The ISME Journal (2019) 13:618–631
https://doi.org/10.1038/s41396-018-0289-4

ARTICLE

Discovery of several novel, widespread, and ecologically distinct marine *Thaumarchaeota* viruses that encode *amoC* nitrification genes

Nathan A. Ahlgren[1,3] · Clara A. Fuchsman[2,4] · Gabrielle Rocap[2] · Jed A. Fuhrman[1]

Received: 21 February 2018 / Revised: 15 June 2018 / Accepted: 19 September 2018 / Published online: 12 October 2018

Abstract

Much of the diversity of prokaryotic viruses has yet to be described. In particular, there are no viral isolates that infect abundant, globally significant marine archaea including the phylum *Thaumarchaeota*. This phylum oxidizes ammonia, fixes inorganic carbon, and thus contributes to globally significant nitrogen and carbon cycles in the oceans. Metagenomics provides an alternative to culture-dependent means for identifying and characterizing viral diversity. Some viruses carry auxiliary metabolic genes (AMGs) that are acquired via horizontal gene transfer from their host(s), allowing inference of what host a virus infects. Here we present the discovery of 15 new genomically and ecologically distinct *Thaumarchaeota* virus populations, identified as contigs that encode viral capsid and thaumarchaeal ammonia monooxygenase genes (*amoC*). These viruses exhibit depth and latitude partitioning and are distributed globally in various marine habitats including pelagic waters, estuarine habitats, and hydrothermal plume water and sediments. We found evidence of viral *amoC* expression and that viral *amoC* AMGs sometimes comprise up to half of total *amoC* DNA copies in cellular fraction metagenomes, highlighting the potential impact of these viruses on N cycling in the oceans. Phylogenetics suggest they are potentially tailed viruses and share a common ancestor with related marine *Euryarchaeota* viruses. This work significantly expands our view of viruses of globally important marine *Thaumarchaeota*.

Introduction

Marine *Thaumarchaeota* are abundant, nitrifying chemolithotrophs that carry out ammonia oxidation and fix inorganic carbon [1–3], and they therefore contribute significantly to important nitrogen and carbon cycles in the oceans. They are often found in high abundances just below the deep chlorophyll maximum in the upper ocean [4, 5], and in the deeper mesopelagic ocean, they can comprise > 25% of the total prokaryotic cells [6, 7]. Understanding the factors that control their growth and abundance are critical to our view of N and C cycling in the oceans [5, 8, 9], but there is limited knowledge about top–down controls of these important microbes. In particular, viruses have yet to be isolated in culture that infect marine *Thaumarchaeota*, or any mesophilic oceanic archaea for that matter. The only culture-based observation of a marine archaeal virus comes from virus-like particles observed in the culture of a hyperthermophilic, deep-sea hydrothermal vent archaeon, *Pyrococcus abyssi* [10, 11]. A wide diversity of archaeal viruses, however, have been isolated from other habitats, including, for example, haloviruses from hypersaline evaporated salterns, but the isolation and knowledge of viruses infecting bacteria (bacteriophage) still far outweighs that of archaeal viruses [12–14].

While cultured-based approaches for discovering new archaeal viruses have had limited success, new culture-independent, high-throughput sequencing approaches provide valuable means for discovering new viruses in general [15].

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41396-018-0289-4) contains supplementary material, which is available to authorized users.

✉ Nathan A. Ahlgren
nahlgren@clarku.edu

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

[2] School of Oceanography, University of Washington, Seattle, WA, USA

[3] Present address: Clark University, Worcester, MA, USA

[4] Present address: Horn Point Laboratory, University of Maryland, Cambridge, MD, USA

Table 1 Amino-acid and structure similarity search results for ETNP_CA_420 predicted proteins

Gene	Top results from blastp search against NCBI *nr*[a]			Top structural similarity results from HHPred[b]	
	E-value	Amino-acid identity	Description of best protein hit	*E*-value	Description of most similar protein structure
ETNP_CA_420_2	n.s.			1.00E–08	*Lactococcus* phage TP901-1 baseplate
ETNP_CA_420_3	n.s.			1.00E–21	*Thermus* phage P7426 portal protein G20C
ETNP_CA_420_6	1.00E–12	33%	Protease domain of thaumarchaeal provirus Pro-Nvie1 protease/major capsid protein	n.s.	
ETNP_CA_420_8	2.00E–41	33%	Capsid domain of thaumarchaeal provirus Pro-Nvie1 protease/major capsid protein	6.00E–07	*Synechococcus* phage Syn5 capsid Gp39
ETNP_CA_420_9	3.00E–139	95%	AmoC, *Ca. Nitrosopelagicus brevis*	n.s.	
ETNP_CA_420_10	4.00E–06	39%	Hypothetical protein, soil Thaumarchaeota 13_1_20CM_2_39_20	n.s.	
ETNP_CA_420_21	4.00E–31	35%	CobS: Candidatus Odinarchaeota archaeon LCB_4	4.00E–14	*Halothiobacillus neapolitanus c2* CbbQ carboxysome subunit
	4.00E–10	29%	ATPase: Candidatus Micrarchaeum acidiphilum ARMAN-2	n.s.	
	4E–6 to 7E–10	24–28%	Cyanophage CobS porphyrin biosynthetic protein	n.s.	

n.s. no significant result (*E*-value ≥ 1E–5)

[a]Only significant results with *E*-value < 1E–5 and bit score ≥50 are listed. The best hit is listed for each gene except for ETNP_CA_420_21, for which multiple top hits are shown

[b]Only significant HHPred results are shown (*E*-value < 1E–5)

Several recent metagenomic and single-cell genome studies specifically have provided evidence for the existence of viruses that infect marine *Thaumarchaeota*. These include recovery of marine prokaryotic fraction fosmids with similarity to a Pro-Nvie1, a probable provirus found in the genome of *Nitrososphaera viennensis* EN76, a terrestrial thaumarchaeon [16, 17]; virus sequences recovered from a thaumarchaeal single-cell amplified genome [18]; probable virus genes in the genome of *Candidatus Nitrosomarinus catalina* SPOT01, a cultured, marine thaumarchaeon [19]; and an 11.6 kb contig, GOV_bin_4552_contig-100_2, assembled from a viral fraction metagenome that contains viral capsid genes and a thaumarchaeal *amoC* gene [21]. AmoC is a subunit of the ammonia monooxygenase responsible for ammonia oxidation from which *Thaumarchaeota* derive energy [22].

Viruses often acquire metabolic genes from their hosts through horizontal gene transfer, and these so-called auxiliary metabolic genes (AMGs) are thought to bolster the metabolism of the infected host cells [23–25]. Identification of AMGs in viral genomes provides a convincing piece of evidence to connect viral sequences to their probable host(s). The recent discovery of a thaumarchaeal *amoC* gene on the viral contig GOV_bin_4552_contig-100_2 therefore provides strong evidence that this contig represents a thaumarchaeal virus. Recently developed *k*-mer-based tools, such as VirHostMatcher, can also help predict the probable host of metagenomic viral sequences by matching them to host genome sequences with which they have the highest similarity in nucleotide word usage patterns [20, 26–28]. These tools take advantage of the phenomenon that many viruses exhibit similar nucleotide usage patterns as their host probably due to strong selective pressures to use similar amino-acid codons as their host.

In order to identify additional new thaumarchaeal virus sequences, we have applied two viral contig identification tools VirSorter [27, 28] and VirFinder [41] to metagenomes from the Eastern Tropical North Pacific (ETNP) and other publically available metagenomes, to identify other viral contigs that encode thaumarchaeal *amoC* genes. In this way, we have identified 32 new probable thaumarchaeal virus contigs, representing 15 putative viral species that are distributed globally; are found in a variety of marine habitats; and appear to occupy distinct marine niches. We have also used the host prediction tool VirHostMatcher [20] to confirm *Thaumarchaeota* as the probable host of these viruses as well corroborate potential specific interactions between corresponding depth-partitioned *Thaumarchaeota* host and viral populations. Finally we provide phylogenetic evidence that these viruses are probably tailed and that some of them share a common evolutionary ancestor with marine *Euryarchaeota* viruses.

Methods

Sample collection and Metagenomic sequencing and assembly

Water samples were collected in April 2012 during cruise TN278 aboard the R/V Thompson using 10 L Niskin bottles on a 24 bottle sampling rosette. A Seabird 911 Conductivity Temperature Density meter and a Seabird SBE 43 Dissolved Oxygen Sensor were attached to the rosette.

DNA samples were obtained from 0.2 µm SUPOR filters from station 136 (–106.543°W 17.043°N) at 10 depths between 60 and 300 m, which included the oxycline and anoxic zones. Metagenomes from these samples have been previously published, and the sampling and processing methods are described therein [29]. Metagenomic reads and assembled contigs for individual samples can be found at GenBank BioProject PRJNA350692.

In this study, metagenomes from the 70 and 90 m samples were co-assembled with IDBA_UA [66] using default settings, and these contigs are also available under Genbank BioProject PRJNA350692 (contigs are named ETNP_CA_X). Protein encoding genes on these contigs were predicted and annotated using Prodigal using default settings [30].

Analysis of viral contigs and delineation of viral populations

Sequence similarity between contigs or genes on contigs were performed with blastn or blastp using default settings. Unless noted, only significant results were considered (*E*-value < 1E–5, Bit score ≥ 50). Structural prediction and similarity analyses were done using HHPred via their online interface (https://toolkit.tuebingen.mpg.de/#/hhpred) using standard settings [31]. Only significant results (*E*-value < 1E–5) were considered. Fifteen distinct viral populations were identified using average nucleotide identity (ANI) values determined from blastn results for predicted coding regions of the contigs and applying a 95% cutoff. Gene pair identities were only included in ANI averages if the alignment was over a minimum of 50% of the query or subject gene and the percent identity was ≥70%.

Fragment recruitment analyses

Reads from metagenomes were mapped to a collection of representative viral contigs, one from each viral species population (typically the longest contig in the species) (Table 1) using the bbsplit.sh script in the BBTools package (https://jgi.doe.gov/data-and-tools/bbtools/) with a minimum percent identity of 95% (minid = 0.95) to match reads to the most similar contig. The script bbsplit.sh assigns each read to the best matching contig. Tara Ocean metagenomes were downloaded from the European Nucleotide Archive using accession numbers provided in [32] and [21]. Read data used from Vik et al. [33], Hollibaugh et al. [34], Thrash et al. [35], and Oulas et al. [36] were downloaded from iVirus, the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA), or the Joint Genome Institute's (JGI) Integrated Microbial Genomes & Microbiomes (IMG/M) website using accessions or file names provided in those studies. *amoC* reads were assigned as cellular or viral by "placement" of reads on the reference *amoC* tree (Fig. 1) [29, 67]. This tree was constructed using the HKY + i + g model using parameters estimated with

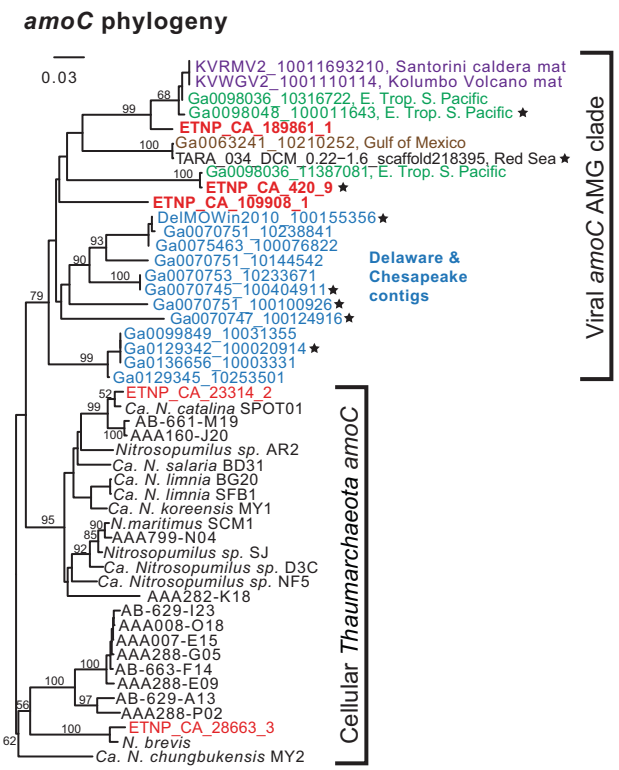

Fig. 1 Phylogeny of *Thaumarchaeota amoC* sequences from cellular *Thaumarchaeota* genomes and thaumarchaeal viral contigs. The tips of the tree are labeled with gene identifiers (viral contigs) or isolate or single-cell amplified genome names. Genes from contigs obtained from co-assembly of ETNP metagenomes from this study are depicted in red; genes from contigs from the Delaware and Chesapeake Bay are depicted in blue; and the locations from which all other contigs were assembled are listed after gene numbers. The tree was constructed using the HKY85 DNA substitution model with invariable sites and gamma distributed rates of evolution and heuristic search of tree space using minimum evolution as the criterion. Numbers at the nodes indicate results from bootstrap analysis (100 replicates). Bootstrap analysis supports that *amoC* sequences from viral contigs sequences ("Viral *amoC* AMG clade") form a phylogenetically distinct clade from cellular *Thaumarchaeota* sequences ("Cellular *Thaumarchaeota amoC*"). Genes marked with stars indicate representative contigs from putative viral species delineated by average nucleotide identity (Fig. S3) that were used for measuring population abundances in various habitats (Fig. 3)

modeltest [37] and minimum evolution as the criterion. Capsid reads were quantified by placement on the Pro-Nvie1-like capsid gene tree. Reads were aligned to *amoC* reference sequences in nucleotide space using PaPaRa: Parsimony-based Phylogeny-Aware Read Alignment program[69] . Paired end reads were combined into the same alignment using a python script and placed as one on the tree using the EPA: Evolutionary Placement Algorithm portion of RAxML [68]. Each read has a number, or "branchlength", which corresponds to the similarity between the read and the sequence to which it is placed. Reads placed with a read "branchlength" longer than 2.0 were removed as erroneous. Spot testing indicated that these reads belonged to different genes than the one examined. Only a small percentage of reads were thus removed (0.1%).

Host prediction analyses with VirHostMatcher

For prediction of the probable host phylum of ≥10 kb thaumarchaeal viruses, nucleotide similarity scores (d_2^*) were calculated using VirHostMatcher for each viral contig against a database of ~5700 possible marine prokaryotic hosts that includes marine host genomes identified in [20] and metagenomically assembled genomes from [38, 39] (listed in Supplemental File 1). For each phylum of hosts in the database, we computed the difference in the mean of scores to that phylum and the mean of scores to all other phyla and normalized this difference with the standard deviation of the scores of the "all other phyla" group. The predicted host was selected as the phylum with the strongest normalized difference in mean scores, i.e., the phylum with the largest negative deviation in similarity when compared with all other phyla. Only phyla with six or more genomes in the database were included ($n = 26$ out of 41 possible phyla), representing 5607 possible host genomes.

For the more specific prediction of whether the viral contigs represent viruses that likely infect *Thaumarchaeota* from the "Deep" or "Shallow" group hosts, VirHostMatcher was applied to a database of *Thaumarchaeota* genomes from isolates or SAGs from the "Deep" ($n = 16$) or "Shallow" ($n = 42$) groups as determined by phylogenomics in [19]. *t*-tests were applied to determine if there were significant differences in the VirHostMatcher score means of comparisons with the Deep and Shallow thaumarchaeal host genomes for each virus.

Results and discussion

In order to identify additional new thaumarchaeal virus sequences, we first analyzed prokaryotic cellular fraction (>0.22 µm) metagenomes collected from above an oxygen minimum zone (OMZ) from the ETNP where *Thaumarchaeota* were prevalent. Thaumarchaeal-specific *amoA* qPCR assays previously showed that *Thaumarchaeota* cells were abundant (1.9×10^5 to 1.0×10^5 copies per mL) and localized at 70–100 m [40] (Fig. S1). They represented up to 12% of prokaryotes at 70–100 m based on the metagenomic analysis of the single copy RNA polymerase gene *rpoB* (Fig. S1) [29]. In congruence, ammonia oxidation rates were highest (14–31 nM/d) at 70–100 m [40]. To identify viral contigs among metagenomic assemblies containing mixtures of host and viral sequences, the viral detection programs VirSorter [27, 28] and VirFinder [41] were applied to pick out viral contigs. Proteins encoded on these viral contigs were then searched against thaumarchaeal genome proteins to identify viral contigs with potential thaumarchaeal AMGs. Among contigs from co-assembly of the ETNP metagenomes from 70 and 90 m, we identified a 26 kb viral contig, named ETNP_CA_420, which encodes a thaumarchaeal-like *amoC* gene (Figs. 1, 2), representing a probable AMG (see below). Although contig ETNP_CA_420 only had a VirSorter category III prediction result ("possible" viral contig), it had a high and statistically significant VirFinder prediction score (score: 0.91, $p = 0.012$, Table S1), highlighting the utility of using multiple virus prediction tools. The discovery of ETNP_CA_420 builds on the recent, similar identification of a 12.2 kb contig, GOV_bin_4552_contig-100_2, which was assembled from viral fraction metagenomes and encodes both a viral capsid gene and a thaumarchaeal *amoC* gene [21]. Note that in our subsequent analyses, we have instead used a longer 14.6 kb contig named TARA_034_DCM_0.22−1.6_scaffold218395_1 because GOV_bin_4552_contig-100_2 is a 99.98% identical subfragment of TARA_034_DCM_0.22−1.6_scaffold218395_1. The latter contig was recovered from individual sample assembly of a cellular fraction (0.22–1.6 µm) Tara Ocean sample from the Red Sea [42].

In concordance with the VirFinder prediction result, all of ETNP_CA_420's 21 genes are encoded on the same strand (Fig. 2), a trait characteristic of viral genomes [27, 28]. More importantly, several of its predicted proteins exhibit sequence and/or structural similarity to known viral structural proteins of previously characterized viruses (Table 1). The proteins to which ETNP_CA_420's predicted genes have similarity include those that make up the main capsid structure; a portal protein, which forms the opening through which DNA moves in and out of the capsid; and a baseplate protein, which occurs at the end of tailed viruses. We highlight genes ETNP_420_6 and ETNP_420_8 that exhibit similarity to the protease and capsid domains, respectively, of the combined protease/capsid protein of Pro-Nvie1, a probable provirus discovered in the genome of the soil thaumarchaeon *Nitrososphaera*

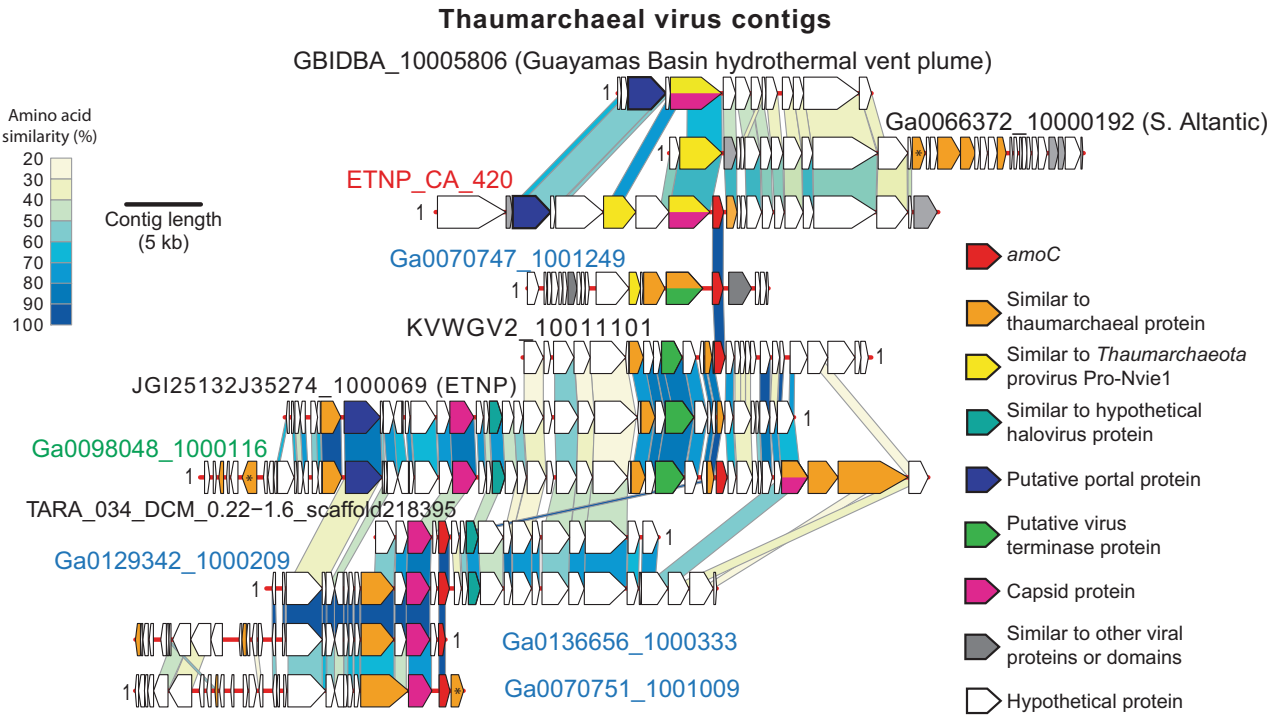

Fig. 2 Contig maps depicting predicted proteins encoded on representative thaumarchaeal viral contigs. Arrows depict the location and direction of predicted proteins on contigs, and the number 1 indicates the end with first nucleotide position of the contig. Fill colors indicate different categories of genes, as indicated in the legend, according to top hits from searches against NCBI's *nr* protein database or protein structure similarity analyses (Table 1). Asterisks denote which genes showed similarity to *Ca. N. catalina* SPOT01 putative viral gene NMSP_1228. The color of the trapezoids connecting genes indicate amino-acid identities between genes.

viennensis EN76 [16] (Table 1). The protease may serve a role in capsid maturation [16]. Also of interest was gene ETNP_CA_420_21 that shows distant (≤28% protein identity) similarity to several cyanophage isolate proteins annotated as CobS, the porphyrin biosynthetic enzyme responsible for the last step in vitamin B_{12} synthesis (Table 1). These cyanophage proteins probably do not synthesize B_{12} because they are phylogenetically distant from any host protein [43]. HHPred results show that ETNP_CA_420_21 has structural similarity instead to a protein involved in carboxysomes, intracellular organelles used to concentrate CO_2 around Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) [44]. In any case, gene ETNP_CA_420_21 has higher sequence similarity to cellular archaeal contigs assembled from a Yellowstone hot spring affiliated with the newly described Asgard 'superphylum' of archaea (*Candidatus Odinarchaeota archaeon* LCB_4) [45] and from an acid mine drainage system (*Candidatus Microarchaeum acidiphilum* ARMAN-2) [46] (Table 1). We suggest genes on these contigs could be viral genes integrated or horizontally transferred into the genomes of these distantly related archaea.

The *amoC* genes and capsid proteins of contigs ETNP_CA_420 and TARA_034_DCM_0.22−1.6_scaffold218395_1 were then used to identify additional thaumarchaeal virus sequences by similarity searches to contigs from publically available metagenomic sequencing projects at the IMG/M database. The resulting contigs were recovered from several diverse marine habitats including other ETNP OMZ samples, low-temperature hydrothermal sediments and plume water, the Gulf of Mexico, and the Delaware and Chesapeake Bay estuaries (Figs. 1, 2 and S2, Table 2). We note that several of these contigs had no or low, category III virus prediction results via VirSorter but had significant VirFinder prediction scores, again highlighting the efficacy of using both tools in tandem to detect novel viral contigs (Table 2). The viral contigs exhibit several distinct genomic architectures (Fig. 2) that broadly belong to two larger groups defined by what type of capsid protein they possess—capsids with distant similarity to caudoviral Haloarchaea virus capsids or to the Pro-Nvie1 capsid (Fig. S2). Members of both capsid types exhibit structural similarity to capsids with HK97-like folds characteristic of the tailed virus order *Caudovirales* [47] (HHPred, *E*-value < 1E–5), suggesting that these thaumarchaeal viruses are tailed. As noted above, contig ETNP_CA_420 encodes a putative baseplate protein (Table 1), also supporting that it likely is a tailed virus. Furthermore, Pro-Nvie1 appears to be a tailed virus [16], and the halovirus-like capsids of the other thaumarchaeal

Table 2 Information about *Thaumarchaeota* viral and cellular contigs that encode *amoC* genes (Fig. 1) or exhibit synteny to *amoC*-encoding viral contigs (Fig. 2)

Species[a]	Contig name	Length (bp)	No. of genes	Has *amoC*?	Source of contig, (JGI IMG/M Genome ID), reference	VirFinder score, *p*-value	VirSorter prediction (I, II, III)[b]	Notes
A	**ETNP_CA_420**	25,994	21	Y	Eastern Tropical North Pacific, this study	0.92, 0.012	III	
A	Ga0098036_1138708	744	2	Y	Marine viral communities from the Subarctic Pacific Ocean - 4_ETSP_OMZ_AT15127 (3300006929)	0.75, 0.06	n.p.	In the JGI/VR viral contig collection
B	**TARA_034_DCM_0.22-1.6_scaffold218395_1**	14,641	17	Y	Tara Oceans Virome contig assemblies [42]	0.91, 0.012	n.p.	GOV_bin_4552_contig-100_2 is a 99.9% identical subfragment of this contig
B	Ga0063241_1021025	3491	6	Y	Northern Gulf of Mexico hypoxic zone (3300003894) [71]	0.76, 0.035	II	
C	**Ga0098048_1000116**	37,657	53	Y	Marine viral communities from the Subarctic Pacific Ocean - 13_ETSP_OMZ_AT15268 (3300006752) [70]	0.76, 0.035	II	In the JGI/VR viral contig collection
C	KVWGV2_10011101	18,034	28	Y	Marine sediment microbial communities from Kolumbo Volcano mats, Greece (3300002242) [36]	0.72, 0.043	II	In the JGI/VR viral contig collection
C	KVRMV2_100116932	12,764	20	Y	Marine sediment microbial communities from Santorini caldera mats, Greece (3300002231) [36]	0.56, 0.074	II	In the JGI/VR viral contig collection
C	Ga0098036_1031672	1659	4	Y	Marine viral communities from the Subarctic Pacific Ocean - 4_ETSP_OMZ_AT15127 (3300006929) [70]	0.77, 0.084	n.p.	In the JGI/VR viral contig collection
C	Ga0080013_1097990	898	2	Y	Marine viral communities from the Pacific Ocean - ETNP_6_85 (3300005862) [33]	0.63, 0.120	n.p.	Assembled by JGI using different pipeline than in
D	**JGI25132-J35274_1000069**	26,317	41	N	Marine viral communities from the Pacific Ocean - ETNP_6_30 (3300002483)[33]	0.88, 0.017	II	ARCHVIR_ETNP_6_30_revised_scaffold28175 (length 12,907 bp) from (18) is a 100% subfragment of this contig, Assembled by JGI using different pipeline than in [33]
E	**Ga0080013_1002070**	24,662	34	N	Marine viral communities from the Pacific Ocean - ETNP_6_85 (3300005862) [33]	0.84, 0.023	II	Assembled by JGI using different pipeline than in [33]
F	**DelMO-Win2010_c10015535**	4132	6	Y	Delaware River and Bay, DEBay_Fall_30_ > 0.8_DNA (3300000117) [70]	0.99, 0.002	II	In the JGI/VR viral contig collection
F	Ga0075463_10007682	3627	6	Y	Delaware Coast, MO Winter December 2010 (3300007236)	0.98, 0.004	n.p.	

Table 2 (continued)

Species[a]	Contig name	Length (bp)	No. of genes	Has *amoC*?	Source of contig, (JGI IMG/M Genome ID), reference	VirFinder score, *p*-value	VirSorter prediction (I, II, III)[b]	Notes
F	Ga0070751_1023884	2887	5	Y	Chesapeake Bay, CPBay_Sum_20_0.8_DNA (3300007640)	0.97, 0.003	I	
G	**Ga0070747_1001249**	12,473	20	Y	Delaware River and Bay, Viral MetaG DEL_Aug_31 (3300007276)	0.75, 0.036	II	
H	**Ga0129342_1000209**	23,286	30	Y	Delaware River and Bay, Viral MetaG DEL_Aug_28 (3300010299)	0.91, 0.012	II	
H	Ga0136656_1000333	16,037	29	Y	Chesapeake Bay, CB_1508_1M Viral MetaG (3300010318)	0.97, 0.005	II	
H	Ga0099849_1003135	7626	11	Y	Chesapeake Bay, CPBay_Sum_15_0.8_DNA (3300007539)	0.76, 0.035	II	
H	Ga0129345_1025350	2308	3	Y	Delaware River and Bay, Viral MetaG DEL_Aug_28 (3300010297)	0.87, 0.024	I	
I	**Ga0070751_1001009**	16,986	27	Y	Delaware River and Bay, Viral MetaG DEL_Mar_4 (3300007640)	0.94, 0.008	II	
J	**Ga0070745_1004049**	7631	13	Y	Delaware River and Bay, Viral MetaG DEL_Aug_28 (3300007344)	0.9, 0.013	II	
J	Ga0070751_1014454	3914	6	Y	Chesapeake Bay, CPBay_Sum_15_0.2_DNA (3300007640)	0.93, 0.01	II	
J	Ga0070753_1023367	2715	4	Y	Delaware River and Bay, Viral MetaG DEL_Mar_31 (3300007346)	0.86, 0.026	I	
K	**Ga0066372_10000192**	21,365	33	N	South Atlantic (3300006902) [70]	0.82, 0.025	III	In the JGI/VR viral contig collection
L	**GBIDBA_10003243**	13,108	14	N	Guaymas Basin Hydrothermal Plume Water (3300001683) [61]	0.89, 0.014	n.p.	Exhibits synteny with ETNP_CA_420, lacks *amoC*
M	**GBIDBA_10004208**	14,033	11	N	Guaymas Basin Hydrothermal Plume Water (3300001683) [61]	0.75, 0.036	n.p.	Exhibits synteny with ETNP_CA_420, lacks *amoC*
N	**GBIDBA_10005806**	13,135	14	N	Guaymas Basin Hydrothermal Plume Water, (3300001683) [61]	0.77, 0.033	III	Exhibits synteny with ETNP_CA_420, lacks amoC
O	**GBIDBA_10128132**	1274	5	N	Guaymas Basin Hydrothermal Plume Water (3300001683) [61]	0.98, 0.001	n.p.	Exhibits synteny with ETNP_CA_420, lacks *amoC*
na	ETNP_CA_109908[c]	1075	3	Y	Eastern Tropical North Pacific, this study	0.68, 0.081	n.p.	The other two genes on this contig have no similarity to any nr sequences
na	ETNP_CA_189861[c]	768	2	Y	Eastern Tropical North Pacific, this study	0.22, 0.45	n.p.	The one other gene on this contig has similarity to a bifunctional tetrahydrofolate synthase/dihydrofolate synthase from Oligella urethralis (Betaproteobacteria)

Table 2 (continued)

Species[a]	Contig name	Length (bp)	No. of genes	Has *amoC*?	Source of contig, (JGI IMG/M Genome ID), reference	VirFinder score, *p*-value	VirSorter prediction (I, II, III)[b]	Notes
na	ETNP_CA_23314[d]	2807	5	Y	Eastern Tropical North Pacific, this study	0.31, 0.275	n.p.	Cellular Thaumarchaeota contig, see Fig. S7
na	ETNP_CA_28663[d]	2468	4	Y	Eastern Tropical North Pacific, this study	0.27, 0.31	n.p.	Cellular Thaumarchaeota contig, see Fig. S7

Contigs listed in bold are the representative contigs from the 15 putative viral species delineated by average nucleotide identity (Fig. S3) that were used for measuring population abundances in various habitats (Fig. 3)

n.p. not predicted as viral

[a]Species are delineated by average nucleotide identity between contigs, ≥ 95% within species (Fig. S7)

[b]Category I = "most confident" predictions; Category II = "likely" predictions, Category III = "possible" predictions

[c]These contigs have viral AMG *amoC* genes (Fig. 1), but the one or two other genes on the contig have no evidence of being viral

[d]For comparison, these contigs are most likely cellular *Thaumarchaeota* contigs based on synteny of other *amo* genes on these contigs to host genomes (see Fig. S7)

viral contigs are most closely related to recently described, tailed *Euryarchaeota* Marine Group II viruses (Magroviruses) (Fig. S2) [48, 49]. More broadly, the phylogeny of Magrovirus, Halovirus, and *Thaumarchaeota* virus capsids point to a shared common ancestor among a larger group of presumably tailed viruses that infect the two archaeal phyla *Euryarchaeota* and *Thaumarchaeota* (Fig. S2). We also note that contig Ga0070747_1001249 from the Chesapeake Bay does not appear to encode a capsid protein. Based on this and its distinct genomic architecture from the other contigs, it could represent an interesting, distinct group of thaumarchaeal viruses.

Similarity between viral sequences found in this study and previously identified, putative thaumarchaeal viral sequences help corroborate that the latter are indeed thaumarchaeal viral sequences. A viral fosmid, Oxic1_7, identified in a fjord and that shares similarity to Pro-Nvie1 sequences [17], also shares similarity to proteins from thaumarchaeal viral contigs found in this study, including the putative portal protein ETNP_CA_420_3 (Table S1). Two proteins from caudoviral contigs recovered from a thaumarchaeal SAG [18] share 30–40% protein identity to thaumarchaeal viral contig sequences (Table S1). This provides additional support that thaumarchaeal viruses identified in this study are probably tailed. Several thaumarchaeal viral contigs carry a gene with similarity to gene NMSP_1228 previously identified as a putative viral gene in the genome of the cultured thaumarchaeon *Ca. N. catalina* SPOT01 [19] (Table S1). Although contig JGI25132J35274_1000069 assembled from a virome at 30 m in the ETNP lacks an *amoC* gene, it also likely represents a thaumarchaeal virus based on synteny to other *amoC*-encoding contigs (Fig. 2) and capsid phylogeny (Fig. S2). Note that ETNP_6_30_revised_scaffold28175_1 from [33], which was previously identified as a putative archaeal contig, is an identical subfragment of JGI25132J35274_1000069 that was obtained using the same raw data with a different assembly pipeline. Finally, host prediction using the *k*-mer similarity tool VirHostMatcher, independently predicted *Thaumarchaeota* as the probable host phylum for all but two of these viral contigs that are >10 kb in length ($n = 17$), including four contigs that lack *amoC* genes (Table S2).

Using a nucleotide identity cutoff of 95% to delineate putative viral species [50], the thaumarchaeal virus contigs represent at least 15 distinct viral species (Fig. S3). These populations exhibit different abundance patterns across various marine habitats and samples, suggesting they are ecologically distinct (Fig. 3). Abundances were determined by metagenomic mapping of reads to one representative contig from each of the species (Fig. S3, Table S1). In two ETNP virome depth profiles from another study [33], our archaeal virus populations exhibit evidence of depth partitioning (Fig. 3a). There was also a

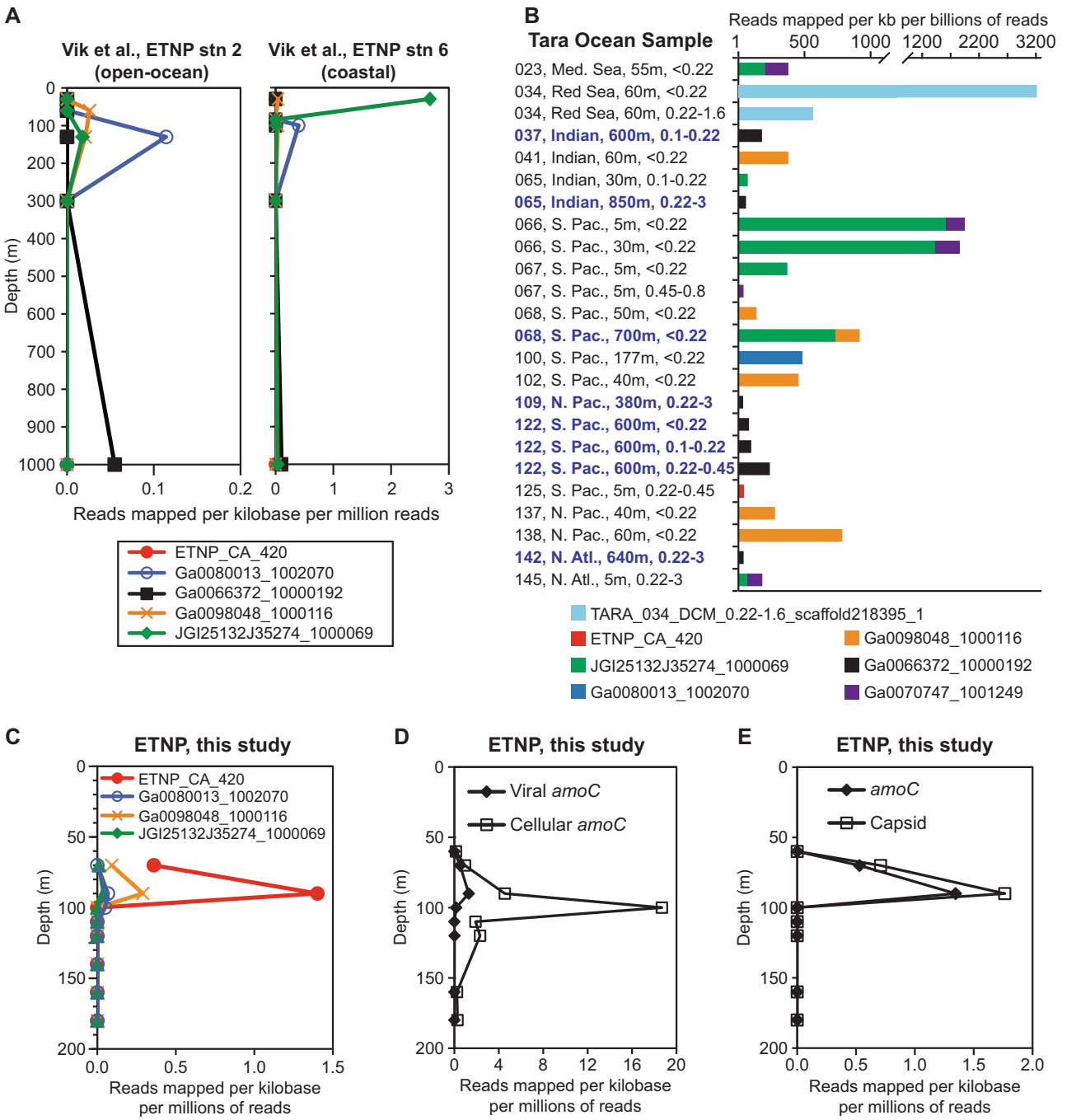

Fig. 3 Abundance of representative *Thaumarchaeota* viral contigs and gene sequences in various marine samples as determined by metagenomic read mapping to **a** Coastal (station [stn] 2) and open-ocean (stn 6) ETNP samples from Vik et al. [33]. **b** Globally distributed Tara Ocean samples and **c-e** ETNP depth profiles from which contig ETNP_CA_420 was identified. For **a**–**c**, metagenomic reads were mapped to representative contigs from the 15 viral species. Normalized read recruitment is depicted as the number of reads mapped per kilobase of the contig per billions of reads in the sample. Contigs that did not meet the mapping criteria in any samples are not depicted in the graphs. For **a** and **c**, recruitment values are only plotted if ≥ 30 reads were mapped to the respective contig. For Tara Ocean samples in **b** recruitment levels are only plotted if contig coverage was ≥1 or the percentage of the contig covered was ≥75%. The following is listed for each Tara sample on the vertical axis: Tara Ocean site number, oceanic region, sampling depth, and size fraction of the sample in µm. Mesopelagic samples are in bold and blue to highlight that contig Ga0066372_10000192 was only detected in mesopelagic samples. For **d**, read recruitment results are shown for reads that mapped to sequences from the viral *amoC* AMG ("Viral *amoC*") or cellular *Thaumarchaeota* ("Cellular *amoC*") clades as defined in Fig. 1. For **e**, recruitment results are shown for the *amoC* and capsid genes on contig ETNP_CA_420

marked shift in dominance between the coastal and open-ocean ETNP sites from Vik et al. (Fig. 3a) and the open-ocean ETNP site from which ETNP_CA_420 was identified (Fig. 3c). Marine *Thaumarchaeota* belong to two major phylogenetic groups ("Deep" [or Water Column B] and "Shallow" [or Water Column A]) according to whether they predominantly occur in the upper water column (<200 m) or deeper mesopelagic depths (>200 m) [5, 19, 51–53]. Contig Ga0066372_10000192 notably dominated at 1000 m in the Vik et al. ETNP Station 2 sample; was absent in surface waters; and thus may specifically infect deep *Thaumarchaeota* populations. Contig Ga0066372_10000192 was also only observed in deep, mesopelagic Tara Ocean metagenomes (Fig. 3b, Fig. S4), and only this contig was significantly correlated to sample depth, as tested with Spearman correlations ($p = 2E{-}8$, $\rho = 0.84$). Contig Ga0066372_10000192 exhibited significantly higher nucleotide similarity to Deep *Thaumarchaeota* genomes than Shallow genomes using VirHostMatcher (*t*-test, $p < 0.05$, Table S2), further supporting that this thaumarchaeal viral population infects Deep group *Thaumarchaeota*. Contig Ga0070751_1001009 likewise has higher *k*-mer similarity to Deep group genomes and likely infects these hosts, but corresponding depth partitioning of this contig to mesopelagic samples could not be corroborated by metagenomic data as it was not significantly detected in any Tara Ocean samples. Most of the other remaining thaumarchaeal contigs that generally were more abundant in epipelagic waters (Fig. 3, S4) had significantly higher nucleotide similarity to Shallow *Thaumarchaeota* genomes (*t*-test, $p < 0.05$, Table S2).

We also used Tara Ocean metagenomes to explore other differences in the distribution of the thaumarchaeal virus populations and thus possible differences in their ecologies. Representative thaumarchaeal virus contigs were detected in 24 samples collected from several ocean basins; at surface, deep chlorophyll maximum, and mesopelagic depths; and in both cellular and viral fraction samples (Fig. 3b). Although the limited number of samples in which these viruses were detected made it difficult to make conclusive inferences about their potential ecological differences, recruitment data allowed for some preliminary insight into their probable niches, in addition to the depth partitioning described above. The underlying assumption here is that these virus populations infect particular host populations such that viral niches reflect the particular niches of their specific hosts. We noticed that contigs Ga00707474_1001249 and JGI25132135274_1000069 appeared to preferentially be found at low latitude sites. Only these two contigs exhibited significant Spearman correlations to latitude ($p < 0.01$, $\rho = 0.54$ and 0.68, respectively), suggesting that they may specifically infect *Thaumarchaeota* found in warmer, low latitude waters. The *Thaumarchaeota* Shallow group does contain several clades, which based on culture studies of a limited number of representative strains, hint that they may occupy distinct niches determined by temperature [19, 54, 55]. VirHostMatcher could potentially be used to test the hypothesis that these two viral populations infect warm-adapted hosts, but the biogeographic ranges of Shallow group clades need to be better constrained and more representative genomes from these clades are needed. We also observed that at a single sampling site and depth (TARA_067, 5 m, Benguela Current), a different population dominated the viral (<0.22 µm) and cellular fractions (0.45–0.8 µm), possibly reflecting the detection of a transition from a recent infection by one viral population, observed as sequences in the viral fraction, to an active infection by another population, observed as sequences in the cellular fraction (Fig. 3b).

We observed that none of the contigs identified from the Delaware and Chesapeake Bay estuaries (Ga0070747_1001249, Ga0129342_1000209, Ga0070751_1001009, DelMOWin2010_c10015535, Ga0070745_1004049) or from hydrothermal plume water (GBIDBA_10003243, GBIDBA_10004208, GBIDBA_10005806, and GBIDBA_10128132) were detected in any of the pelagic Tara Ocean samples. This is perhaps not surprising if these viral populations specifically infect *Thaumarchaeota* hosts adapted to estuarine and vent plume habitats, because the Tara Ocean samples did not sample such habitats. On the other hand, contig Ga0098036_10316722 identified from pelagic waters of the South Pacific was detected in several Tara Ocean samples (Fig. 3b, S4) and is closely related to contigs recovered from low-temperature hydrothermal mat samples (Figs. 1, 2, Fig. S3), suggesting that these viruses perhaps infect hosts that occupy these two disparate habitats. This is rather unexpected given a model that viruses generally infect specific hosts and the expectation that pelagic and hydrothermal mat *Thaumarchaeota* are probably not closely related. The latter assumption may not be true as low-temperature hydrothermal mats at the Kolumbo Volcano site do support abundant *Thaumarchaeota* that are closely related (99% identity by 16S rRNA) to pelagic strain *Nitrosopumilus maritimus* SCM1 [56]; however, 16S rRNA can fail to resolve closely related *Thaumarchaeota* clades that probably are ecologically distinct [19]. There are examples of some marine cyanophage that have somewhat broad host ranges and infect multiple genera, *Prochlorococcus* and *Synechococcus*, [57] but both of these genera occupy broadly similar niches—both are pelagic and mesophilic. The curious observations of these *Thaumarchaeota* viruses require further investigation.

The predicted AmoC protein sequences from viral contigs have high amino-acid identity (>90%) to marine

Thaumarchaeota AmoC sequences. Although it is difficult to establish the functionality of proteins by sequence analysis alone, the high degree of sequence similarity to cellular *Thaumarchaeota* proteins suggests that these AmoC proteins are functional. There was no significant difference in protein length between viral and host AmoC sequences (*t*-test, $p < 0.05$), nor was there any clear difference in particular amino-acid motifs used by each group. Several non-marine (soil and freshwater) *Thaumarchaeota* genomes of the genus *Candidatus Nitrososphaera* possess several copies of *amoC*, but copies of this gene from non-marine genomes form a separate lineage from marine host and viral sequences (Figs. S5, S6).

amoC nucleotide sequences from viral contigs, however, are quite dissimilar to marine thaumarchaeal *amoC* sequences (<80% identity) (Fig. S6). Note that we also recovered host contigs whose *amoC* sequences cluster with other host sequences and which contain genes for the other *amo* subunits (Fig. S7). The *amoC* sequences from viral contigs form a distinct phylogenetic group that we argue represents viral *amoC* AMGs (Fig. 1, S6). This is consistent with cyanophage photosystem *psbA* sequences that also form phylogenetically distinct clades from host versions of the gene [58]. This phylogenetic distinction made it possible for us to discriminate viral and cellular *amoC* sequences in metagenomic data. In Tara Ocean metagenomes, the ratio of viral *amoC* AMG to cellular sequences was significantly higher in viral fraction samples (<0.22 µm) than cellular fraction samples (average: 0.7 vs. 0.1, $p < 0.05$), supporting that *amoC* AMG clade sequences are associated with viral-sized particles. Furthermore, we found a handful of metatranscriptomic reads from the Gulf of Mexico [35] and coastal Georgia, USA (Sapelo Island) [34] that matched viral *amoC* and capsid genes with ≥90% identity, demonstrating active RNA expression of these viral genes.

Analysis of cellular fraction (>0.22 µm) ETNP metagenomic reads showed that viral *amoC* sequences surprisingly comprised 54 and 29% of total *amoC* reads at 70 and 90 m, respectively (Fig. 3d). The high abundance of viral *amoC* genes supports two non-mutually exclusive scenarios: (1) these samples have captured an active infection of *Thaumarchaeota*, whereby replicating viruses within cells were recovered from >0.22 µm fraction metagenomes or (2) the ETNP_CA_420 contig represents a highly prevalent integrated provirus. In either case, it represents a highly successful virus in these samples. It has been similarly reported that viral versions of the *psbA* gene can comprise a large portion (up to 60%) of total *psbA* abundance in natural communities [59]. The fraction of viral to total *amoC* genes was more modest in the hydrothermal vent mat samples from which contigs KVWGV2_10011101 and KVRMV2_100116932 were assembled respectively (Santorini Caldera: 10% and Kolumbo Volcano: 20%). Viral *amoC* genes were detected in 20% of Tara Ocean viromes (at least 10 reads mapped, ≥95% nucleotide identity). Viral *amoC* genes, however, were only detected in four >0.22 µm, prokaryotic fraction Tara Oceans samples (3% of samples for which any *amoC* gene was detected), and viral *amoC* genes never exceeded 2.2% of total *amoC* DNA abundance (range: 0.12–2.2%). The comparable levels of recruitment for ETNP_CA_420 capsid and *amoC* genes measured for ETNP samples also implies that most (~75%) thaumarchaeal viruses carried an *amoC* gene at that location (Fig. 3e). Contigs JGI25132J35274_1000069 and Ga0066372_10000192 notably lack *amoC* genes. These contigs are otherwise syntenous with *amoC*-encoding viral contigs and have a few genes with similarity to *Thaumarchaeota*, supporting that they too represent thaumarchaeal viruses. These contigs highlight along with the recruitment data (Fig. 3e) that not all thaumarchaeal viruses appear to carry *amoC* genes.

The discovery of widespread *amoC*-encoding viral sequences and in some cases high viral *amoC* abundances in metagenomic samples has potential implications for our understanding of N cycling in marine systems, to which *Thaumarchaeota* are major contributors. AMGs have been previously described for several enzymes involved in the biochemical cycles for most of the major elements comprising life including C, P, and S [23, 60–62]. Notably missing from this list were prominent N-related AMGs, but the discovery the nitrogen regulators *ntcA* in cyanophage [63]; P_{II} and ammonia transporters (*amt*) on viral contigs [21]; *amoC* AMGs (in [21] and here); and more recently nitrate reductase genes in deep-sea vent viromes [64], have expanded the known set of key biogeochemical pathways impacted by viral AMGs. Viral infection should generally limit the abundance and thus contribution of *Thaumarchaeota* to nitrification, but cells infected by *amoC*-carrying viruses presumably still can contribute to nitrification. Analogous *psbA* AMGs in marine cyanophage are thought to contribute to the photosynthetic functioning of infected, natural populations [24]. Further work is needed to assess what fraction of *Thaumarchaeota* are infected by viruses at any given time, what fraction of viruses encode *amoC*, and to what degree cells infected with *amoC*-carrying viruses have enhanced contribution to N cycling. Our initial metagenomic survey suggests that viral *amoC* genes are distributed globally (Table 2, Fig. 3, S4), and at certain key sites of nitrification, they may be very abundant and could have an important impact on N cycling. Directed *amoC*-specific assays (e.g., amplicon sequencing) may be better equipped to assess *amoC* contribution than metagenome approaches that often yield low coverage results for individual genes. These potential biogeochemical implications may also extend to marine sediments, another important location of nitrification, because *amoC* AMGs were recovered from sediment samples as well.

The fact that thaumarchaeal viruses carry the gene for the AmoC subunit rather the other two subunits (A and B) gives potential insight into the biochemistry of ammonia monooxygenase. It is surmised that cyanobacterial viruses carry the *psbA* AMG encoding the D1 protein of the photosystem complex because this is the subunit most susceptible to damage and has a high turnover rate [59, 65]. We therefore hypothesize that AmoC likewise has a high turnover rate, such that expression of AmoC provides these viruses a selective advantage in maintaining cellular energy during infection.

The discovery of these novel thaumarchaeal virus sequences adds to the small but growing dataset of *Thaumarchaeota* virus sequences identified by isolation-independent approaches [17, 18, 21]. Because of the paucity of cultured archaeal viruses and particularly for marine archaea (see Introduction), these contigs help enlarge the known diversity of archaeal viruses and elucidate broader evolutionary relationships of viruses infecting *Euryarchaeota* and *Thaumarchaeota* (Fig. S2), two prominent phyla of archaea in the oceans. These new sequences will also assist in the discovery of new archaeal viruses. Case in point is the discovery of potential viral genes in Asgard superphylum archaeal sequences. This study also highlights the usefulness of analyzing cellular fraction metagenomes, not just viromes, for discovery new viral sequences, especially with availability of robust tools like VirFinder and VirSorter to distinguish viral sequences from prokaryotic host sequences.

Acknowledgements We thank Cameron Thrash, Barbara Campbell, and Feng Chen for permission to include in our analysis select contigs from viral metagenomes sequenced at the Joint Genome Institute for which they are principal investigators. This work was supported by funding from The Gordon and Betty Moore Foundation Marine Microbiology Initiative (GBMF3779) to JAF; from the National Institutes of Health to NAA and JAF (R01 GM120624-01A1); and from the National Science Foundation OCE-1138368 to GR.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Könneke M, Bernhard AE, De La Torre JR, Walker CB, Waterbury JB, Stahl DA. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature. 2005;437:543–6.
2. Könneke M, Schubert DM, Brown PC, Hugler M, Standfest S, Schwander T, et al. Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. Proc Natl Acad Sci USA. 2014;111:8239–44.
3. Walker CB, De La Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, et al. *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proc Natl Acad Sci USA. 2010;107:8818–23.
4. Santoro AE, Casciotti KL, Francis CA. Activity, abundance and diversity of nitrifying archaea and bacteria in the central California Current. Environ Microbiol. 2010;12:1989–2006.
5. Beman JM, Popp BN, Francis CA. Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. ISME J. 2008;2:429–41.
6. Karner MB, Delong EF, Karl DM. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature. 2001;409:507–10.
7. Teira E, Lebaron P, Van Aken H, Herndl GJ. Distribution and activity of Bacteria and Archaea in the deep water masses of the North Atlantic. Limnol Oceanogr. 2006;51:2131–44.
8. Santoro AE, Saito MA, Goepfert TJ, Lamborg CH, Dupont CL, Ditullio GR. Thaumarchaeal ecotype distributions across the equatorial Pacific Ocean and their potential roles in nitrification and sinking flux attenuation. Limnol Oceanogr. 2017;62:1984–2003.
9. Sintes E, Bergauer K, De Corte D, Yokokawa T, Herndl GJ. Archaeal amoA gene diversity points to distinct biogeography of ammonia-oxidizing Crenarchaeota in the ocean. Environ Microbiol. 2013;15:1647–58.
10. Geslin C, Le Romancer M, Erauso G, Gaillard M, Perrot G, Prieur D. PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, "*Pyrococcus abyssi*". J Bacteriol. 2003a;185:3888–94.
11. Geslin C, Le Romancer M, Gaillard M, Erauso G, Prieur D. Observation of virus-like particles in high temperature enrichment cultures from deep-sea hydrothermal vents. Res Microbiol. 2003b;154:303–7.
12. Pietila MK, Demina TA, Atanasova NS, Oksanen HM, Bamford DH. Archaeal viruses and bacteriophages: comparisons and contrasts. Trends Microbiol. 2014;22:334–44.
13. Prangishvili D. The wonderful world of archaeal viruses. Annu Rev Microbiol. 2013;67:565–85.
14. Prangishvili D, Forterre P, Garrett RA. Viruses of the archaea: a unifying view. Nat Rev Microbiol. 2006;4:837–48.
15. Hurwitz BL, Ponsero A, Thornton J Jr., U'ren JM. Phage hunters: computational strategies for finding phages in large-scale 'omics datasets. Virus Res. 2017;244:110–5.
16. Krupovic M, Spang A, Gribaido S, Forterre P, Schleper C. A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. Biochem Soc Trans. 2011;39:82–8.
17. Chow CET, Winget DM, White RA, Hallam SJ, Suttle CA. Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. Front Microbiol. 2015;6:1–15.
18. Labonte JM, Swan BK, Poulos B, Luo HW, Koren S, Hallam SJ, et al. Single-cell genomics-based analysis of virus-host interactions in marine surface bacterioplankton. ISME J. 2015;9:2386–99.
19. Ahlgren NA, Chen Y, Needham DM, Parada AE, Sachdeva R, Trinh V, et al. Genome and epigenome of a novel marine *Thaumarchaeota* strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems. Environ Microbiol. 2017a;19:2434–52.
20. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Res. 2017b;45:39–53.
21. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature. 2016;537:689–93.
22. Stahl DA, De La Torre JR. Physiology and diversity of ammonia-oxidizing archaea. Annu Rev Microbiol. 2012;66:83–101.
23. Hurwitz BL, U'ren JM. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol. 2016;31:161–8.

24. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. Photosynthesis genes in marine viruses yield proteins during host infection. Nature. 2005;438:86–9.
25. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, et al. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci USA. 2011;108:E757–64.
26. Edwards RA, Mcnair K, Faust K, Raes J, Dutilh BE. Computational approaches to predict bacteriophage-host relationships. FEMS Microbiol Rev. 2015;40:258–72.
27. Roux S, Enault F, Hurwitz BL.Sullivan MB, VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015a;3:e985.
28. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. eLife. 2015b;4:e08490.
29. Fuchsman CA, Devol AH, Saunders JK, Mckay C, Rocap G. Niche partitioning of the N cycling microbial community of an offshore oxygen deficient zone. Front Microbiol. 2017;8:2384.
30. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 2010;11:119.
31. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics. 2005;21:951–60.
32. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Patterns and ecological drivers of ocean viral communities. Science. 2015;348:1261498–91.
33. Vik DR, Roux S, Brum JR, Bolduc B, Emerson JB, Padilla CC, et al. Putative archaeal viruses from the mesopelagic ocean. PeerJ. 2017;5:e3428.
34. Hollibaugh JT, Gifford SM, Moran MA, Ross MJ, Sharma S, Tolar BB. Seasonal variation in the metatranscriptomes of a Thaumarchaeota population from SE USA coastal waters. ISME J. 2014;8:685–98.
35. Thrash JC, Seitz KW, Baker BJ, Temperton B, Gillies LE, Rabalais NN, et al. Metabolic roles of uncultivated bacterioplankton lineages in the Northern Gulf of Mexico Dead Zone. mBio. 2017;8:e01017–17.
36. Oulas A, Polymenakou PN, Seshadri R, Tripp HJ, Mandalakis M, Paez-Espino AD, et al. Metagenomic investigation of the geologically unique Hellenic Volcanic Arc reveals a distinctive ecosystem with unexpected physiology. Environ Microbiol. 2016;18:1122–36.
37. Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics. 1998;14:817–8.
38. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533–42.
39. Tully BJ, Sachdeva R, Graham ED, Heidelberg JF. 290 metagenome-assembled genomes from the Mediterranean Sea: a resource for marine microbiology. PeerJ. 2017;5:e3558.
40. Peng X, Fuchsman CA, Jayakumar A, Oleynik S, Martens-Habbena W, Devol AH, et al. Ammonia and nitrite oxidation in the Eastern Tropical North Pacific. Glob Biogeochem Cycles. 2015;29:2034–49.
41. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5:69.
42. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348:1261351–59.
43. Ignacio-Espinoza JC, Sullivan MB. Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes. Environ Microbiol. 2012;14:2113–26.
44. Heinhorst S, Cannon GC, Shively JM. Carboxysomes and carboxysome-like inclusions. Complex intracellular structures in prokaryotes. Berlin, Heidelberg, New York, NY: Springer; 2006. p. 141–66.
45. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Backstrom D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017;541:353–8.
46. Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, et al. Community-wide analysis of microbial genome sequence signatures. Genome Biol. 2009;10:R85.
47. Krupovic M, Koonin EV. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci USA. 2017;114:E2401–10.
48. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M, Rodriguez-Valera F. Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters. PLoS Genet. 2017;13:e1007018.
49. Philosof A, Yutin N, Flores-Uribe J, Sharon I, Koonin EV, Beja O. Novel abundant oceanic viruses of uncultured Marine Group II *Euryarchaeota*. Curr Biol. 2017;27:1362–8.
50. Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, et al. Viral tagging reveals discrete populations in *Synechococcus* viral genome sequence space. Nature. 2014;513:242–5.
51. Francis CA, Roberts KJ, Beman JM, Santoro AE, Oakley BB. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc Natl Acad Sci USA. 2005;102:14683–88.
52. Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM, et al. Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine *Crenarchaeota*. PLoS Biol. 2006;4:520–36.
53. Luo HW, Tolar BB, Swan BK, Zhang CLL, Stepanauskas R, Moran MA, et al. Single-cell genomics shedding light on marine Thaumarchaeota diversification. ISME J. 2014;8:732–6.
54. Bayer B, Vojvoda J, Offre P, Alves RJE, Elisabeth NH, Garcia JaL, et al. Physiological and genomic characterization of two novel marine thaumarchaeal strains indicates niche differentiation. ISME J. 2016;10:1051–63.
55. Qin W, Amin SA, Martens-Habbena W, Walker CB, Urakawa H, Devol AH, et al. Marine ammonia-oxidizing archaeal isolates display obligate mixotrophy and wide ecotypic variation. Proc Natl Acad Sci USA. 2014;111:12504–9.
56. Kilias SP, Nomikou P, Papanikolaou D, Polymenakou PN, Godelitsas A, Argyraki A, et al. New insights into hydrothermal vent processes in the unique shallow-submarine arc-volcano, Kolumbo (Santorini), Greece. Sci Rep. 2013;3:2421.
57. Sullivan MB, Waterbury JB, Chisholm SW. Cyanophages infecting the oceanic cyanobacterium *Prochlorococcus*. Nature. 2003;424:1047–51.
58. Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. 2006;4:e234.
59. Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, Rusch DB, et al. Viral photosynthetic reaction center genes and transcripts in the marine environment. ISME J. 2007;1:492–501.
60. Hurwitz BL, Hallam SJ, Sullivan MB. Metabolic reprogramming by viruses in the sunlit and dark ocean. Genome Biol. 2013;14:R123.
61. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. Sulfur oxidation genes in diverse deep-dea viruses. Science. 2014;344:757–60.
62. Roux S, Hawley AK, Beltran MT, Scofield M, Schwientek P, Stepanauskas R, et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and metagenomics. eLife. 2014;3:e03125.

63. Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, et al. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environ Microbiol. 2010;12:3035–56.
64. He T, Li H, Zhang X. Deep-Sea hydrothermal vent viruses compensate for microbial metabolism in virus-host interactions. mBio. 2017;8:e00893–17.
65. Mulo P, Sicora C, Aro EM. Cyanobacterial *psbA* gene family: optimization of oxygenic photosynthesis. Cell Mol Life Sci. 2009;66:3697–10.
66. Peng y, Leung HCM, Yiu SM, Chin FYL IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 2012;28:1420–1428.
67. Jaclyn K Saunders, Gabrielle Rocap, Genomic potential for arsenic efflux and methylation varies among global Prochlorococcus populations. The ISME Journal 2016;10:197–209.
68. Alexandros Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014;30:1312–1313.
69. Simon A. Berger, Alexandros Stamatakis, Aligning short reads to reference alignments and trees. Bioinformatics 2011;27:2068–2075.
70. David Paez-Espino, I.-Min A. Chen, Krishna Palaniappan, Anna Ratner, Ken Chu, Ernest Szeto, Manoj Pillay, Jinghua Huang, Victor M. Markowitz, Torben Nielsen, Marcel Huntemann, T. B. K. Reddy, Georgios A. Pavlopoulos, Matthew B. Sullivan, Barbara J. Campbell, Feng Chen, Katherine McMahon, Steve J. Hallam, Vincent Denef, Ricardo Cavicchioli, Sean M. Caffrey, Wolfgang R. Streit, John Webster, Kim M. Handley, Ghasem H. Salekdeh, Nicolas Tsesmetzis, Joao C. Setubal, Phillip B. Pope, Wen-Tso Liu, Adam R. Rivers, Natalia N. Ivanova, Nikos C. Kyrpides, IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. Nucleic Acids Research 2017;45:D457–D465.
71. Thrash JC, Baker BJ, Seitz KW, Temperton B, Campbell LG, Rabalais NN, Henrissat B, Mason OU. Metagenomic assembly and prokaryotic metagenome-assembled genome sequences from the Northern Gulf of Mexico "Dead Zone". Microbial Research Announcements. 2018;7:9.