environmental microbiology

Environmental Microbiology (2018) **00**(00), 00–00 doi:10.1111/1462-2920.14462

Novel *Caudovirales* associated with Marine Group I Thaumarchaeota assembled from metagenomes

Mario López-Pérez,[1†*] **Jose M. Haro-Moreno**,[1†] **José R. de la Torre**[2] **and Francisco Rodriguez-Valera**[1]
[1]*Evolutionary Genomics Group, División de Microbiología, Universidad Miguel Hernández, Apartado 18, San Juan, Alicante, 03550, Spain.*
[2]*Department of Biology, San Francisco State University, San Francisco, CA, 94132, USA.*

Summary

Marine Group I (MGI) Thaumarchaeota are some of the most abundant microorganisms in the deep ocean and responsible for much of the ammonia oxidation occurring in this environment. In this work, we present 35 sequences assembled from metagenomic samples of the first uncultivated *Caudovirales* viruses associated with Thaumarchaeota, which we designated marthavirus. Most of the sequences were obtained from cellular metagenomes confirming that they represent an important tool to study environmental viral communities due to cells retrieved while undergoing viral lysis. Metagenomic recruitment showed that this viral population is formed by very divergent entities with high intrapopulation homogeneity. However, metatranscriptomic analyses revealed the same differential expression profile with the capsid as major transcript, indicative of viruses during the lytic cycle. The cobalamine biosynthesis gene *cob*S, an auxiliary metabolic gene, was also highly expressed during the infection. These analyses expand our understanding of the global diversity of archaeal viruses.

Introduction

Marine Thaumarchaeota, initially discovered 26 years ago by 16S rRNA gene surveys (Fuhrman and McCallum, 1992; Delong, 1992), are some of the most abundant microorganisms in the deep ocean, accounting for up to 40% of the bacterioplankton below the euphotic zone (Karner et al., 2001; Fuhrman and Ouverney, 1998; Church et al., 2003). This abundance and the discovery that members of this lineage derive energy from the oxidation of ammonia (Könneke et al., 2005) and are able to fix inorganic forms of carbon (Berg et al., 2010) argue that the marine Thaumarchaeota are important players in global Carbon (C) and Nitrogen (N) biogeochemical cycles. Recent studies have shown that these marine archaea are responsible for the majority of the aerobic nitrification measured in marine environments and may be a significant source of the greenhouse gas nitrous oxide (Santoro et al., 2011). The cultivation of numerous strains, as well as sequences from environmental metagenomes and single-cell genomes have provided invaluable information on the ecology and evolution of this diverse lineage (Luo et al., 2014; Swan et al., 2014; Santoro et al., 2015). However, despite these efforts, remarkably little is known about their viruses. The vast majority of archaeal viruses that have been isolated so far came from either hyperthermophilic or hyperhalophilic environments, where Crenarchaeota or Euryarchaeota dominate (Snyder et al., 2015). Conversely, in the case of the mesophilic archaea, only the advent of high-throughput sequencing has provided novel information of the unknown viruses infecting archaea (Vik et al., 2017; Roux et al., 2016), expanding viral diversity far beyond that established by traditional methods for virus isolation. The recently discovered Marine Group II Euryarchaeota viruses (magrovirus) group from assembled sequences (Philosof et al., 2017), which infects the ubiquitous and abundant but yet uncultured Marine Group II Euryarchaeota, is an example of the benefits of metagenomics. However, to date no marine thaumarchaeal virus has yet been isolated probably because their host are also difficult to obtain in pure culture. Only two putative viral sequences have been retrieved by single-cell genomics (Labonté et al., 2015) and fosmid libraries (Chow et al., 2015). Additionally, a putative provirus has been found within the genome of *Ca.* Nitrosomarinus catalina SPOT01 (Ahlgren et al., 2017). Previous studies showed that viruses infecting Thaumarchaeota from the deep ocean were more active than bacterial viruses, contributing to their cell lysis and, hence, modifying the biogeochemical cycles of N and C (Danovaro et al., 2016).

Received 19 September, 2018; revised 23 October, 2018; accepted 24 October, 2018. *For correspondence. E-mail mario.lopezp@umh.es; Tel. +34-965919313; Fax +34-965 919457. †These authors contributed equally to this work.

In this work, we present 35 sequences assembled from metagenomic samples of the first uncultivated viruses associated with marine Thaumarchaeota. Sequences of this newly identified viral population are locally distributed with low intra-population diversity. Metatranscriptomic analyses showed that they were retrieved while undergoing viral lysis and apart from the capsid, the essential structural component, expression of the cobalamine biosynthesis gene *cob*S, an auxiliary metabolic gene, was also high during the infection. Our analyses provide important insights into the genomic diversity of this new marine viral population, which remain uncultivated, expanding our understanding of the global diversity of archaeal viruses.

Results and discussion

Recently, we characterized variations in the marine microbiome at different depths within the photic zone during a period of strong thermal stratification of the water column (Haro-Moreno et al., 2018). Results showed that marine Thaumarchaeota were only found below the deep chlorophyll maximum (accounted for up to 10% of the community at 90 m), coinciding with the increase of available ammonia that is practically non-existent at shallower depths. Analyses of the assembled contigs from these samples showed the presence of a 69 kb contig that had hits to a few Thaumarchaeota genomes (the majority of these hits were to the genus *Ca.* Nitrosopumilus) but also to viral-related genes, including predicted major capsid proteins (MCP), portal proteins, tail tape measure proteins and the large subunit of viral terminases. Both the terminase and the MCP proteins gave hits with low identity (32–35%) to a complete, unclassified archaeal virus (KY229235) recovered from a metagenomic assembly of a sample ~550 m below the seafloor (Nigro et al., 2017). Like in the case of KY229235, our contig had identical repeated sequences (>30 nucleotides) at the 5′ and 3′ terminal regions suggesting a complete viral genome. It has been demonstrated that cellular metagenomes (>0.22 μm size fraction) are a source of bacterial and archaeal viruses that are undergoing the lytic cycle and actively replicating their DNA (López-Pérez et al., 2017).

New Thaumarchaeota viruses recovered from metagenomic samples

In order to expand the repertoire of putative Thaumarchaeota viruses we used the MCP, terminase and portal proteins of this new contig and KY229235 sequences as queries to search against several marine metagenomes and viromes, including the Mediterranean Sea dataset (Haro-Moreno et al., 2018; López-Pérez et al., 2017), *Tara* Oceans (Sunagawa et al., 2015) and Malaspina expeditions (Duarte, 2015) and datasets publicly available at the Joint Genome Institute (JGI) database (https://img.jgi.doe.gov/). In the end, we identified 35 putative viral contigs (Supporting Information Table S1) manually curated to check for similarity to these reference proteins and thaumarchaeal genes (see Material and Methods). Most contigs with similarity to these proteins (#18) were found in our own dataset (Med-OCT2015-75m and Med-OCT2015-90m), 17 in the cellular and one in the viral fraction (MedVir-OCT2015-60m). Interestingly, another large batch (9 genomes) was found in viromes from the Chesapeake Bay, an estuary where strong ammonia gradients are also found (Maresca et al., 2018). All had a GC content varying from 30% to 37%, as expected from the low-GC content of marine Thaumarchaeota genomes (Ahlgren et al., 2017; Supporting Information Table S1). Based on the presence of terminal inverted repeats >30 nucleotides only two viral genomes were complete. A total of 1289 open reading frames were identified in all the sequences. However, only 14% showed significant homology to sequences present in the pVOGs (Prokaryotic Virus Orthologous Groups) database (Grazziotin et al., 2017), as typical for novel viruses. Clustering of the sequences based on similarity resulted in 684 protein clusters, 11 of which formed the viral 'soft' core (they were present in at least half of the sequences) (Supporting Information Table S2). Five of the 'soft' core protein clusters contained proteins involved in DNA metabolism (terminase, RadA, ATPase, PD-D/EXK nuclease and Ribonuclease H) and one sequence was an auxiliary metabolic gene (AMG), *cob*S, that encodes a protein that catalyses the final step in cobalamin (vitamin B_{12}) biosynthesis in prokaryotes, which has previously only been found in cyanophages (Sullivan et al., 2005). Unfortunately, the remaining five 'soft' core clusters were hypothetical proteins and no function could be inferred. Furthermore, no tRNA-encoding sequences or hallmarks of temperate phage, such as integrase or excisionase genes, were detected in any of the recovered genomes. In addition to the terminase, we found other clusters indicating the *Caudovirales* affiliation of these viruses such as prohead or portal proteins. This is to our knowledge the first group of head-tail viruses described for the Crenarchaeal superphylum, although they are relatively common in Euryarchaea (Rachel et al., 2002; Pietilä et al., 2014).

Phylogeny and host assignment

We next sought to establish the phylogenetic affiliation of these sequences and their relationship with other archaeal virus sequences. We used two characteristic *Caudovirales* marker genes, the terminase and the MCP. Homologous (although less than 30% nucleotide identity)

terminases and MCPs were found in the fosmid Oxic1_7 (Chow et al., 2015) and the provirus Nvie-Pro1 present in the genome of *Nitrososphaera viennensis*, a soil Thaumarchaeota (Krupovic et al., 2011), but could not be identified in the putative thaumarchaeal virus found in the single-cell genome AAA160-J20 (Labonté et al., 2015) or in the putative provirus of *Ca.* Nitrosopumilus catalina SPOT01 (Ahlgren et al., 2017). Both Nvie-Pro1 and Oxic1_7 encoded a multifunctional MCP fused with a protease sequence. This particular fusion between the two domains has not been seen in any of our recovered virus genomes. For the phylogenetic analyses, we selected only the MCP domain of the mentioned reference sequences. Results showed a similar phylogenetic pattern for both terminase and MCPs, where the new sequences identified here formed a separate lineage from haloviruses and magroviruses (Fig. 1A and B). Only the viral genome KY229235 was found close to our sequences, while Oxic1_7 and Nvie-Pro1, which clustered together, were found more closely related to haloviruses. Additionally, we carried out phylogenetic analyses of the viral RadA and PD-D/EXK nuclease genes (both of which are also present in the genomes of Thaumarchaeota cells) and appeared to be strongly associated with the Thaumarchaeota, and distinct from Euryarchaeota group II and their viruses (Fig. 1C and D). Our results confirm the association of these novel viruses to the marine archaeal phylum. Neither Oxic1_7 or Nvie-Pro1 encoded these genes in their sequence. Remarkably, all 35 putative viral sequences clustered as a single, monophyletic lineage in all four phylogenetic trees, indicating that they are a novel clade of marine archaea-infecting *Caudovirales*, evolutionarily distinct from previous putative Thaumarchaeota viruses. These findings have led us to name this group of new viruses marthavirus (MARine THAumarchaea viruses). The terminase, PD-D/EXK nuclease and the combination of RadA/ATPase protein sequences of only the marthavirus were aligned, and a phylogenetic tree was constructed for each of them (Supporting Information Fig. S1). However, we could not identify any clustering of the sequences or a distinct pattern linking genomic phylogeny and place of isolation.

In order to gain more insights into the putative host of the new viruses, we first identified all the Thaumarchaea genomes including pure culture, Single-Cell Genomes (SAGs) and Metagenomic Assembled-Genomes (MAGs) available. Only those SAGs and MAGs with an estimated completion ≥70% and ≤ 5% contamination were considered. In total, we analysed 94 genomes that were classified into 12 clusters based on pairwise comparisons of average nucleotide identity (ANI; Supporting Information Fig. S2). Clusters A-G form an independent clade composed of strains from marine origin belonging to the order Nitrosopumilales and unclassified Thaumarchaeota. The other clade contained a mix between genomes recovered from soil metagenomes, mostly members of the genus Nitrososphaera (clusters H-J) and marine samples (K and L) (Supporting Information Fig. S2). Metagenomic recruitment of a representative of each cluster in the same metagenomic samples where marthavirus recruited showed that clusters F and G were the most prevalent (Supporting Information Fig. S3). Cluster G is represented by members of the genus *Ca.* Nitrosopelagicus. Only one representative of this group has been recovered by pure culture and showed a ubiquitous distribution in oligotrophic marine waters (Santoro et al., 2015).

Viral genomic features

Despite the high degree of sequence and gene-content divergence ([ANI 70.8%, coverage 6.16%]; [Average Amino Acid Identity (AAI) 53.5%; percentage of common proteins 40.43%]), the alignment of the two complete genomes showed that synteny was remarkably well preserved, also among the other sequences, with two clearly conserved genomic regions (structural and DNA related) (Fig. 2A), separated by a variable region that in Marthavirus-1 contains the auxiliary metabolic gene cobalamin biosynthesis protein (*cob*S). This gene catalyses the final step in the cobalamin biosynthesis in Thaumarchaeota, but not in marine Group II/III Euryarchaeota. Cobalamin (vitamin B_{12}) plays an important role in all three domains of life as a cofactor in the synthesis of amino acids (cobalamin-dependent methionine synthase) or DNA (ribonucleotide reductase—RNR), as well as in other metabolic pathways (Doxey et al., 2015), but only a few taxa are capable to synthetize it (Doxey et al., 2015). A recent study has implicated the Thaumarchaeota as important producers of cobalamin in aquatic environments (Doxey et al., 2015). In fact, some studies have reported a relationship between the availability of vitamin B_{12} and the distribution and growth of phytoplankton and bacterioplankton blooms (Sañudo-Wilhelmy et al., 2006). Furthermore, this gene has been found in cyanophages, suggesting that it could be potentially associated with RNR during nucleotide metabolism (Helliwell et al., 2016) boosting the replication of viral DNA. Phylogenetic analysis showed that Marthavirus-encoded *cob*S is not related to archaeal *cob*S (Supporting Information Fig. S4). Remarkably, viral *cob*S sequences clustered together and separated from their hosts, suggesting a different evolutionary history.

Distribution and genomic diversity

To assess the abundance, distribution and genomic diversity of the novel group of viruses, we performed fragment recruitment analysis by comparing each sequence to

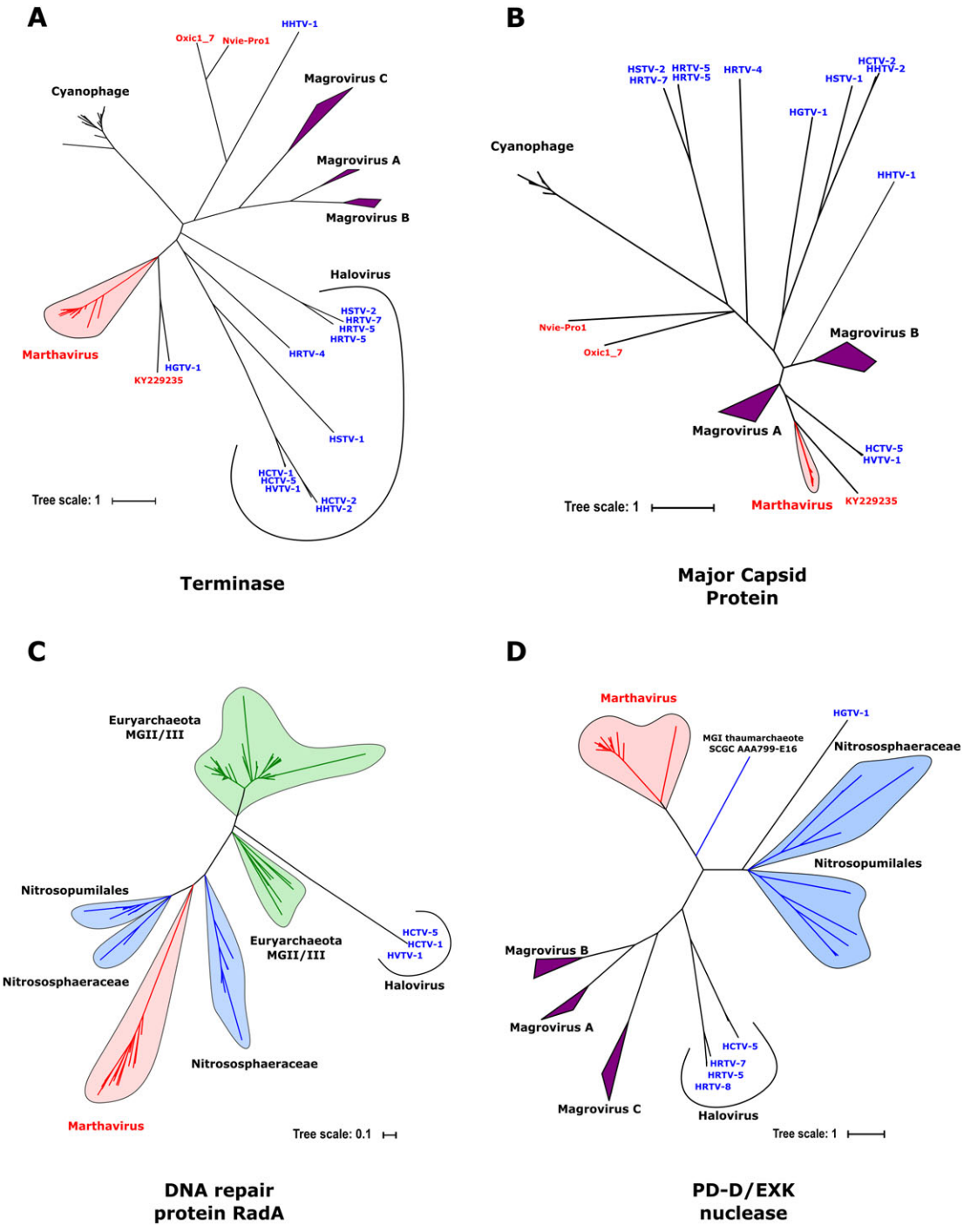

Fig. 1. Unrooted Maximum Likelihood phylogenetic trees of the (A) terminase large subunit, (B) major capsid, (C) DNA repair RadA and (D) PD-D/EXK nuclease proteins. Marthavirus gene sequences were compared against the putative thaumarchaeal reference genomes KY229235, Oxic1_7 and Nvie-Pro1 (coloured in red). Additionally, viral sequences of magrovirus, halovirus and cyanophages and archaeal cellular sequences of MG-I Thaumarchaeota and MGII/III Euryarchaeota were also included in the analysis.

314 metagenomes from Mediterranean, *Tara* Oceans and Malaspina datasets (cellular and viral fraction) with a sequence identity threshold of 70%. We considered only those samples where these viral genomes recruited more than 10 Reads per Kilobase of genome and Gigabase of metagenome (RPKG). As expected, the marthavirus genomes recruited from metagenomes containing Thaumarchaeal genomes, albeit at significantly lower levels and with a more restricted distribution (Fig. 2B). While reference genomes of Thaumarchaeota were detected in 65% of the

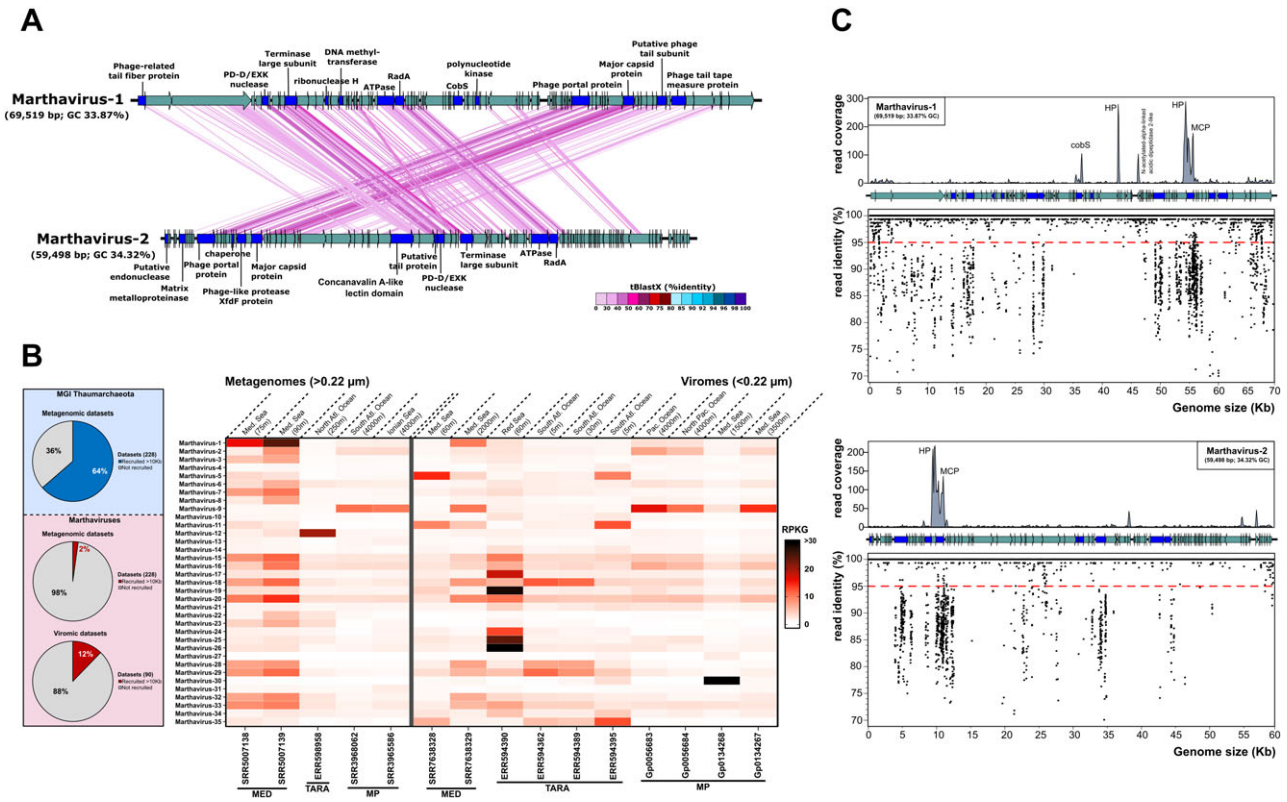

Fig. 2. A. Whole-genome translated nucleotide (tBlastX) comparison between the two complete Thaumarchaeota viruses, Marthavirus-1 and Marthavirus-2. Genome size and GC content are indicated between brackets. Hypothetical and annotated proteins are coloured in green and blue respectively.B. Recruitments of the novel Marthaviruses within the different metagenomic and metaviromic datasets of the Mediterranean Sea (MED), *Tara* Oceans (TARA) and Malaspina (MP) expeditions. On the left, pie charts indicating the percentage of datasets where MG-I Thaumarchaeota and Marthaviruses recruited >10RPKG with a threshold of 70% identity. On the right, a heatmap showing the abundance, measured in RPKG of the selected metagenomes and viromes where the genomes recruited >10RPKG with a threshold of 70% identity.C. Metagenome and Metatranscriptome analysis of Marthavirus-1 and -2. In the upper panel, a mapping of the metatranscriptomic raw reads from the sample Med-OCT2015-90m_MT (>99% identity, 100 bp window) is represented. In the lower panel, a recruitment plot using metagenomic raw reads from the sample Med-OCT2015-90 m (>70% identity, >50 bp long) is shown. Each dot represents a mapped raw read. Red line indicates the species threshold (95% identity). Genes are coloured according to (A). HP: hypothetical protein, MCP: major capsid protein.

metagenomic samples analysed, marthaviruses were found only in 2% of the metagenomes and 12% of the viromes contrasting with the global abundance of magroviruses in the *Tara* Oceans samples (Philosof et al., 2017), marthaviruses showed a patchy distribution. In fact, the majority of the samples where these viruses recruited came from the Mediterranean Sea and the South Atlantic Ocean (Fig. 2B).

Interestingly, most of the viral genomes recruited reads at more than 99% nucleotide identity, with minimal coverage below 95% identity (Supporting Information Fig. S5). These results suggest that marthaviruses may form a population with low intra-population diversity, but with significant divergence among groups.

Metatranscriptome analysis

From the same seawater sample (Western Mediterranean Sea, 90 m) where we obtained 18 marthavirus genomes (Supporting Information Table S1) we also performed a metatranscriptome sequencing. These data could provide clues about the prevailing activities during infection. cDNA reads were mapped onto the two complete genomes assembled from this sample (Marthavirus-1 and -2). Most abundant transcripts in both viruses corresponded to the MCPs, which is required for viral assembly (Fig. 2C). In cyanophages, transcription of the structural genes, including MCP, tail and putative tail fiber proteins, is highest during the final phase of infection (Doron et al., 2016). These data confirm the active viral replication in our sample.

Remarkably, we observed that the *cob*S, encoded within Marthavirus-1 genome, was also highly expressed in the metatranscriptome. Although no study of the structure and activity of the CobS-like viral proteins has been done, results of the mRNA transcripts indicate that the presence of this gene may have an important role during the infection process. The acquisition of AMGs has been repeatedly seen in both bacterial and archaeal viral genomes (Rosenwasser et al., 2016), and their presence

modulate host metabolism to favour a more efficient viral replication.

Nine marthavirus genomes were recovered from viromic samples from the Chesapeake estuary. Consequently, we used the metagenomic, viromic and metatranscriptomic datasets collected there (Maresca et al., 2018). Similar results were obtained after analysing the transcripts for the two different genomes (Supporting Information Fig. S6) that recruited the most (Supporting Information Fig. S7). Again, the MCP was the most expressed gene in Marthavirus-4. The *cob*S gene encoded within the Marthavirus-10 genome was expressed as well, although several genes, mostly hypothetical proteins but also an adhesin, which might mediate the virus-host adhesion, and a metallo-phosphatase were expressed.

In summary, this study characterized several uncultivated viruses assembled from metagenomic samples that infect marine Thaumarchaeota, which we designated marthavirus. It is important to emphasize that several (23 out of 35) of the sequences were obtained from cellular (>0.2 μm) metagenomes reinforcing the idea that they are an important tool to study environmental viral communities containing complementary information which is sometimes missing in viromes. The cellular fraction obviously contains abundant viral material due to the cells retrieved while undergoing viral lysis (López-Pérez et al., 2017). Due to the ecological importance of marine Thaumarchaeota, which are important components in the global nitrogen and carbon nutrient cycling, the study of the thaumarchaea-infecting viruses comprises a key element to understand the dynamics of marine Thaumarchaeota in the ocean.

While this article was in revision, another set of viruses linked to Thaumarchaeaota was reported. They were identified as contigs that encode the viral capsid and thaumarchaeal ammonia monooxygenase genes (*amo*C), highlighting the potential impact of these viruses on N cycling in the oceans (Ahlgren et al., 2018). However, those genomes are very different from the ones described here (only one sequence had 3.3 Kb 99% similar to the marthavirus-13). In addition, we have not found in our dataset any gene encoding AmoC that was the search criterium used by these authors (Ahlgren et al., 2018). Together with our results, the discovery of these viruses highlights the likely enormous diversity of Thaumarchaeota viruses present in the ocean.

Experimental procedures

Sample collection and processing

Six metagenomic samples from a depth profile in the Mediterranean Sea were taken on 15 October 2015. Information about the location and sampling procedure can be found in Haro-Moreno et al. (2018). Additionally, in the same cruise, a metatranscriptome was made from a sample collected at 90 m deep. For the RNA sample, 200 liter of seawater was collected and immediately filtered in a shaded area onboard through a 0.22 μm polyethersulfone filter that was suspended with RNAlater and kept on dry ice until storage at −80 °C. RNA extraction was performed according to the phenolic PGTX (Miller et al., 2017). Metatranscriptome was sequenced using Illumina Hiseq-4000 (150 bp, paired-end read) (Macrogen, Republic of Korea).

Genome annotation

The resulting genes on the assembled contigs were predicted using Prodigal (Hyatt et al., 2010). tRNA and rRNA genes were predicted using tRNAscan-SE (Lowe and Eddy, 1996), ssu-align (Nawrocki, 2009) and meta-RNA (Huang et al., 2009). Predicted protein sequences were compared against NCBI NR databases using USEARCH6 (Edgar, 2010) and against COG (Tatusov et al., 2001) and TIGFRAM (Haft et al., 2001) using HMMscan (Eddy, 2011) for taxonomic and functional annotation.

Identification of novel archaeal viruses

MCP, terminase and portal proteins of marthavirus-1 and KY229235 sequences were used as queries to search against several marine metagenomes (Haro-Moreno et al., 2018; López-Pérez et al., 2017; Duarte, 2015; Sunagawa et al., 2015) using DIAMOND (blastp option, top hit, ≥ 30% identity, ≥ 50% alignment length, *E* value $< 10^{-5}$; Buchfink et al., 2015). Only contigs larger than 8Kb were taken into account. These sequences were also filtered using VirFinder (Ren et al., 2017) to confirm the viral origin.

Metagenomic read recruitments

Genomes of known marine Thaumarchaeota (available up to May 2018 in the NCBI database) and the Marthavirus recovered in this work were used to recruit reads from our metagenomic and metaviromic datasets (Haro-Moreno et al., 2018; López-Pérez et al., 2017), together with those retrieved from the *Tara* Oceans (Sunagawa et al., 2015) and Malaspina expeditions (Duarte, 2015) and the Chesapeake estuary (Maresca et al., 2018), using BLASTN (Altschul et al., 1997), with a cut-off of 70% nucleotide identity over a minimum alignment length of 50 nucleotides. Metagenomic samples where archaeal and viral genomes recruited less than 10 reads per kilobase of genome per gigabase of metagenome (RPKG) were discarded.

Phylogenetic trees of hallmark proteins

Manual inspection of the viral genomes was used to retrieve the amino acid sequences of the Terminase, Major Capsid, RadA, PD-D/EXK nuclease and CobS proteins. To infer their taxonomic relationships, sequences coming from marine Thaumarchaeota and Euryarchaeota genomes, as well as from other archaeal viruses (magroviruses and haloviruses) were used. For the CobS protein, we also included sequences coming from cyanobacterial genomes and phages. Sequences were aligned with MUSCLE (Edgar, 2004) and a Maximum-Likelihood tree was constructed with MEGA 7.0 (Kumar et al., 2016). Jones-Taylor-Thornton model, gamma distribution with five discrete categories, 100 bootstraps, positions with less than 80% site coverage were eliminated.

Thaumarchaeota diversity

Genome completeness and degree of contamination was estimated with CheckM (Parks et al., 2015). The ANI between strains was calculated using JSpecies software package v1.2.1 using default parameters (Richter and Rossello-Mora, 2009).

Acknowledgements

This work was supported by grants 'MEDIMAX' BFPU2013-48007-P, 'VIREVO' CGL2016-76273-P [AEI/FEDER, EU], (co-founded with FEDER funds); Acciones de dinamización 'REDES DE EXCELENCIA' CONSOLIDER CGL2015-71523-REDC from the Spanish Ministerio de Economía, Industria y Competitividad and PROMETEO II/2014/012 'AQUAMET' from Generalitat Valenciana. JHM was supported with a Ph.D. fellowship from the Spanish Ministerio de Economía y Competitividad (Grant No. BES-2014-067828). MLP was supported with a Postdoctoral fellowship from the Valencian Consellería de Educació, Investigació, Cultura i Esport (Grant No. APOSTD/2016/051).

Author contributions

MLP conceived the project. MLP and JHM performed bioinformatic analyses. MLP, JHM, FRV and JRT wrote the manuscript with contributions from all authors to data analysis, figure generation and the final manuscript.

Data availability

Data (viral sequences and metatranscriptome) presented in this manuscript has been submitted to NCBI and are available under BioProject accession numbers PRJNA352798 and PRJNA484324. The metatranscriptomic sample has been deposited in the SRA database (Med-OCT2015-90m_MT – SRR7633016).

References

Ahlgren, N. A., Chen, Y., Needham, D. M., Parada, A. E., Sachdeva, R., Trinh, V., *et al*. (2017) Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems. *Environ Microbiol* **19**: 2434–2452.

Ahlgren, N. A., Fuchsman, C. A., Rocap, G., and Fuhrman, J. A. (2018) Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *ISME J* **1**. https://doi.org/10.1038/s41396-018-0289-4.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Berg, I. A., Kockelkorn, D., Ramos-Vera, W. H., Say, R. F., Zarzycki, J., Hügler, M., *et al*. (2010) Autotrophic carbon fixation in archaea. *Nat Rev Microbiol* **8**: 447–460.

Buchfink, B., Xie, C., and Huson, D. H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

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.

Church, M. J., DeLong, E. F., Ducklow, H. W., Karner, M. B., Preston, C. M., and Karl, D. M. (2003) Abundance and distribution of planktonic Archaea and bacteria in the waters west of the Antarctic peninsula. *Limnol Oceanogr* **48**: 1893–1902.

Danovaro, R., DellAnno, A., Corinaldesi, C., Rastelli, E., Cavicchioli, R., Krupovic, M., *et al*. (2016) Virus-mediated archaeal hecatomb in the deep seafloor. *Sci Adv* **2**: 1–10.

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

Doron, S., Fedida, A., Hernández-Prieto, M. A., Sabehi, G., Karunker, I., Stazic, D., *et al*. (2016) Transcriptome dynamics of a broad host-range cyanophage and its hosts. *ISME J* **10**: 1437–1455.

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.

Duarte, C. M. (2015) Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. *Limnol Oceanogr Bull* **24**: 11–14.

Eddy, S. R. (2011) Accelerated profile HMM searches. *PLoS Comput Biol* **7**: e1002195.

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

Edgar, R. C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

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

Fuhrman, J. A., and Ouverney, C. C. (1998) Marine microbial diversity studied via 16S rRNA sequences: cloning results from coastal waters and counting of native archaea with fluorescent single cell probes. *Aquat Ecol* **32**: 3–15.

Grazziotin, A. L., Koonin, E. V., and Kristensen, D. M. (2017) Prokaryotic virus orthologous groups (pVOGs): a resource

for comparative genomics and protein family annotation. *Nucleic Acids Res* **45**: D491–D498.

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.

Haro-Moreno, J. M., López-Pérez, M., de la Torre, J. R., Picazo, A., Camacho, A., and Rodriguez-Valera, F. (2018) Fine metagenomic profile of the Mediterranean stratified and mixed water columns revealed by assembly and recruitment. *Microbiome* **6**: 128.

Helliwell, K. E., Lawrence, A. D., Holzer, A., Kudahl, U. J., Sasso, S., Kräutler, B., *et al*. (2016) Cyanobacteria and eukaryotic algae use different chemical variants of vitamin B12. *Curr Biol* **26**: 999–1008.

Huang, Y., Gilna, P., and Li, W. (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**: 1338–1340.

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

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.

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.

Krupovic, M., Spang, A., Gribaldo, 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.

Kumar, S., Stecher, G., and Tamura, K. (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**: msw054.

Labonté, J. M., Swan, B. K., Poulos, B., Luo, H., 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.

López-Pérez, M., Haro-Moreno, J. M., Gonzalez-Serrano, R., Parras-Moltó, M., and Rodriguez-Valera, F. (2017) Genome diversity of marine phages recovered from Mediterranean metagenomes: size matters. *PLoS Genet* **13**: e1007018.

Lowe, T. M., and Eddy, S. R. (1996) TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

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

Maresca, J. A., Miller, K. J., Keffer, J. L., Sabanayagam, C. R., and Campbell, B. J. (2018) Distribution and diversity of rhodopsin-producing microbes in the Chesapeake Bay. *Appl Environ Microbiol* **84**: 00137–00118.

Miller, D. R., Pfreundt, U., Elifantz, H., Hess, W. R., and Berman-Frank, I. (2017) Microbial metatranscriptomes from the thermally stratified gulf of Aqaba/Eilat during summer. *Mar Genomics* **32**: 23–26.

Nawrocki, E. P. (2009) *Structural RNA Homology Search and Alignment Using Covariance Models.* PhD Thesis, Washington University, St. Louis. p. 256. Available at https://openscholarship.wustl.edu/etd/256. doi:https://doi.org/10.7936/K78050MP

Nigro, O. D., Jungbluth, S. P., Lin, H. T., Hsieh, C. C., Miranda, J. A., Schvarcz, C. R., *et al*. (2017) Viruses in the oceanic basement. *MBio* **8**: e02129-16.

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

Philosof, A., Yutin, N., Flores-Uribe, J., Sharon, I., Koonin, E. V., and Béjà, O. (2017) Novel abundant oceanic viruses of uncultured Marine Group II Euryarchaeota. *Curr Biol* **27**: 1362–1368.

Pietilä, M. K., Demina, T. A., Atanasova, N. S., Oksanen, H. M., and Bamford, D. H. (2014) Archaeal viruses and bacteriophages: comparisons and contrasts. *Trends Microbiol* **22**: 334–344.

Rachel, R., Bettstetter, M., Hedlund, B. P., Häring, M., Kessler, A., Stetter, K. O., and Prangishvili, D. (2002) Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. *Arch Virol* **147**: 2419–2429.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**: 69.

Richter, M., and Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* **106**: 19126–19131.

Rosenwasser, S., Ziv, C., van Creveld, S. G., and Vardi, A. (2016) Virocell metabolism: metabolic innovations during host–virus interactions in the ocean. *Trends Microbiol* **24**: 821–832.

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 viruses. *Nature* **537**: 689–693.

Santoro, A. E., Buchwald, C., McIlvin, M. R., and Casciotti, K. L. (2011) Isotopic signature of N2O produced by marine ammonia-oxidizing Archaea. *Science* **333**: 1282–1285.

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.

Sañudo-Wilhelmy, S. A., Gobler, C. J., Okbamichael, M., and Taylor, G. T. (2006) Regulation of phytoplankton dynamics by vitamin B12. *Geophys Res Lett* **33**: L04604.

Snyder, J. C., Bolduc, B., and Young, M. J. (2015) 40 years of archaeal virology: expanding viral diversity. *Virology* **479–480**: 369–378.

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., and Chisholm, S. W. (2005) Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS Biol* **3**: 0790–0806.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., *et al*. (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**: 1261359.

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**: e95380.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., *et al*. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**: 22–28.

Vik, D. R., Roux, S., Brum, J. R., Bolduc, B., Emerson, J. B., Padilla, C. C., *et al*. (2017) Putative archaeal viruses from the mesopelagic ocean. *PeerJ* **5**: e3428.

Supporting Information

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

Table S1. Genomic properties, isolation source and accession number for the 35 marthaviruses.

Table S2. Protein clusters present in at least half of the Mathavirus genomes

Fig. S1. Maximum likelihood phylogenetic trees for terminase, RadA+ATPase and PD-D/EXK nuclease protein. ColoredColoured circles indicate origin of the metagenome

Fig. S2. Pairwise comparison among all the available Thaumarchaeota genomes and the ones assembled here using average nucleotide identity (ANI) distances.

Fig. S3. Metagenomic recruitment of a representative of each cluster in the same metagenomics samples where mathavirus recruited the most.

Fig. S4. Unrooted Maximum Likelihood phylogenetic tree of the CobS protein.

Fig. S5. Recruitments of selected marthaviruses among several metagenomic and metaviromic datasets. Red line represents the 95% species identity threshold. MED-MG: Mediterranean metagenomes; MED-MV: Mediterranean metavirome; T-MV: TARA metavirome; T-MG: TARA metagenome; MP-MG: Malaspina metagenome; MP-MV: Malaspina metavirome. M#: Marthavirus-#.

Fig. S6. Metagenome and Metatranscriptome analyses of Marthavirus-4 and -10. In the upper panel, a mapping of the metatranscriptomic raw reads from the sample SRR5830089 (>99% identity, 100 bp window) is represented. In the lower panel, a recruitment plot using metagenomic raw reads from the sample SRR5468101 (>70% identity, >50 bp long) is shown. Each dot represents a mapped raw read. Red line indicates the species threshold (95% identity). HP: hypothetical protein.

Fig. S7. A) Heatmap showing the abundance, measured in RPKG, of the novel marthaviruses within the different metagenomic and metaviromic samples collected in the Chesapeake Bay. Only those samples where at least one of the genomes recruited >10 RPKG are considered. Red circles indicate those viral genomes recovered from the Chesapeake Bay. B) Box plot of the selected samples in A).