nr blastp hit. This was visualized by a running average of the gene-by-gene ratios of normalized bit scores for the best nr result to the best viral database result. These ratios were highest in the viral region (Fig. 5). These multiple observations—overlapping regions predicted by independent virus identification programmes, significant homology of a strain SPOT01 gene to a viral isolate gene, and enrichment of genes with significant similarity to vetted viral metagenomes—support that the predicted region in strain SPOT01 contains probable viral genes.

Additional analysis for viral elements. Clustered regularly-interspaced short palindromic repeats (CRISPRs) and adjacent Cas proteins comprise a defence system against viral infection (Bhaya *et al.*, 2011). No definitive CRISPRs were found by the CRISPRFinder program. We did not find any genes homologous to Cas proteins, even though Thaumarchaeota strains *Ca. N. sediminis* AR2 and *Ca. N.*

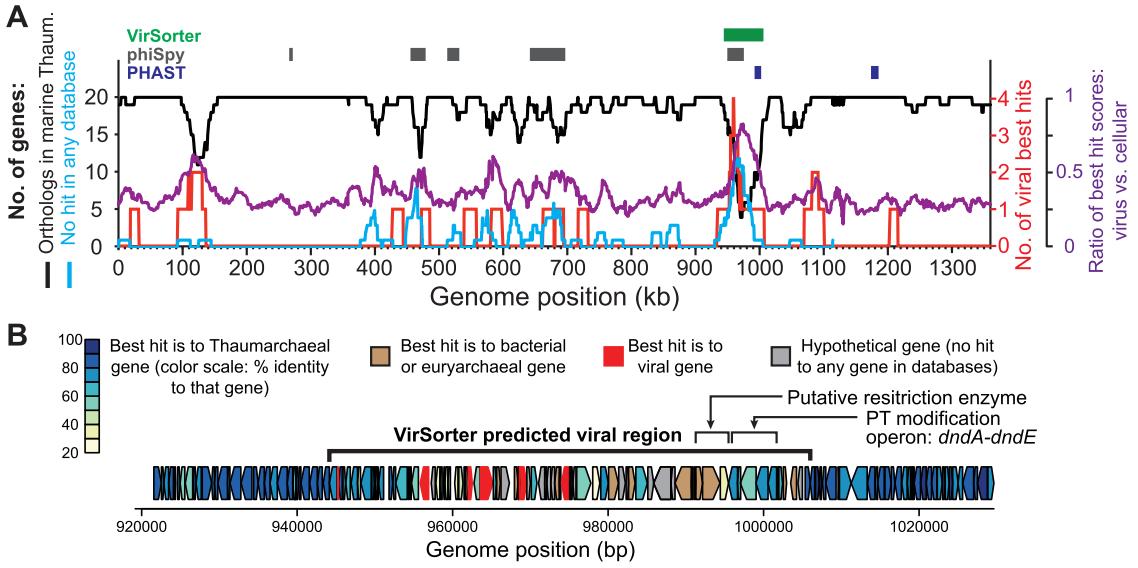

Fig. 5. Evidence for the presence of viral genes in strain SPOT01.
A. Regions predicted as viral by three programmes are shown at the top: VirSorter (green), PHAST (blue) and phiSpy (grey) using an alternative prediction threshold (see Supporting Information Fig. S6). Running averages over a 20 gene window are depicted along the length of the genome for four analyses: the number of genes with a more similar (higher normalized bit score) to viral (virus sequences in nr and viral metagenomes) than cellular sequences in nr (red); the number of genes with orthologues in at least one other complete thaumarchaeal genome (black); the number of genes with no significant match to viral or cellular sequence databases (blue); and the ratio of normalized bit scores for the best viral match to the best cellular match (purple).
B. Genes in and flanking the VirSorter predicted viral region. Genes are coloured according to the type of gene to which each has a best blastp hit: Gene in another thaumarchael genome (blue-yellow colour scale depicts identity to that thaumarchaeal gene), viral gene (red), cellular genes (bacterial or euryarchaea, brown), or no hit (hypothetical gene, grey). The location of PT modification and restriction enzyme genes are also indicated. [Color figure can be viewed at wileyonlinelibrary.com]

Table 4. Genes in SPOT01 for which the best blast hit was more similar (higher normalized bit score) to viral sequences (marine viral sequences or viral sequence in nr) than cellular sequences in nr.

Locus tag	Gene start position	Best viral database hit[a]	Annotated function
NMSP_0040	29687	POV	GTP cyclohydrolase I
NMSP_0127	107686	POV	Bifunctional 3-demethylubiquinone-9 3-methyltransferase
NMSP_0144	126794	POV	Bifunctional UDP-glucuronic acid decarboxylase
NMSP_0149	131829	POV	Molybdenum cofactor biosynthesis protein A
NMSP_0525	440234	POV	Plastocyanin
NMSP_0581	480927	POV	Uracil DNA glycosylase superfamily protein
NMSP_0673	554531	POV	Membrane ATPase/protein kinase
NMSP_0733	594810	POV	tRNA (mo5U34)-methyltransferase
NMSP_0850	677835	BMVD: Hydrothermal Vent/Seep virome (CAM_SMPL_A0003)	Hypothetical protein
NMSP_0873	689900	BMVD: Uncultured virus Saanichss Achan-JL3 (CAM_SMPL_001004)	Hypothetical protein
NMSP_0919	727537	POV	Hypothetical protein
NMSP_1196	944144	BMVD: Marine sediment virome VAGALB1/1 (CAM_SMPL_000842)	Hypothetical protein
NMSP_1215*	954718	Viral protein in nr	NUMOD3 motif protein
NMSP_1217*	957247	POV	Hypothetical protein
NMSP_1226*	961678	BMVD: Hydrothermal Vent/Seep virome (CAM_SMPL_A0003)	Hypothetical protein
NMSP_1228*	962751	POV	Hypothetical protein
NMSP_1234*	968411	POV	Chaperone protein DnaJ
NMSP_1253*	982316	POV	Hypothetical protein
NMSP_1272	1002532	POV	Pentapeptide repeats
NMSP_1371	1085468	POV	Hypothetical protein
NMSP_1381	1095550	POV	30S ribosomal protein S4
NMSP_1531	1211588	POV	Putative aspartate aminotransferase 2

*gene is located within VirSorter predicted viral region.
a. Source sample of the best hit is listed, POV = Pacific Ocean Virome, BMVD = Broad Marine Viral database.

koreensis AR1 contain Cas1 domain proteins (Luo *et al*., 2016). Neither did we find any recognizable integration sites or viral integrases.

Urea utilization. It has been recently recognized that some but not all marine Thaumarchaeota contain transporters and urease genes for the utilization of urea (Tully *et al*., 2012; Luo *et al*., 2014; Qin *et al*., 2014; Bayer *et al*., 2016) (Fig. 1A), an abundant form of organic N in the oceans. Strain SPOT01 contains the complete urease operon for urea utilization (NMSP_0016-0018, *ureA-C*), and these predicted proteins have 86%, 73% and 86% identity to the UreA-C proteins in *Ca. N piranensis* D3C, a strain demonstrated to utilize urea. Strain SPOT01 also contains the accessory urease genes *ureD-G* (NMSP_0012-0015). Strain SPOT01 also has a probable urea transporter (NMSP_0019) adjacent to the urease operon that has 39% identity to the characterized DUR3 urea transporter in yeast (Navarathna *et al*., 2011) and high identity to the two annotated urease transporters in *Ca. Nitrosopumilus piranensis* (56% and 84%). Curiously, strain SPOT01 did not grow appreciably on urea as a sole N source when tested at concentrations of 4, 20 and 100 μM (data not shown), nor did it show enhanced growth when provided urea and ammonium in comparison to media with ammonium only (data not shown).

Natural prevalence of the viral region and PT and urease operons

Metagenomes from 150 m at the SPOT study site were used to estimate the prevalence of PT, viral region, and urease genes in natural communities by calculating the

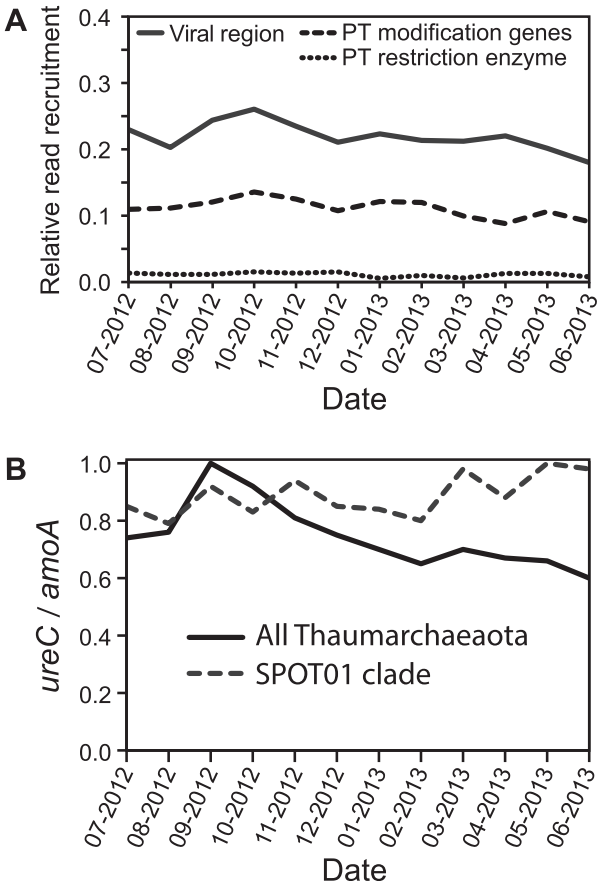

Fig. 6. Estimated relative abundance of viral region and PT genes (A) and the urease operon (B) in natural Thaumarchaeota at 150 m at the San Pedro Ocean Time-series (SPOT) site as determined by metagenomic read recruitment. Relative recruitment for PT and viral regions was calculated as the ratio of length-normalized read coverage for the region of interest versus that of the whole strain SPOT01 genome for reads that were ≥85% identical for ≥ 100 bp. The relative ratio of *ureC* to *amoA* genes was calculated as the length-normalized coverage for reads with ≥85% identity for ≥ 100 bp to marine thaumarchaeal *ureC* and *amoA* genes. The estimated ratio of *ureC* to *amoA* genes for clade SPOT01 was similarity calculated but only for reads that whose top hit was a SPOT01 clade *ureC* and *amoA* with ≥85% and ≥95% identity respectively, which reflect appropriate thresholds of between clade sequence divergence.

ratio of reads that mapped with ≥85% identity over- ≥ 100 bp to the region of interest divided by reads that mapped to the whole strain SPOT01 genome with the same criterion. The VirSorter predicted region and PT modification genes had recruitment ratios of 0.18–0.26 and 0.08–0.14 respectively, indicating they occur at modest levels (∼22 and ∼10% respectively) in natural SPOT01 populations (Fig. 6A). The putative PT RE genes had lower ratios (0.005 to 0.015) (Fig. 6A), indicating the large majority (∼90%) of individuals with the PT modification genes lacked the putative RE region. Similar recruitment analysis of thaumarchaeal *ureC* versus *amoA* genes indicated a large portion (∼60–100%) of natural Thaumarchaeota cells contain urease genes (Fig. 6B). Recruitment just using clade SPOT01 *ureC* and *amoA* sequences also indicated high prevalence (∼80–100%) of the urease operon in members of this clade (Fig. 6B).

Discussion

The new SPOT01 strain represents a previously uncultured clade of marine Thaumarchaeota that based on genomic analyses (ANI values) likely represents a new genus or higher level taxon. This culture is distinguished by being less tolerant of warmer temperatures and at temperatures < 20°C, growing faster than previous strains tested. It is worth noting that there was a noticeable shift in the bacterial genotypes found in the enrichment at 23 and 26°C compared with lower temperatures (Supporting Information Fig. S1). It is possible that these bacteria could influence the physiology of strain SPOT01, but we surmise the bacteria act as commensals that do not significantly affect the upper and lower limits of temperature at which strain SPOT01 can grow.

Given its divergence from other Thaumarchaeota genomes, we propose this strain be classified as a new genus with the name *Candidatus Nitrosomarinus catalina* SPOT01. The name *Nitrosomarinus* describes this organism's ability to oxidize ammonia to nitrite (nitrosus is the Latin masculine adjective for nitrous) and that it lives in the ocean (*marinus* is the Latin masculine adjective for marine). The species name *catalina* is in reference to Santa Catalina Island that is near the San Pedro Ocean Time-series site where the SPOT01 clade was found to be abundant.

Metagenomic recruitment analysis demonstrates strain SPOT01 is ecologically relevant since the clade to which it belongs can comprise a large portion of Thaumarchaeota in temperate waters (Fig. 3). In comparison to samples at other sites, the *N. brevis* clade instead was the more abundant clade below the surface mixed layer in the warmer waters at HOT and BATS. This difference in dominance between the SPOT01 and *N. brevis* clades is broadly consistent with laboratory assessment of temperature adaptation of strains from each clade (Fig. 2B, Supporting Information Fig. S2). While the specific growth rates for *Ca. N. brevis* are unfortunately not well constrained with measurements at fewer temperatures (Fig. 2), strain SPOT01 again seems to be less tolerant of warmer temperatures than *Ca. N. brevis*. It is worth noting that optimum growth temperatures determined in the lab are generally a few degrees above typical temperatures organisms experience in the field (Fuhrman and Azam, 1983). Closely related strains within clades can have diverse temperature responses (Lehtovirta-Morley *et al.*, 2014; Pittera

et al., 2014; Qin *et al*., 2014; Bayer *et al*., 2016). Therefore, it is difficult to make concrete extrapolations from single strains, but our physiology and metagenomic results are initially congruent in suggesting that water temperature is directly or indirectly (through correlated factors) an important driver of Thaumarchaeota clade niche partitioning. Further work is needed to more fully characterize how the biogeographic distribution of clades may correlate to temperature.

The strain SPOT01 genome interestingly contains *dnd* genes for phosphorothioation DNA modification, and a possible associated restriction system for defence from foreign DNA. *dnd* genes occur in many lineages of archaea and bacteria (Zhou *et al*., 2005; Yao *et al*., 2009; Wang *et al*., 2011) but strain SPOT01 represents the first report of a marine archaeon possessing PT modification. We subsequently found that the genome of *Ca. N. salaria* BD31 (Mosier *et al*., 2012) and two marine thaumarchaeal fosmids, KM3_186_C08 and KM_47_F06 (Deschamps *et al*., 2014) also contain *dnd* homologues (Supporting Information Fig. S5), not previously recognized by these studies. We infer from PacBio sequencing that the guanine in the motif, TGAC, is the site of PT modification since methylation only occurs at adenine and cytosine bases. If this motif is confirmed by other experimental means to be PT modified, it would be a unique PT motif and add to the known diversity of these systems (Xu *et al*., 2009; Wang *et al*., 2011).

We suggest that the probable RE operon next to the *dnd* genes in strain SPOT01 functions as a restriction modification (RM) system to protect the cell from infection by non-PT modified viruses or other foreign DNA. The adjacent location of the putative RE and *dnd* operons is homologous to the same organization of PT modification and RE genes in *S. enterica*, that function together as a PT-specific host restriction system (Xu *et al*., 2010). In *S. enterica*, foreign DNA lacking PT modifications is degraded by the PT-specific RE encoded next to the *dndB-E* genes. Only a subset of bacteria carrying *dnd* genes also contain adjacent RE operons (Xu *et al*., 2010), and strain SPOT01 is the first report of any archaeon or marine prokaryote to potentially possess both operons and this new class of RM system. We subsequently noted that the thaumarchaeal fosmid KM3_186_C08 (Deschamps *et al*., 2014) and *Ca. N. salaria* BD31 also appears to possess putative RE genes adjacent to *dnd* genes (Supporting Information Fig. S5). PT modification may additionally or alternatively function in epigenetic control of gene expression (Low *et al*., 2001; Marinus and Casadesus, 2009), although the relatively high degree of PT-modified sites that show modification on both strands is consistent with their use in a RM system. Further tests are needed to test our hypothesis that these genes in strain SPOT01 function as an RM system.

About 20% of possible TGAC motifs in strain SPOT01 were PT modified, similar to the frequency observed for *E. coli* B7A (12%) that possesses a PT-specific RM system (Cao *et al*., 2015). The overall frequency of PT-modified sites in strain SPOT01 of $\sim$ 1 in 350 bp (Table 2) is within the reported range for other PT systems (1 in 322 to 3500 bp) (Wilson and Murray, 1991; Wang *et al*., 2011). The apparent low modification frequency for PT systems however differs from methylation RM systems that typically exhibit modification of nearly all possible sites (but fewer overall instances of those motifs in the genome). It is thought that PT and associated PT REs may require recognition of sequence features beyond the core four base motif for modification and cleavage activity (Cao *et al*., 2014; Gan *et al*., 2014).

Strain SPOT01 likely also possesses at least one, maybe two, methylation-based RM systems. PacBio sequencing revealed nearly complete modification of two motifs in strain SPOT01. This is only the third study to report genome-wide methylation patterns for archaea (Ouellette *et al*., 2015; Blow *et al*., 2016). The m6A GATC sites are likely methylated by the product of NMSP_0260, an orthologue of a predicted GATC-specific methyltransferase in *N. maritimus* (Nmar_1499) (Roberts *et al*., 2015). Strain SPOT01 and *N. maritimus* also possess another pair of orthologous, putative methylases (NMSP_0378 and Nmar_1319). Both do not possess obvious cognate restriction enzymes, but it is often difficult to identify restriction enzymes because they are so diverse (Roberts *et al*., 2015, R. Roberts pers. comm.). The adjacent *N. maritimus* gene Nmar_1320 and strain SPOT01 gene NMSP_0379 possess the catalytic motif PD.D/EXK or slightly modified motif PN.D/EXK respectively present in many restriction endonucleases (Pingoud and Jeltsch, 2001). While this requires further confirmation, NMSP_0378 and NMSP_0379 may represent an additional methylation-specific restriction modification system in strain SPOT01. The nearly complete methylation of the two motifs and modification on both strands is indirect evidence of their role in a restriction modification system, but this does not rule out that methylation may be used for purposes other than defence (Vasu and Nagaraja, 2013).

Another exciting discovery is the presence of probable viral genes in the genome of strain SPOT01, supported by converging results of multiple virus prediction tools and follow-up analysis of the 80 kb region predicted as viral by VirSorter, in particular. A potential caution in using such similarity searches against viral metagenomes is that these metagenomes could be potentially contaminated with cellular DNA. 'Viral fraction' metagenomes are often constructed using DNA extracted from particles that pass through a 0.2 μm filter to exclude most cellular organisms, but this method can sometimes collect small cells or free prokaryotic DNA. This may be problematic in particular for

Thaumarchaeota that can be smaller than 0.2 μm (Santoro *et al*., 2015). The sliding window analyses that we have conducted controls for potential contamination by examining regions enriched for genes with stronger similarity to viral genes above a 'background' of cellular sequence reads (Fig. 5). Furthermore, five proteins in the viral region with best hits to viromes (NMSP_1217, NMSP_1226, NMSP_1228, NMSP_1234 and NMSP _1253; Table 4) were from metagenomes that were processed to specifically capture viral particles and produce metagenomes with little cellular DNA contamination (Anderson *et al*., 2011; Hurwitz and Sullivan, 2013).

It is inconclusive if this region represents a viable lysogenic provirus or perhaps the remnants of lytic virus that recombined with the host genome. Exposure to UV or mitomycin C can often induce lysogenic viruses to enter the lytic cycle, but initial attempts with the latter method did not produce an obvious induction of viruses (data not shown). The ends of the VirSorter predicted region are closely flanked by tRNAs, a common feature of integrated bacteriophages, but the region lacks a recognizable, canonical integrase. We surmise that the larger VirSorter region is more broadly a hypervariable region that contains probable viral genes, notably concentrated at the front portion of the region (Fig. 5), that are the remnants of a defunct provirus or were laterally transferred into the genome during infection(s). Similarly, it is unclear to what group of viruses these genes may belong, but the presence of the gene with similarity to a T4-like endonuclease suggests a possible connection to myoviruses. Most archaeal virus isolates belong to groups distinct from bacterial (Fusello, Lipothrixviridae and Rudiviridae) (Prangishvili *et al*., 2006), so it is perhaps unusual that one of the recognizable viral genes in strain SPOT01 shows similarity to genes from classic myoviridae bacteriophage. However, a handful of viruses resembling bacteriophage myoviruses and siphoviruses are known to infect euryarchaeal hosts (Pfister *et al*., 1998; Prangishvili *et al*., 2006; Pagaling *et al*., 2007), and a nearly complete T4-like myovirus genome was recovered from a thaumarchaeal SAG (Labonte *et al*., 2015).

Regardless of their source, the presence of viral genes in the strain SPOT01 genome adds to the growing evidence that viruses infect mesophilic marine archaea. Remarkably little is known about viruses infecting marine Thaumarchaeota or marine mesophilic archaea in general. The putative provirus in the soil thaumarchaeon *N. viennensis* EN76 was recently shown to have homology to a marine, viral fraction (< 0.22 um) fosmid, suggestive of a related marine Thaumarchaeota virus (Chow *et al*., 2015), and the same study found viral fraction marine metagenomic sequences recruited to the T4-like virus recovered from a thaumarchaeal SAG (Labonte *et al*., 2015). While these studies point to the presence of Thaumarchaeota viruses, they are not definitive. As noted above, metagenomes from 'viral' size fractions may contain cellular material, especially from cells as small as Thaumarchaeota. SAGs can sometimes recover artifactual, non-specific viral sequences—for example the LaBonte *et al*. (2015) study recovered an algal virus from a bacterial SAG, probably due to non-specific attachment of virus particles to the host cell. More convincing evidence is the recent discovery of a contig assembled from virus fraction metagenomes (< 0.22 μm) that harbours both probable viral structural genes and a copy of the *amoC* ammonia oxidation gene with high similarity to cellular thaumarchaeal *amoC* genes (Roux *et al*., 2016). Our study adds evidence to the existence of marine thaumarchaeal viruses by demonstrating probable viral genes in a Thaumarchaeota isolate genome.

Metagenomic recruitment demonstrated that both the probable viral region and PT modification genes are maintained in sizeable portions of natural Thaumarchaeota populations (Fig. 6). Given that Thaumarchaeota exhibit streamlined genomes, similar to other highly abundant, marine free-living bacteria like *Prochlorococcus* and the SAR11 clade, we suggest that the moderate prevalence of these genes in natural populations is ecologically important, or otherwise would they would be purged. Interestingly, the presence of PT RE genes is rare in natural populations of close relatives of strain SPOT01 at the SPOT site but appear to be prevalent enough in marine Thaumarchaeota in general since they have been recovered in three other cases (*N. salaria* BD31 and two fosmids). Still, the presence of two or three different restriction modification systems in very small genome suggests strain SPOT01 places high priority on keeping out foreign DNA, presumably by viral infection and other forms of horizontal gene transfer, assuming this is the function of the modification system. The need for such defences when it lives in ocean midwaters at abundances typically near 10^3 cells per ml, and thus has extremely low encounter frequencies with other cells or viruses, is unexpected. Perhaps the methylation and phosphorothioation systems have other important functions in these organisms.

Strain SPOT01 also contains the genes for utilization of urea. There is no coherent phylogenetic pattern for which clades possess urease genes (Fig. 1). Strain SPOT01 strangely could not grow on urea as a sole N source when tested at concentrations from 4 to 100 μM, despite possessing urease genes (*ureA-C*), urease accessory genes (*ureD-G*) and a probable urea transporter that all show strong identity to proteins in *Ca. N. piranensis* D3C that grows on urea. The reason for this discrepancy is unclear, but perhaps it recently acquired mutation(s) that rendered this operon non-functional. Urea frequently occurs in the upper ocean at concentrations of 100s of nM (Painter *et al*., 2008), and thus potentially provides a significant pool of reduced N for growth of Thaumarchaeota. Utilization of urea by Thaumarchaeota was inferred to be

important in arctic communities (Alonso-Sáez *et al*., 2012), and thaumarchaeal urease genes were frequently found in the Antarctic (Tolar *et al*., 2016), Arctic (Pedneault *et al*., 2014) and the northeast Pacific (Smith *et al*., 2016). Congruent with these studies, a significant portion of cells of Thaumarchaeota and specifically those closely related to strain SPOT01 appear to contain the urease operon at the SPOT site (Fig. 6), suggesting that urea utilization may be important in lower latitude waters as well. The work of Smith *et al*. and Pedneault *et al*., however, found that expression of urease as assessed by *ureC* transcription was rarely detected or not detected at all. It therefore will be important in on-going work to directly measure expression of urease and rates of urea utilization and its subsequent contribution to ammonia oxidation rates at sites like SPOT and polar regions.

Overall, the novel thaumarchaeal strain with the proposed name *Ca. Nitrosomarinus catalina* SPOT01 and its genome provides valuable insight into this novel and abundant lineage (clade SPOT01) of marine Thaumarchaeota. It will also provide a valuable resource for continued work on organic N utilization of Thaumarchaeota and interactions with viruses. We also lay out a phylogenomic framework for classifying thaumarchaeal lineages, and metagenomic recruitment approaches to further biogeographic studies of these clades and specific functional genes (e.g. urease genes). This study more importantly adds to the growing evidence of interactions between viruses and mesophilic marine archaea. We provide one of the first epigenetic analyses archaea and highlight their probable importance in archaea-virus interactions. This work also raises important questions for future studies concerning why streamlined marine thaumarcaheal genomes apparently maintain multiple defence systems even though they are predicted to rarely encounter viruses or other cells that could introduce DNA into their cells.

Materials and methods

Enrichment and growth of the SPOT01 strain

A sample of the Thaumarchaeota enrichment culture CN75 provided by Alyson Santoro (Santoro and Casciotti, 2011) was maintained in our lab using 0.2 μm filtered seawater from the San Pedro Ocean Time-Series (SPOT) site off of California amended with nutrients as described in (Santoro and Casciotti, 2011). After a few months of transfers, we found the culture was dominated by cells with a 16S rDNA sequence different from that of CN75 and other cultured thaumarchaeal isolates (Fig. 1). We suspect that this new strain came from cells from the < 0.2 μm filtrate of SPOT seawater, since Thaumarchaeota cells can be < 0.2 μm in diameter (Santoro *et al*., 2015), and they supplanted the CN75 strain. The strain SPOT01 enrichment has subsequently been maintained on media using seawater passed twice through a 0.02 μm filter to avoid further similar changes.

For growth experiments, ammonium and nitrite concentrations were measured in duplicate using standard fluorescence (Holmes *et al*., 1999) or colorimetric methods (Strickland and Parsons, 1968) respectively. Total prokaryote cell abundances were determined by flow cytometry: unpreserved culture was stained with 1X SYBR Green I for 10 min and then run on a Becton Dickinson Accuri C6 machine. For 16S rRNA analysis 1 to 10 ml of culture were collected on a 0.1 μm polycarbonate Poretics filter and stored at −80°C in a cryovial for later extraction using a standard sodium dodecyl sulfate lysis and phenol-chloroform purification protocol (Fuhrman *et al*., 1998). Specific growth rates were determined at several temperatures by measuring the accumulation of nitrite. Triplicate cultures were acclimated to each temperature and maintained in logarithmic growth for at least three culture transfers after which stable specific growth rates were measured for three or four batch culture transfers.

For control DNA used for phosphorothioation assays, we grew up *Escherichia coli* BH10b with plasmid pJTU1238 expressing the phosphorothioation operon *dptB-E* from *Salmonella enterica* serovar Cerro 87 (Xu *et al*., 2010). This strain was provided by M. DeMott and grown in LB broth with 100 μg/ml ampicillin.

Genomic DNA from cultures was extracted using a standard sodium dodecyl sulfate lysis and phenol-chloroform purification protocol (Fuhrman *et al*., 1998). The 16S rDNA gene was PCR amplified with primers 515F-Y and 926R for sequencing on an Illumina MiSeq sequencer as described in (Parada *et al*., 2016) except that 0.5 ng of genomic DNA was used as input and only a single PCR reaction was run and purified per sample. Twenty bases were trimmed off the 3' end of the reverse reads to remove the primer sequence and then were merged with their corresponding forward reads requiring a minimum overlap of 50 bp and a maximum of two differences in the overlap. Merged reads were subsequently quality filtered requiring that the expected number of errors be ≤ 0.1, using usearch (http://drive5.com/usearch/) and the–fastq_maxee flag. Unique ribotypes were identified using Minimum Entropy Decomposition (MED) (Eren *et al*., 2015). Absolute thaumarchaeal cell concentrations were calculated by multiplying total prokaryotic abundances, measured by flow cytometry, by the relative abundance of thaumarchaeal 16S rRNA sequences.

Genome sequencing, assembly, annotation and modification analysis

Approximately 500 ml of enrichment culture from late exponential phase was extracted using the same phenol chloroform protocol above with an additional chloroform extraction step to remove any residual phenol. DNA was sheared with a Covaris S2 to generate fragments of roughly 500 bp that were used for library construction using the NEBNext kit following the manufacturer's instructions (New England Biolabs, Ipswitch, MA, USA). This library was sequenced on a MiSeq, 2×300 bp run that produced a total of 14.6 M paired reads. Genome assembly was done using Spades (Bankevich *et al*., 2012) with default settings on reads with $\leq 50\%$ GC. This yielded six contigs with length > 1 kb and coverage of >1300X, totalling 1.356 Mb. PCR using primers

designed for the ends of each contig were used to determine the orientation of these contigs and subsequent sequencing of PCR products was used to close the genome. The genome was annotated using prokka (Seemann, 2014) using the '–proteins' option to first annotate predicted genes based on the *Ca. N. brevis* and *N. maritimus* genomes. The strain SPOT01 genome has been deposited under the BioProject accession PRJNA341864.

PacBio libraries were constructed using the standard 20 kb library preparation protocol, and two cells were sequenced with the P6/C4 chemistry. A complete Thaumarchaeota genome was assembled using the Hierarchical Genome Assembly tool in the PacBio SMRT Analysis Software (RS_HGAP_Assembly.3) with default settings except that the expected genome size was set to 1.4 Mb and the minimum seed read length was set to 4 kb. Detection of methylation and phosphorothioation modification were performed using the RS_Modification_and_Motif_Analysis.1 protocol in the SMRT Analysis Software using the genome assembled from MiSeq data as the reference genome. Sites were called as modified if their QV scores were ≥ 30.

Sequence analysis and phylogenetic trees

Homologous proteins from thaumarchaeal genomes were identified as those that are best reciprocal blastp (RBH) hits to each other, requiring an E-value of $\leq 1e^{-10}$ and homology over $\geq 67\%$ of the length both proteins. A set of 38 orthologous core genes were selected as those that were found in $\geq 20\%$ of all thaumarchaeal genomes (Supporting Information Table S2). Each set of orthologous proteins were aligned using CLUSTALW and the best tree was found via Phylip using distances computed with the JTT distance matrix and the FITCH algorithm (Felsenstein, 2005). The phylogeny of the 16S rRNA gene was constructed in the same manner except using the F84 DNA substitution model (Kishino and Hasegawa, 1989).

Viral regions were predicted using VirSorter (Roux *et al*., 2015), PHAST (Zhou *et al*., 2011), and phiSpy (Akhter *et al*., 2012). The online tool CRISPRfinder was used to find Clustered regularly interspaced short palindromic repeats (CRISPRs) (Grissa *et al*., 2007). To assess if predicted proteins of strain SPOT01 are likely to be cellular or viral, each protein was searched with blastp or tblastn using default settings against nr representing mostly cellular proteins and two viral databases: the Pacific Ocean Virome (POV) protein clusters (Hurwitz and Sullivan, 2013) (downloaded from http://data.imicrobe.us/project/view/94) and the Moore Foundation Marine Page/Virus metagenome database which includes viral isolate genomes and marine viral metagenomes (http://data.imicrobe.us/project/view/11). Only blast results with an E-value of $\leq 1e^{-5}$ and a bit score of ≥ 50 were considered significant. To compare results, we computed normalized bit scores, the bit score divided by the length of the blast search alignment to control for the fact that the viral databases generally have shorter sequences than those in the nr database. Conserved domains of the putative restriction enzyme operon were identified using CD-Search at The National Center for Biotechnology Information (NCBI).

Competitive blast searches

Following the competitive fragment recruitment methods of (Santoro *et al*., 2015), metagenomic reads were searched against a database of all available thaumarchaeal genomes (Supporting Information Table S3) with their 5S, 16S and 23S rRNA operons excluded. Requiring a maximum E-value of $1e^{-10}$, $\geq 85\%$ identity to any thaumarchaeal genome and an alignment length of ≥ 100 bp, the top matches were retained. The number of reads matching genomes from each clade were normalized to genome length to calculate the relative abundance of each clade in a sample. Genome length was taken as the average genome length for complete genomes within each clade. Since the Deep group consists of only incomplete SAGs, the complete genome size was estimated as 1.63 Mb, the average genome size of all complete marine thaumarchaeal genomes. For metagenomic analysis of communities at the San Pedro Time-series (SPOT) site off of Los Angeles, CA, water samples were collected monthly from June 2012 to July 2013 at 150 m and from 5 m, 45 m (the deep chlorophyll maximum), 500 and 890 m on September 28, 2012. Microbial cells were collected on 0.22 μm filters and DNA was extracted following the protocol in (Cram *et al*., 2015). Illumina compatible metagenomic libraries were constructed as described above and were sequenced on an Illumina HiSeq sequencer, 2 X 250 bp run. Reads were quality filtered with using the Minoche method as described in (Minoche *et al*., 2011) and Eren *et al*. (Eren *et al*., 2013), before recruitment analysis. These metagenomes have been submitted to the European Nucleotide Archive accession PRJEB17887. Metagenomes from the Bermuda Atlantic Times Series (BATS) from August 2002 (samples BATS-167) were downloaded from http://data.imicrobe.us (project CAM_PROJ_BATS) (Glass *et al*., 2015). Metagenomes from the Hawaii Ocean Time-series (HOT) on March 2006 were downloaded from http://data.imicrobe.us (project CAM_PROJ_HOT, 'HOT179_SG' samples) (Martinez *et al*., 2010).

Abundance of PT, viral and urease genes

To estimate the prevalence of PT genes and the viral region in natural populations of close relatives to strain SPOT01, we first identified metagenomic reads with $\geq 85\%$ identity over ≥ 100 bp to the strain SPOT01 genome with the 5S, 16S and 23S rRNA genes excluded, and then calculated the ratio of length normalized read coverage of the gene region to that of the whole genome. The average fraction of Thaumarchaeota that possess the urease operon was estimated by calculating the ratio of length-normalized *ureC* to *amoA* coverage for reads that had $\geq 85\%$ identity for ≥ 100 bp to any marine Thaumarchaeota *ureC* gene found in genomes or environmental clones in Smith *et al*. (2016) and *amoA* genes shown in the tree in Supporting Information Fig. S3. The prevalence of urease operons in the SPOT01 clade was similarly calculated only when using reads whose best hit was to a strain SPOT01 *ureC* or *amoA* sequence and requiring $\geq 85\%$ and $\geq 95\%$ identity respectively, which represent appropriate thresholds of sequence divergence between clades for each gene (Supporting Information Figs S3 and S4).

Acknowledgements

This work was supported by grants from the National Science Foundation (OCE 1136818) and the Gordon and Betty Moore Foundation Marine Microbiology Initiative (GBMF3779) to J.F. D.N. and A.P. each were supported by an NSF Graduate Student Fellowship. We thank Alyson Santoro for kindly providing a sample of the original CN75 culture and growth instructions. We thank Michael S. DeMott for providing us the pJTU1238 strain. The authors declare no competing financial interests or conflicts of interest.

References

Akhter, S., Aziz, R.K., and Edwards, R.A. (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* **40:** 1–13.

Alonso-Sáez, L., Waller, A.S., Mende, D.R., Bakker, K., Farnelid, H., Yager, P.L., *et al.* (2012) Role for urea in nitrification by polar marine Archaea. *Proc Natl Acad Sci USA* **109:** 17989–17994.

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS Microbiol Ecol* **77:** 120–133.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., *et al.* (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19:** 455–477.

Bayer, B., Vojvoda, J., Offre, P., Alves, R.J.E., Elisabeth, N.H., Garcia, J.A.L., *et al.* (2016) Physiological and genomic characterization of two novel marine thaumarchaeal strains indicates niche differentiation. *ISME J* **10:** 1051–1063.

Bhaya, D., Davison, M., and Barrangou, R. (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet* **45:** 273–297.

Biller, S.J., Mosier, A.C., Wells, G.F., and Francis, C.A. (2012) Global biodiversity of aquatic ammonia-oxidizing archaea is partitioned by habitat. *Front Microbiol* **3:** 1–15.

Blow, M.J., Clark, T.A., Daum, C.G., Deutschbauer, A.M., Fomenkov, A., Fries, R., *et al.* (2016) The epigenomic landscape of prokaryotes. *PLoS Genet* **12:** 1–28.

Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008) Mesophilic crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* **6:** 245–252.

Cao, B., Chen, C., DeMott, M.S., Cheng, Q.X., Clark, T.A., Xiong, X.L., *et al.* (2014) Genomic mapping of phosphorothioates reveals partial modification of short consensus sequences. *Nat Commun* **5:** 3951.

Cao, B., Zheng, X.Q., Cheng, Q.X., Yao, F., Zheng, T., Babu, I.R., *et al.* (2015) In vitro analysis of phosphorothioate modification of DNA reveals substrate recognition by a multiprotein complex. *Sci Rep-Uk* **5:** 12513.

Chow, C.E.T., Winget, D.M., White, R.A., Hallam, S.J., and Suttle, C.A. (2015) Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front Microbiol* **6:** 1–15.

Cram, J.A., Chow, C.E.T., Sachdeva, R., Needham, D.M., Parada, A.E., Steele, J.A., and Fuhrman, J.A. (2015) Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J* **9:** 563–580.

Delong, E.F. (1992) Archaea in coastal marine environments. *Proc Natl Acad Sci USA* **89:** 5685–5689.

DeLong, E.F., Taylor, L.T., Marsh, T.L., and Preston, C.M. (1999) Visualization and enumeration of marine planktonic archaea and bacteria by using polyribonucleotide probes and fluorescent in situ hybridization. *Appl Environ Microbiol* **65:** 5554–5563.

Deschamps, P., Zivanovic, Y., Moreira, D., Rodriguez-Valera, F., and Lopez-Garcia, P. (2014) Pangenome evidence for extensive interdomain horizontal transfer affecting lineage core and shell genes in uncultured planktonic Thaumarchaeota and Euryarchaeota. *Genome Biol Evol* **6:** 1549–1563.

Doxey, A.C., Kurtz, D.A., Lynch, M.D.J., Sauder, L.A., and Neufeld, J.D. (2015) Aquatic metagenomes implicate *Thaumarchaeota* in global cobalamin production. *ISME J* **9:** 461–471.

Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2015) Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9:** 968–979.

Eren, A.M., Vineis, J.H., Morrison, H.G., and Sogin, M.L. (2013) A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One* **8:** 1–6.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package), Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J., Olivares, E.C., Clark, T.A., *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7:** 461–465.

Francis, C.A., Roberts, K.J., Beman, J.M., Santoro, A.E., and Oakley, B.B. (2005) Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci USA* **102:** 14683–14688.

Fuhrman, J.A., and Azam, F. (1983) Adaptations of bacteria to marine subsurface waters studied by temperature response. *Mar Ecol Prog Ser* **13:** 95–98.

Fuhrman, J.A., Comeau, D.E., Hagstrom, A., and Chan, A.M. (1998) Extraction from natural planktonic microorganisms of DNA suitable for molecular biological studies. *Appl Environ Microbiol* **54:** 1426–1429.

Fuhrman, J.A., Mccallum, K., and Davis, A.A. (1992) Novel major archaebacterial group from marine plankton. *Nature* **356:** 148–149.

Fuhrman, J.A., Mccallum, K., and Davis, A.A. (1993) Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. *Appl Environ Microbiol* **59:** 1294–1302.

Fuhrman, J.A., and Ouverney, C.C. (1998) Marine microbial diversity studied via 16S rRNA sequences: coastal cloning results and counting of native archaea with fluorescent single cell probes. *Aquat Ecol* **32:** 3–5.

Gan, R., Wu, X.L., He, W., Liu, Z.H., Wu, S.J., Chen, C., *et al.* (2014) DNA phosphorothioate modifications influence the

global transcriptional response and protect DNA from double-stranded breaks. *Sci Rep-Uk* **4:** 6642.

Garcia-Martinez, J., and Rodriguez-Valera, F. (2000) Microdiversity of uncultured marine prokaryotes: the SAR11 cluster and the marine Archaea of Group I. *Mol Ecol* **9:** 935–948.

Geslin, C., Le Romancer, M., Erauso, G., Gaillard, M., Perrot, G., and Prieur, D. (2003) PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, "*Pyrococcus abyssi*". *J Bacteriol* **185:** 3888–3894.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., *et al*. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309:** 1242–1245.

Glass, J.B., Kretz, C.B., Ganesh, S., Ranjan, P., Seston, S.L., Buck, K.N., *et al*. (2015) Meta-omic signatures of microbial metal and nitrogen cycling in marine oxygen minimum zones. *Front Microbiol* **6:** 1–13.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35:** W52–W57.

Haft, D.H., Loftus, B.J., Richardson, D.L., Yang, F., Eisen, J.A., Paulsen, I.T., and White, O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* **29:** 41–43.

Hallam, S.J., Mincer, T.J., Schleper, C., Preston, C.M., Roberts, K., Richardson, P.M., and DeLong, E.F. (2006) Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine *Crenarchaeota*. *PLoS Biol* **4:** 520–536.

Harrison, W.G., Head, E.J.H., Conover, R.J., Longhurst, A.R., and Sameoto, D.D. (1985) The distribution and metabolism of urea in the Eastern Canadian Arctic. *Deep-Sea Res* **32:** 23–42.

Heal, K.R., Qin, W., Ribalet, F., Bertagnolli, A.D., Coyote-Maestas, W., Hmelo, L.R., *et al*. (2017) Two distinct pools of B-12 analogs reveal community interdependencies in the ocean. *Proc Natl Acad Sci USA* **114:** 364–369.

Holmes, R.M., Aminot, A., Kerouel, R., Hooker, B.A., and Peterson, B.J. (1999) A simple and precise method for measuring ammonium in marine and freshwater ecosystems. *Can J Fish Aquat Sci* **56:** 1801–1808.

Hurwitz, B.L., and Sullivan, M.B. (2013) The Pacific Ocean Virome (POV): A marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8:** 1–12.

Ingalls, A.E., Shah, S.R., Hansman, R.L., Aluwihare, L.I., Santos, G.M., Druffel, E.R.M., and Pearson, A. (2006) Quantifying archaeal community autotrophy in the mesopelagic ocean using natural radiocarbon. *Proc Natl Acad Sci USA* **103:** 6442–6447.

Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., and Armbrust, E.V. (2012) Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota. *Science* **335:** 587–590.

Karner, M.B., DeLong, E.F., and Karl, D.M. (2001) Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409:** 507–510.

Kishino, H., and Hasegawa, M. (1989) Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA-sequence data, and the branching order in hominoidea. *J Mol Evol* **29:** 170–179.

Könneke, M., Bernhard, A.E., de la Torre, J.R., Walker, C.B., Waterbury, J.B., and Stahl, D.A. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437:** 543–546.

Könneke, M., Schubert, D.M., Brown, P.C., Hugler, M., Standfest, S., Schwander, T., *et al*. (2014) Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. *Proc Natl Acad Sci USA* **111:** 8239–8244.

Konstantinidis, K.T., and Tiedje, J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102:** 2567–2572.

Krupovic, M., Spang, A., Gribaido, S., Forterre, P., and Schleper, C. (2011) A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. *Biochem Soc Trans* **39:** 82–88.

Labonte, J.M., Swan, B.K., Poulos, B., Luo, H.W., Koren, S., Hallam, S.J., *et al*. (2015) Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. *ISME J* **9:** 2386–2399.

Lehtovirta-Morley, L.E., Ge, C.R., Ross, J., Yao, H.Y., Nicol, G.W., and Prosser, J.I. (2014) Characterisation of terrestrial acidophilic archaeal ammonia oxidisers and their inhibition and stimulation by organic compounds. *FEMS Microbiol Ecol* **89:** 542–552.

Low, D.A., Weyand, N.J., and Mahan, M.J. (2001) Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect Immun* **69:** 7197–7204.

Luo, H., Sun, Y., Hollibaugh, J.T., and Moran, M.A. (2016) Low Genome Content Diversity of Marine Planktonic Thaumarchaeota. *Environ Microbiol Rep* **8:** 501–507.

Luo, H.W., Tolar, B.B., Swan, B.K., Zhang, C.L.L., Stepanauskas, R., Moran, M.A., and Hollibaugh, J.T. (2014) Single-cell genomics shedding light on marine Thaumarchaeota diversification. *ISME J* **8:** 732–736.

Marinus, M.G., and Casadesus, J. (2009) Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more. *FEMS Microbiol Rev* **33:** 488–503.

Martens-Habbena, W., Qin, W., Horak, R.E.A., Urakawa, H., Schauer, A.J., Moffett, J.W., *et al*. (2015) The production of nitric oxide by marine ammonia-oxidizing archaea and inhibition of archaeal ammonia oxidation by a nitric oxide scavenger. *Environ Microbiol* **17:** 2261–2274.

Martinez, A., Tyson, G.W., and DeLong, E.F. (2010) Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol* **12:** 222–238.

Miller, E.S., Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Durkin, A.S., Ciecko, A., *et al*. (2003) Complete genome sequence of the broad-host-range vibriophage KVP40: Comparative genomics of a T4-related bacteriophage. *J Bacteriol* **185:** 5220–5233.

Minoche, A.E., Dohm, J.C., and Himmelbauer, H. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* **12:** R112.

Mosier, A.C., Allen, E.E., Kim, M., Ferriera, S., and Francis, C.A. (2012) Genome sequence of "*Candidatus Nitrosopumilus salaria*" BD31, an ammonia-oxidizing archaeon from the San Francisco Bay Estuary. *J Bacteriol* **194:** 2121–2122.

Navarathna, D.H.M.L.P., Das, A., Morschhauser, J., Nickerson, K.W., and Roberts, D.D. (2011) DUR3 is the major urea transporter in *Candida albicans* and is co-regulated with the urea amidolyase DUR1,2. *Microbiol-Sgm* **157:** 270–279.

Ouellette, M., Jackson, L., Chimileski, S., and Papke, R.T. (2015) Genome-wide DNA methylation analysis of *Haloferax volcanii* H26 and identification of DNA methyltransferase related PD-(D/E)XK nuclease family protein HVO_A0006. *Front Microbiol* **6:** 1–11.

Pagaling, E., Haigh, R.D., Grant, W.D., Cowan, D.A., Jones, B.E., Ma, Y., *et al.* (2007) Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. *BMC Genomics* **8:** 410.

Painter, S.C., Sanders, R., Waldron, H.N., Lucas, M.I., and Torres-Valdes, S. (2008) Urea distribution and uptake in the Atlantic Ocean between 50 degrees N and 50 degrees. *Mar Ecol Prog Ser* **368:** 53–63.

Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18:** 1403–1414.

Park, S.J., Ghai, R., Martin-Cuadrado, A.B., Rodriguez-Valera, F., Chung, W.H., Kwon, K., *et al.* (2014) Genomes of two new ammonia-oxidizing archaea enriched from deep marine sediments. *PLoS One* **9:** 1–10.

Pedneault, E., Galand, P.E., Polvin, M., Tremblay, J.E., and Lovejoy, C. (2014) Archaeal *amoA* and *ureC* genes and their transcriptional activity in the Arctic Ocean. *Sci Rep-UK* **4:** 4661.

Pfister, P., Wesserfallen, A., Stettler, R., and Leisinger, T. (1998) Molecular analysis of *Methanobacterium* phage Psi M2. *Mol Microbiol* **30:** 233–244.

Pingoud, A., and Jeltsch, A. (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Res* **29:** 3705–3727.

Pittera, J., Humily, F., Thorel, M., Grulois, D., Garczarek, L., and Six, C. (2014) Connecting thermal physiology and latitudinal niche partitioning in marine Synechococcus. *ISME J* **8:** 1221–1236.

Prangishvili, D., Forterre, P., and Garrett, R.A. (2006) Viruses of the Archaea: a unifying view. *Nat Rev Microbiol* **4:** 837–848.

Qin, W., Amin, S.A., Martens-Habbena, W., Walker, C.B., Urakawa, H., Devol, A.H., *et al.* (2014) Marine ammonia-oxidizing archaeal isolates display obligate mixotrophy and wide ecotypic variation. *Proc Natl Acad Sci USA* **111:** 12504–12509.

Remsen, C.C. (1971) Distribution of urea in coastal and oceanic waters. *Limnol Oceanogr* **16:** 732–740.

Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2015) REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **43:** D298–D299.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424:** 1042–1047.

Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., *et al.* (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537:** 689–693.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3:** e985.

Santoro, A.E., and Casciotti, K.L. (2011) Enrichment and characterization of ammonia-oxidizing archaea from the open ocean: phylogeny, physiology and stable isotope fractionation. *ISME J* **5:** 1796–1808.

Santoro, A.E., Casciotti, K.L., and Francis, C.A. (2010) Activity, abundance and diversity of nitrifying archaea and bacteria in the central California Current. *Environ Microbiol* **12:** 1989–2006.

Santoro, A.E., Dupont, C.L., Richter, R.A., Craig, M.T., Carini, P., McIlvin, M.R., *et al.* (2015) Genomic and proteomic characterization of "*Candidatus Nitrosopelagicus brevis*": An ammonia-oxidizing archaeon from the open ocean. *Proc Natl Acad Sci USA* **112:** 1173–1178.

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

Smith, J.M., Damashek, J., Chavez, F.P., and Francis, C.A. (2016) Factors influencing nitrification rates and the abundance and transcriptional activity of ammonia-oxidizing microorganisms in the dark northeast Pacific Ocean. *Limnol Oceanogr* **61:** 596–609.

Strickland, J., and Parsons, T. (1968). "A practical handbook of seawater analysis." *Fish Res Bd Can Bull* **167:** 71–75.

Swan, B.K., Chaffin, M.D., Martinez-Garcia, M., Morrison, H.G., Field, E.K., Poulton, N.J., *et al.* (2014) Genomic and metabolic diversity of Marine Group I Thaumarchaeota in the mesopelagic of two subtropical gyres. *PLoS One* **9:** 1–9.

Teira, E., Lebaron, P., van Aken, H., and Herndl, G.J. (2006) Distribution and activity of Bacteria and Archaea in the deep water masses of the North Atlantic. *Limnol Oceanogr* **51:** 2131–2144.

Teira, E., Reinthaler, T., Pernthaler, A., Pernthaler, J., and Herndl, G.J. (2004) Combining catalyzed reporter deposition-fluorescence in situ hybridization and microautoradiography to detect substrate utilization by bacteria and archaea in the deep ocean. *Appl Environ Microbiol* **70:** 4411–4414.

Teira, E., van Aken, H., Veth, C., and Herndl, G.J. (2006) Archaeal uptake of enantiomeric amino acids in the meso- and bathypelagic waters of the North Atlantic. *Limnol Oceanogr* **51:** 60–69.

Tolar, B.B., Ross, M.J., Wallsgrove, N.J., Liu, Q., Aluwihare, L.I., Popp, B.N., and Hollibaugh, J.T. (2016) Contribution of ammonia oxidation to chemoautotrophy in Antarctic coastal waters. *ISME J* **10:** 2605–2619.

Tully, B.J., Nelson, W.C., and Heidelberg, J.F. (2012) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. *Environ Microbiol* **14:** 254–267.

Vasu, K., and Nagaraja, V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* **77:** 53–72.

Walker, C.B., de la Torre, J.R., Klotz, M.G., Urakawa, H., Pinel, N., Arp, D.J., *et al.* (2010) *Nitrosopumilus maritimus*

genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* **107:** 8818–8823.

Wang, L.R., Chen, S., Vergin, K.L., Giovannoni, S.J., Chan, S.W., DeMott, M.S., *et al*. (2011) DNA phosphorothioation is widespread and quantized in bacterial genomes. *Proc Natl Acad Sci USA* **108:** 2963–2968.

Wang, L.R., Chen, S., Xu, T.G., Taghizadeh, K., Wishnok, J.S., Zhou, X.F., *et al*. (2007) Phosphorothioation of DNA in bacteria by dnd genes. *Nat Chem Biol* **3:** 709–710.

Wilson, G.G., and Murray, N.E. (1991) Restriction and Modification Systems. *Annu Rev Genet* **25:** 585–627.

Wuchter, C., Abbas, B., Coolen, M.J.L., Herfort, L., van Bleijswijk, J., Timmers, P., *et al*. (2006) Archaeal nitrification in the ocean. *Proc Natl Acad Sci USA* **103:** 12317–12322.

Xu, T.G., Liang, J.D., Chen, S., Wang, L.R., He, X.Y., You, D.L., *et al*. (2009) DNA phosphorothioation in *Streptomyces lividans*: mutational analysis of the dnd locus. *BMC Microbiol* **9:** 41.

Xu, T.G., Yao, F., Zhou, X.F., Deng, Z.X., and You, D.L. (2010) A novel host-specific restriction system associated with DNA backbone S-modification in *Salmonella*. *Nucleic Acids Res* **38:** 7133–7141.

Yao, F., Xu, T.G., Zhou, X.F., Deng, Z.X., and You, D.L. (2009) Functional analysis of spfD gene involved in DNA phosphorothioation in *Pseudomonas fluorescens* Pf0-1. *FEBS Lett* **583:** 729–733.

Yool, A., Martin, A.P., Fernandez, C., and Clark, D.R. (2007) The significance of nitrification for oceanic new production. *Nature* **447:** 999–1002.

Zhou, X.F., He, X.Y., Liang, J.D., Li, A.Y., Xu, T.G., Kieser, T., *et al*. (2005) A novel DNA modification by sulphur. *Mol Microbiol* **57:** 1428–1438.

Zhou, Y., Liang, Y.J., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011) PHAST: a fast phage search Tool. *Nucleic Acids Res* **39:** W347–W352.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Table S1. Fraction of shared genes (reciprocal best blast hits, RHBs) shared between isolate or enrichment thaumarchaeal genomes. Values below the diagonal are the fraction of shared genes divided by the number of genes in the genomes listed by row and values above the diagonal are the fraction of shared genes divided by the number of genes in the genomes listed by columns.

Table S2. The 38 core genes shared among marine thaumarchaeal genomes and used to construct the concatenated phylogenomic tree in Figure 1. See spreadsheet file available online: Suppelemental_Table_2.xlsx.

Table S3. Thaumarchaeal genomes from enrichments or single-cell amplified genomes (SAGs) used in this study. Accessions with letters are NCBI accessions, 10 digit numbers are IMG accessions.

Table S4. Average nucleotide identity (ANI) between genomes of 'Shallow' group marine thaumarchaeal strains. ANI is calculated as the average nucleotide identity of reciprocal best blast hits (RBHs) between two genomes whereby RBHs must have an evalue $\leq$ 1e-10 over $\geq$ 67% of the length of the gene and only genes with a nucleotide identity of $\geq$70% are included. Within clade ANI values are shaded in colors corresponding to the clade colors designated in Fig. 1, to highlight that within clade ANI values are typically $\geq$ 84% whereas between clade ANI values are almost always $<$ 84%. For reference, ANI values between all 'Deep' group SAG genomes are $\geq$ 84% (on average 90%) and ANI values between 'Deep' group SAG genomes and 'Shallow' group genomes are $\leq$ 78.3% (data not shown for simplicity).

Table S5. Table adapted from Santoro *et al*. (2015) that identifies in SPOT01, homologs of *N. brevis* proteins involved in key metabolic pathways. Spreadsheet file available online: Supplemental_Table_5_Nbrevis_homologs.xlsx.

Fig. S1. Prokaryotic composition of SPOT01 enrichment cultures assessed by 16S rRNA amplicon sequencing (Illumina MiSeq). **A**) In the representative culture experiments from Fig. 2A, nine ribotypes were found in the SPOT01 enrichment cultures: seven belonged to Thaumarchaeota, and one each belong to the genera *Sphingomonas* and *Erythrobacter*. The dominant ribotype corresponding to the strain SPOT01 ('SPOT01') represented $\geq$ 97% of the Thaumarchaeota in the enrichment culture. Community composition is shown for samples from the representative growth curve shown in Fig. 2 ('Day *n*') and the DNA used for PacBio sequencing and the PT iodine assay in Fig. 4 ('PacBio'). **B)** Composition of archaeal and bacterial oligotypes in enrichment cultures grown at different temperatures and sampled during exponential phase. Twelve thauarchaeal oligotypes were found (each labeled with a unique node number, e.g. '_N101' or with 'SPOT01' for the oligotype corresponding to strain SPOT01). Sixteen bacterial oligotypes were found and belonged to the genera *Sphignomonas* ('Sphingo_'), *Erythrobacter* ('Erythro_'), *Mesorhizobium* ('Mesorhiz_'); the SAR86 or SAR11 marine bacteria clades; the classes Gammaproteobacteria ('Gamma') or Alphaproteobacteria ('Alpha'); or unknown phylogeny ('Unkn_').

Fig. S2. Temperature versus growth rate curves from Figure 2B whereby for each strain points are replotted as relative growth rates normalized to the maximum growth rate (set as 1) for each strain.

Fig. S3. Phylogeny of representative *amoA* nucletoide sequences from isolate or enrichment genomes (red), single cell amplified genomes (blue) or environmental clone sequences (blue). The corresponding Water Column groups A and B (WCA and WCB) as defined by Francis *et al*. (2005) are shown. Note that the SPOT01 *amoA* sequence belongs to a distinct clade of sequences that was previously only represented by uncultured sequences (SAG or environmental clones). The tree was constructed using the F84 nucleotide substitution model and the FITCH algorithm in phylip. Values at the nodes indicate bootstrap values (100 replicates) when $>$ 50. Listed in parentheses are gene locus tags for sequences from genomes or NCBI protein accessions for environmental clone sequences.

Fig. S4. Phylogeny of thaumarchaeal *ureC* sequences from enrichment or isolate genomes (red), SAGs (blue) and environmental clones (black) from Smith *et al*. (2016). *ureC* sequences were aligned and a 998 nucleotide region for which most of positions were present was used for construction of the tree. The tree was constructed using the HKY85 nucleotide substitution model with gamma

distributed rates and invariable sites (HKY85+i+g) and minimum evolution as the search criterion. Values at the nodes indicate bootstrap values (100 replicates) when >50. Listed in parentheses are gene locus tags for sequences from genomes or NCBI nucleotide accession numbers for environmental clone sequences. For simplicity, clades with several closely sequences are depicted with trapezoids. To the right of each trapezoid is listed the number of sequences in that clade and names of representative genome and/or cloned sequenced in that clade.

Fig. S5. The organization of *dnd* homologs and adjacent restriction enzyme genes in SPOT01, other bacteria, and archaeal contigs. For thaumarchaeal fosmids KM3_186_C08 and KM3_47_F06, the *dnd* genes are located near the end of the fosmid, depicted as a vertical bar at position zero, and the dashed line on the opposite end indicates that the fosmid continues on in that direction.

Fig. S6. Results from the virus prediction tool phiSpy. The purple line depicts gene ranks output from phiSpy and the dotted lined shows the default threshold used by phiSpy (half of the maximum rank score) to predict viral regions is shown as a dashed line. The primary peak in phiSpy scores (red line) is centered on an operon of ribosomal proteins and potentially represents a false positive result. The dashed line represents an alternative threshold: half of the maximum rank score when excluding the primary peak in rank scores. Predicted regions are shown at the top of the graph: phiSpy using the default threshold (red), phiSpy using the alternate threshold (grey), and the VirSorter predicted region (green) for reference. The resulting phiSpy regions using this alternative threshold are those depicted in Fig. 5.