Report

Current Biology

Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota

Highlights

- A novel viral group, magroviruses, likely infects marine group II archaea
- Magroviruses are highly abundant in oceanic surface waters worldwide
- Magroviruses have linear, double-stranded DNA genomes of about 100 kilobases
- Magroviruses encode a near complete replication apparatus of apparent archaeal origin

Authors

Alon Philosof, Natalya Yutin, José Flores-Uribe, Itai Sharon, Eugene V. Koonin, Oded Béjà

Correspondence

aphilosof@gmail.com (A.P.),
beja@tx.technion.ac.il (O.B.)

In Brief

Philosof et al. report on newly identified viruses (magroviruses), detected using metagenomics, that most likely infect the uncultured marine group II Euryarchaeota. Magroviruses encode a structural module characteristic of tailed viruses and unexpectedly, a nearly complete replication apparatus of apparent archaeal origin.

Philosof et al., 2017, Current Biology *27*, 1362–1368
May 8, 2017 © 2017 The Author(s). Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.cub.2017.03.052

Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota

Alon Philosof,[1,*] Natalya Yutin,[2] José Flores-Uribe,[1] Itai Sharon,[3,4] Eugene V. Koonin,[2] and Oded Béjà[1,5,*]
[1]Faculty of Biology, Technion – Israel Institute of Technology, Haifa 32000, Israel
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[3]Migal Galilee Research Institute, Kiryat Shmona 11016, Israel
[4]Tel Hai College, Upper Galilee 12210, Israel
[5]Lead Contact
*Correspondence: aphilosof@gmail.com (A.P.), beja@tx.technion.ac.il (O.B.)
http://dx.doi.org/10.1016/j.cub.2017.03.052

SUMMARY

Marine group II Euryarchaeota (MG-II) are among the most abundant microbes in oceanic surface waters [1–4]. So far, however, representatives of MG-II have not been cultivated, and no viruses infecting these organisms have been described. Here, we present complete genomes for three distinct groups of viruses assembled from metagenomic sequence datasets highly enriched for MG-II. These novel viruses, which we denote magroviruses, possess double-stranded DNA genomes of 65 to 100 kilobases in size that encode a structural module characteristic of head-tailed viruses and, unusually for archaeal and bacterial viruses, a nearly complete replication apparatus of apparent archaeal origin. The newly identified magroviruses are widespread and abundant and therefore are likely to be major ecological agents.

RESULTS AND DISCUSSION

Marine members of the archaeal phylum Euryarchaeota are divided into three groups: marine groups II (MG-II), III (MG-III), and IV (MG-IV) [5]. MG-II archaea dominate the photic zone of oligotrophic oceans [1, 2], show seasonal variation [6], and comprise up to 90% of the total archaea and one-third of all microbial cells during spring blooms in the Atlantic [7]. Metatranscriptomic analyses show that MG-II archaea are among the most transcriptionally active microbial groups in the coastal Pacific Ocean, with transcription levels and patterns similar to those of *Pelagibacter ubique* and SAR86 [8, 9].

Despite their high abundance and transcription activity, not a single representative of MG-II has been cultured. Nevertheless, MG-II genomes have been assembled for MG-II [10, 11], and the analyses suggest that MG-II are motile, photoheterotrophic, and capable of degrading polymers such as proteins and lipids [5].

To gain further insight into the diel activity of marine Euryarchaeota in the Red Sea, we examined metagenomic bins containing MG-II signatures. The contigs in these bins were retrieved from a cross assembly of microbial, viral, and transcriptomic samples collected at four time points during a single day in the Gulf of Aqaba in the Red Sea (ENA: PRJEB19060). Manual inspection showed that one bin (169) contained a metagenome-assembled genome (MAG) (156409) carrying hallmark viral genes including predicted major capsid protein (MCP), portal protein, and large subunit of the terminase, as well as DNA polymerase of the B family (DNAP). This contig contained overlapping terminal regions, suggesting that it represents a complete, terminally redundant viral genome.

Viruses have been previously isolated from members of several euryarchaeal groups. Euryarchaeal dsDNA viruses show diverse morphologies including spindle-shaped, icosahedral, pleomorphic, and head-tailed viruses. The latter group resembles the bacterial head-tailed phages (order *Caudovirales*), both in morphology and genome organization, and is currently classified into the families *Myoviridae*, *Podoviridae*, and *Siphoviridae*, each of which includes bacterial and archaeal viruses [12–17]. No viruses infecting MG-II have been isolated [12], and despite the increasing amount of viral metagenomic data [18, 19], candidate viral contigs from the MG-II group have not been reported either.

The protein sequence of the predicted MCP from MAG 156409 showed significant, albeit moderate (33% identity), similarity solely to the MCPs of haloarchaeal siphoviruses (haloviruses) (Figure 1A). Three additional proteins encoded in MAG 156409, namely, primase (Figure 1C), portal protein, and prohead protease (Figure S2), showed comparable levels of similarity to homologs from haloviruses, indicating a relatively distant evolutionary link. We used these protein sequences as queries in BLAST searches against the complete Red Sea assembly dataset, in an attempt to expand the repertoire of virus-related sequences. This search resulted in additional nine viral MAGs (Table S1), all showing the same pattern of homology to haloviruses.

To validate and expand our observations on Red Sea metagenomes, we used the same query sequences in a BLAST search against both the original *Tara* Oceans assembly datasets [20, 21] and a reassembly of the raw sequences from this project (see Supplemental Experimental Procedures). This search yielded 15 additional putative viral genomes related to the Red Sea MAGs (Table S1). All together, we identified 26 putative viral MAGs from two independent metagenomic projects (Red Sea and *Tara* Oceans).

In the phylogenetic tree of the euryarchaeal virus and provirus MCPs, and their environmental homologs (Figure 1A), the MAG

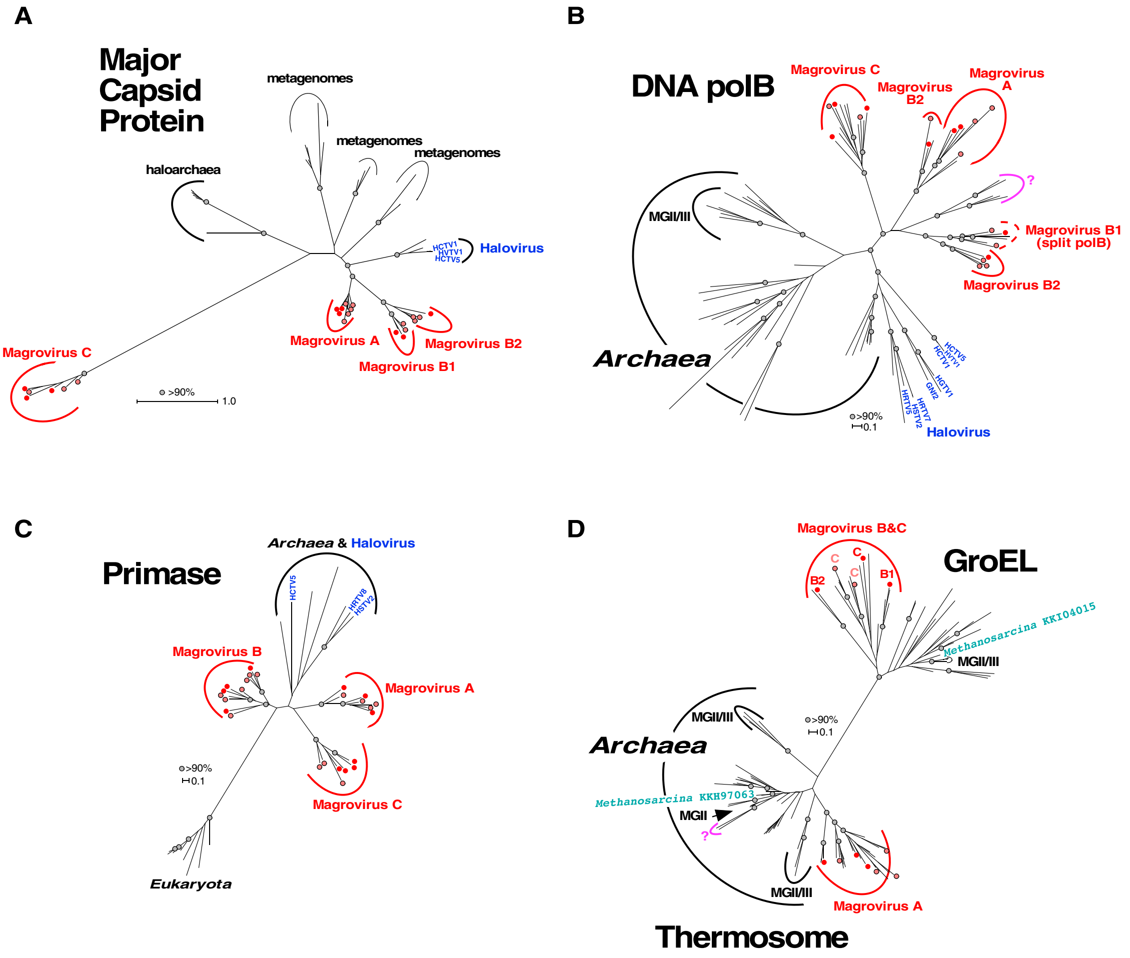


Figure 1. Unrooted Maximum Likelihood Phylogenetic Trees of Conserved Magrovirus Genes

(A) Major capsid protein (MCP).

(B) DNA polymerase B (DNAP).

(C) Archaeo-eukaryotic primase (AEP).

(D) Chaperonins (thermosome subunit and GroEL).

Metagenome-assembled complete or nearly complete genomes (MAGs) of magrovirus from the Red Sea and *Tara* Oceans metagenomes are marked with red and light red circles, respectively. Bootstrap support values greater than 90 are marked with gray circles. See also Figures S2 and S3.

proteins split into three distinct groups, two of which (A, B) join in a clade affiliated with the halovirus MCPs, whereas the third group (C) forms a long branch with an uncertain affiliation. A similar phylogenetic pattern was observed for other hallmark caudoviral genes of the MAGs, namely prohead protease, portal protein, and large subunit of the terminase. Whereas groups A and B cluster together in all these trees, the position of group C changed from tree to tree, suggesting rapid evolution. These findings, along with the fact that 11 MAGs are terminally redundant linear genomes of about 100 kbp in size (Table S1), suggest that these MAGs represent a novel family of head-tailed archaeal viruses. Because, as shown below, these MAGs appear to be strongly associated with MG-II, we provisionally denote them magroviruses (MArine GROup II viruses).

We then compared the gene complements and genome organizations of the three groups of magroviruses in detail. Although, as with many other archaeal viruses [22], most of the magrovirus genes encode proteins without significant similarity to any sequences in current databases, the genomes include two readily identifiable, compact gene blocks, the structural-morphogenetic module and the replicative module (Figure 2). The structural module consists of the genes encoding MCP, portal protein, prohead protease, and terminase and closely resembles the corresponding genomic module of haloviruses (Figure 2). The distinctive feature of magroviruses is the presence of a suit of genes for proteins involved in the genome replication (Figure 2B). All 26 genomes encode DNAP, sliding clamp, clamp loader ATPase (replication factor C), archaeo-eukaryotic primase (AEP), replicative helicase (MCM protein), RadA-like ATPase, single-stranded DNA-binding protein (ssb), and several nucleases; all viruses except for group C also encode one or two ATP-dependent DNA ligases (Figure 2B). Taken together, these proteins could comprise a nearly complete archaeal-type replisome [23], which is unusual among currently known viruses of

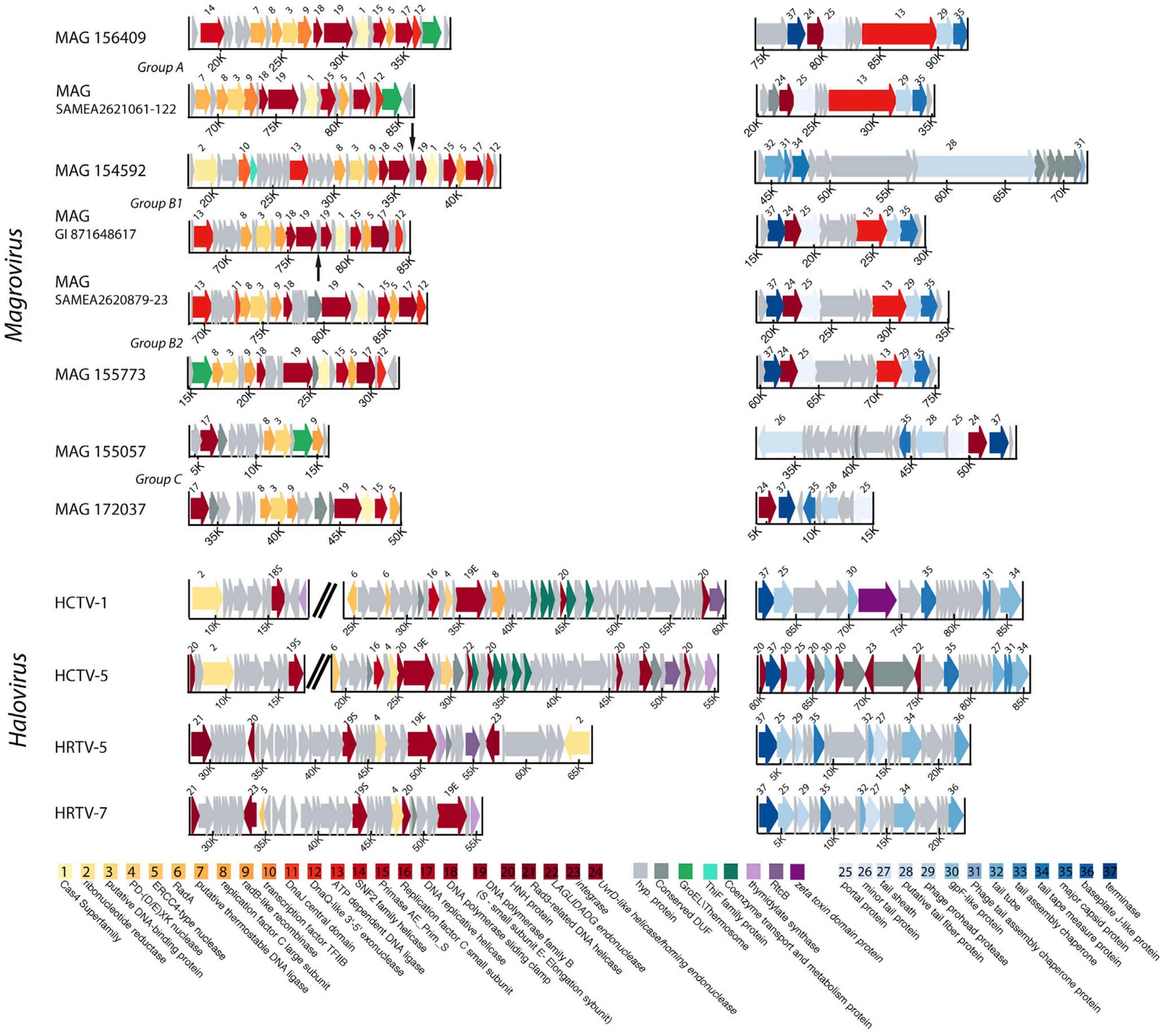

Magrovirus
MAG 156409
Group A
MAG SAMEA2621061-122
MAG 154592
Group B1
MAG GI 871648617
MAG SAMEA2620879-23
Group B2
MAG 155773
MAG 155057
Group C
MAG 172037
Halovirus
HCTV-1
HCTV-5
HRTV-5
HRTV-7
1 Cas4 Superfamily
2 ribonucleotide reductase
3 putative DNA-binding protein
4 PD-(D/E)XK nuclease
5 ERCC4-type nuclease
6 RadA
7 putative thermostable DNA ligase
8 replication factor C large subunit
9 radB-like recombinase
10 transcription factor TFIIB
11 DnaJ central domain
12 DnaQ-like 3'-5' exonuclease
13 ATP dependent DNA ligase
14 SNF2 family helicase
15 Primase AE_Prim_S
16 Replication factor C small subunit
17 DNA replicative helicase
18 DNA polymerase sliding clamp
19 DNA polymerase family B (S - small subunit E- Elongation sybunit)
20 HNH protein
21 Rad3-related DNA helicase
22 LAGLIDADG endonuclease
23 integrase
24 UvrD-like helicase/homing endonuclease
hyp. protein
Conserved DUF
GroEL\Thermosome
ThiF family protein
Coenzyme transport and metabolism protein
thymidylate synthase
RtcB
zeta toxin domain protein
25 portal protein
26 minor tail protein
27 tail sheath
28 putative tail fiber protein
29 phage prohead protease
30 gpF-like protein
31 Phage tail assembly chaperone protein
32 tail tube
33 tail assembly chaperone
34 tail tape measure protein
35 major capsid protein
36 baseplate J-like protein
37 terminase

Figure 2. Genome Organization in Different Groups of Magroviruses and Haloviruses

For each group, detailed genome schemes of the replicative gene block (left, yellow to red) and the structural gene block (right, different shades of blue) are shown. Homologous genes with predicted functions are shown using color code (see key at the bottom). Green arrows indicate thermosome genes, and gray arrows indicate hypothetical proteins. Split DNAP genes from group B1 are marked with a narrow black arrow pointing to the regions between the split genes. See also Figure S1 and Table S1.

archaea and bacteria, although some recently discovered bacteriophages also encode expansive suites of replication proteins [24]. Haloviruses that share the morphogenetic gene block with the magroviruses encompass a smaller complement of replicative genes (Figures 2 and S1). The gene order within both the replicative and the structural modules of magroviruses is highly conserved, with limited rearrangements, largely in group C (Figure 2B), possibly, because operonic organization of functionally linked genes is important for virus reproduction. No auxiliary metabolic genes [25] could be observed in magroviruses.

With the exception of the DNAP, AEP, and ligases, the replicative proteins of magroviruses show low sequence similarity to homologs from cellular organisms that could be detected only through sensitive profile-profile searches. Nevertheless, most of these proteins show signs of archaeal origin as indicated by the provenance of the closest homologs. Due to the low sequence conservation, informative phylogenetic analysis was feasible only for DNAP and AEP. In the phylogenetic tree of DNAPs, the magrovirus polymerases cluster with the halovirus DNAPs [17, 26], and, together, these viral polymerases are related to archaeal DNApolB III that is involved in lagging strand replication (Figure 1B). This gene is apparently subject to frequent horizontal transfer among archaea that is likely to be at least partially mediated by viruses [27, 28], so that the

DNApolB III phylogeny does not follow the archaeal evolutionary tree. Groups A and C that were delineated by analysis of morphogenetic genes retain monophyly in the DNAP tree (Figure 1B), whereas group B splits between two branches, one of which is affiliated with group A (Figure 1B). This discrepancy between the MCP and DNAP phylogenies suggests genetic exchange among diverse magroviruses. In addition to groups A, B, and C, a putative new group (labeled ''?'' in Figures 1B and 1D) was observed in the DNAP tree. The MAGs encoding this group of DNAPs (tentatively, group D) are not represented in the Red Sea metagenomes but are present among scaffolds originating from the *Tara* Oceans project. Group D genomes contain genes for DNAP (Figure 1B), thermosome (Figure 1C), resolvase, and DNA-dependent RNA polymerase A, all with the highest similarity to archaeal homologs. So far no closed genomes of this group were assembled, and a morphogenetic gene block was not identified. Therefore, although a connection to magroviruses is apparent, the nature of group D (virus, provirus, mobile element, megaplasmid, or uncultured archaea) remains to be determined.

Notably, seven group B magroviruses possess a split DNAP gene, whereas in all other viral MAGs, the DNAP gene is uninterrupted (Figures 2B and S3). The split is located within the sequence encoding the catalytic domain, similarly to the DNAP genes of the cyanophage P-SSP7 [29] and *Methanobacterium thermoautotrophicum*. Interrupted DNAP genes are also found in other Euryarchaeota in which the inserts consist of post-transcriptionally excised inteins [30, 31]. The positions of the splits in these archaeal polymerases and the group B magrovirus DNAP are similar, but not identical, suggesting independent, convergent evolutionary events resulting in gene fragmentation. Regardless of the exact evolutionary scenario, the split DNAP is so far unique among archaeal viruses and supports the monophyly of magrovirus group B.

The phylogenetic tree of the AEP supports the monophyly of all three groups of magroviruses as well as the common origin of primases in magroviruses and haloviruses; in this case, however, magrovirus groups A and C cluster with haloviruses, suggesting the possibility of gene exchange between these archaeal viruses (Figure 1C).

Except for group C, all magroviruses encode a DnaQ-like exonuclease that can be implicated in proofreading during viral genome replication. Unlike most of the other magrovirus genes, this nuclease shows significant sequence similarity only to bacterial homologs. Thus, somewhat unexpectedly, magroviruses appear to have acquired genes not only from archaea but also from bacteria.

A notable feature of magroviruses is the presence of two genes encoding distinct ATP-dependent DNA ligases in groups A and B, one within the replicative module and the other one, unexpectedly, embedded in the structural module (Figure 2B). The provenances of these ligases are different as indicated by phylogenetic analysis: the ligase encoded within the structural module represents a distinct family that along with the magrovirus proteins includes ligases from uncharacterized bacteria; the ligase in the replicative block of group B is a typical archaeal variety, whereas the one in group A replicative block belongs to the distinct family known as ''thermostable ligases'' (Figure S1). Thus, apparently, ligases have been acquired by magroviruses on three independent occasions.

The structural modules of all magroviruses also contain another inserted gene that in different groups encodes distinct nucleases (Figure 2B). In groups A, B, and D, this is a homing endonuclease (LAGLI-DADG family) homologous to nucleases of group I self-splicing introns and inteins, which are also present in many bacteriophages. In contrast, in group C, the inserted gene encodes a homolog of the exonuclease subunit of the archaeal DNA polymerase D [32]. An intriguing possibility is that the two nucleases play analogous roles in magrovirus replication. In addition to the conserved replicative genes, several genes implicated in replication are found in individual groups of magroviruses, e.g., ribonucleotide reductase in group B and SNF2 family helicase in group A.

Apart from the replicative and structural-morphogenetic proteins, 12 of the 26 magroviruses encode either a bacterial-type chaperonin GroEL (groups B, B1, and C) or a thermosome subunit, the archaeal homolog of GroEL (groups A and D) (Figure 2B). Unlike the replicative genes, these magrovirus proteins are highly similar to the archaeal and bacterial homologs. Phylogenetic analysis confirmed the sharp split between GroEL and thermosome subunits (Figure 1D). The thermosome subunits of magroviruses group with homologs from MG-II, which is compatible with relatively recent acquisition of these genes by the viruses. A subset of MG-II archaea encode GroEL instead of the thermosome subunit, conceivably owing to displacement of the ancestral archaeal chaperonin by a bacterial homolog. Although the magrovirus GroEL do not group with those of MG-II in the phylogenetic tree, this could be due to the accelerated evolution in the viruses; acquisition of this gene from MG-II archaea remains likely. Comparison of the topology of the chaperonin tree with those of the MCP, DNAP, and AEP trees suggests that the common ancestor of the magroviruses acquired a GroEL gene that was replaced by the thermosome subunit in group A. Chaperonins are not encoded by any known archaeal viruses but have been detected in several bacteriophages [33, 34]. Notably, these phages do not encode the co-chaperonin GroES, and GroES is not required for the phage chaperonin activity [35]. This is likely to be the case for the magrovirus chaperonins as well. Given that, in magroviruses, the chaperonin genes reside in the replicative gene cluster, the chaperonins might facilitate folding of the replicative proteins and replisome assembly.

Mapping reads from the Red Sea and from the *Tara* Oceans on the 26 magrovirus genomes indicate that these viruses are globally widespread, with high abundance in the Indian Ocean, the Red Sea, and the South Pacific and Atlantic Oceans (Figures 3A and 3B), and are composed of different ecotypes (Figure 3A). Magroviruses are highly abundant in the marine environment, third only to SAR11 phages [36] and cyanophages (Figure 4). Surprisingly, the overall abundance of magroviruses was found to be higher than that of SAR116 phages (Figure 4), previously identified as the second most abundant phage group in the oceans [37].

So far, based on microscopic measurements [26], head-tailed viruses appeared to be the least abundant morphotype of archaeal viruses. However, our abundance estimates for magroviruses suggest that these previously uncharacterized

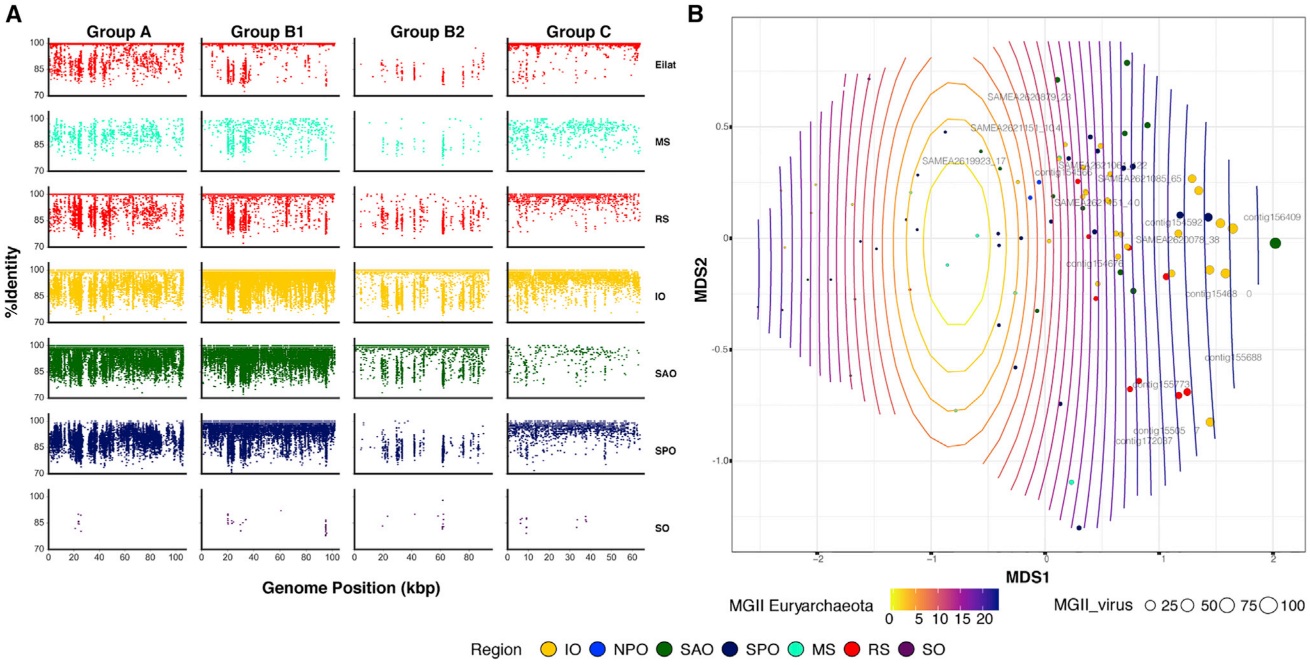


Figure 3. Global Abundance of Magroviruses and MG-II and/or MG-III

(A) Recruitment plots of representatives from different magrovirus groups (A: MAG 154566, B1: MAG 154680, B2: MAG SAMEA2620879_23, C: MAG 155057), using reads (viral fraction) from surface waters (top 5 m) from the Red Sea and different *Tara* Oceans stations.

(B) Non-metric multidimensional scaling (NMDS) plot of sampling sites magrovirus abundance. Circle diameter indicates magrovirus abundance at a specific site. MG-II abundance is represented by a gradient of white to red contour lines. A linear response between the magrovirus abundance ordination and the MG-II abundance variable is represented by fitted contours that are equally spaced parallel lines perpendicular to the MG-II abundance vector (R-sq.[adj] = 0.432; deviance explained = 48%; p value = 4.82e−10).

Region abbreviations are as follows: IO, Indian Ocean; RS, Red Sea; SO, Southern Ocean; MS, Mediterranean Sea; NPO, North Pacific Ocean; SAO, South Atlantic Ocean; SPO, South Pacific Ocean.

head-tailed viruses dominate the archaeal virome in surface marine environments. In contrast to the cyanophages, and the SAR11 phages, normalized counts of magroviruses are almost negligible in the microbial fraction (>0.2 μm) but high in the virus-enriched fraction (<0.2 μm) (Figure 4). Thus, the principal source of the magrovirus reads appear to be free virus particles rather than cell-associated viruses, proviruses or megaplasmids. A possible explanation to this observation is that MG-II show seasonal blooming patterns [6]. However, the analysis was done with samples spanning multiple seasons and regions. The low magrovirus signal in the microbial fraction raises interesting questions on how magrovirus virions are maintained in the marine surface waters for long periods of time.

Despite the abundance of MG-II in the oceans and their apparent ecological importance, no viruses infecting these organisms have been identified so far. Here, using metagenomic

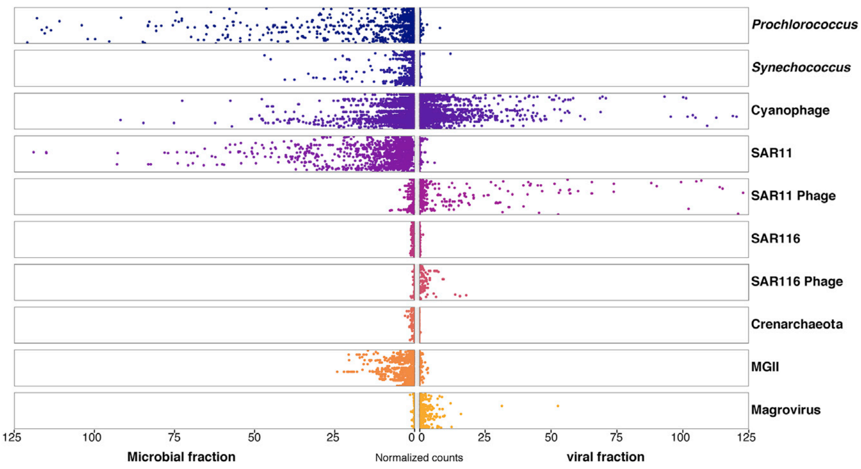


Figure 4. Total Abundance of Magroviruses

The abundance of the Magrovirus reads in Red Sea samples and all *Tara* Oceans microbiomes [21] and viromes [18] is shown along with the abundances of the putative host MG-II, marine group I Crenarchaeota, marine Cyanobacteria (*Synechococcus* and *Prochlorococcus*) and their phages, SAR11 bacteria and their phages, and SAR116 and their phages. Plots for halovirus and representatives from cultured Euryarchaeota are not shown as the signals were close to zero. Horizontal axis: the normalized count (genome fragments per kilobase reference sequence per million library reads [GFPM]; see Supplemental Experimental Procedures). Vertical axis: the sampling stations.

approaches, we describe three distinct groups of viruses associated with MG-II. Unequivocal demonstration of the virus-host relationship between magroviruses and MG-II is currently unfeasible due to the lack of cultivable MG-II representatives. Nevertheless, several lines of evidence strongly suggest that MG-II archaea are indeed the hosts of magroviruses. First, our abundance estimates show that MG-II is the dominant archaeal group in the samples from which the magrovirus genomes were assembled (Figures 3A and 3B). Second, most of the replicative genes of magroviruses and especially the viral chaperonins show clear signs of archaeal provenance, and in some cases, a specific connection with homologs from MG-II. Furthermore, group B magroviruses encode two tRNAs (tRNA^{Leu} and tRNA^{Arg}) that are most similar to the respective tRNAs of MG-II archaea. Some of the marine Euryarchaeal fosmids deposited in GenBank lack 16S rDNA genes, therefore their affiliation is not resolved, and they are deposited in GenBank as MG-II/III. In addition, some representatives of MG-III are abundant in surface waters [38], therefore we cannot rule out the possibility that MG-III are the hosts of (some) magroviruses.

The discovery of the magroviruses is part of the growing trend in virology whereby viruses are recognized solely from metagenomics sequence analysis, as recently formalized by the International Committee for Taxonomy of Viruses [39]. Genome analysis of the magroviruses is consistent with modular evolution of viruses whereby the structural and replicative modules have distinct origins. Magroviruses are unusual among viruses of bacteria and archaea in that they encode an elaborate, apparently (almost) self-sufficient replication apparatus. Given the high abundance of both MG-II archaea and magroviruses, the latter are likely to be important ecological agents, similar to cyanophages or archaeal viruses that apparently infect Thaumarchaeota [40].

While this paper was in revision, identification of a group of putative viruses overlapping with the magrovirus set described here has been reported independently [41].

ACCESSION NUMBERS

The raw sequencing data and assembled contigs reported in this paper have been deposited in the European Nucleotide Archive under project ENA: PRJEB19060.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, three figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.03.052.

AUTHOR CONTRIBUTIONS

A.P. and O.B designed the project. A.P., N.Y., J.F.-U., I.S., E.V.K., and O.B. performed bioinformatic analyses. A.P., E.V.K., and O.B. wrote the manuscript with contributions from all authors to data analysis, figure generation, and the final manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the captain and crew of the R/V *Sam Rothberg* of the Inter University Institute in Eilat, Israel for their expert assistance at sea and Hagay Enav and Idan Bodaker for their help with sampling. The authors would also like to thank David Cohen of the Physics Department at the Technion for his help with the HPC ATLAS cluster and Mart Krupovic of the Institut Pasteur (Paris) for critical reading on the manuscript and helpful suggestions. This work was funded by a European Commission ERC Advanced Grant (no. 321647), the Technion's Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering and the Russell Berrie Nanotechnology Institute, and the Louis and Lyra Richmond Memorial Chair in Life Sciences (to O.B.). N.Y. and E.V.K. are supported by intramural funds of the US Department of Health and Human Services (to the National Library of Medicine).

Received: February 5, 2017
Revised: March 9, 2017
Accepted: March 21, 2017
Published: April 27, 2017

REFERENCES

1. Massana, R., Murray, A.E., Preston, C.M., and DeLong, E.F. (1997). Vertical distribution and phylogenetic characterization of marine planktonic *Archaea* in the Santa Barbara Channel. Appl. Environ. Microbiol. *63*, 50–56.
2. Massana, R., DeLong, E.F., and Pedrós-Alió, C. (2000). A few cosmopolitan phylotypes dominate planktonic archaeal assemblages in widely different oceanic provinces. Appl. Environ. Microbiol. *66*, 1777–1787.
3. Galand, P.E., Casamayor, E.O., Kirchman, D.L., Potvin, M., and Lovejoy, C. (2009). Unique archaeal assemblages in the Arctic Ocean unveiled by massively parallel tag sequencing. ISME J. *3*, 860–869.
4. Lincoln, S.A., Wai, B., Eppley, J.M., Church, M.J., Summons, R.E., and DeLong, E.F (2014). Planktonic Euryarchaeota are a significant source of archaeal tetraether lipids in the ocean. Proc. Natl. Acad. Sci. USA *111*, 9858–9863.
5. Zhang, C.L., Xie, W., Martin-Cuadrado, A.-B., and Rodriguez-Valera, F. (2015). Marine Group II Archaea, potentially important players in the global ocean carbon cycle. Front. Microbiol. *6*, 1108.
6. Murray, A.E., Blakis, A., Massana, R., Strawzewiski, S., Passow, U., Alldredge, A., and DeLong, E.F. (1999). A timeseries assessment of planktonic archaeal variability in the Santa Barbara Channel. Aquat. Microb. Ecol. *20*, 129–145.
7. Pernthaler, A., Preston, C.M., Pernthaler, J., DeLong, E.F., and Amann, R. (2002). Comparison of fluorescently labeled oligonucleotide and polynucleotide probes for the detection of pelagic marine bacteria and archaea. Appl. Environ. Microbiol. *68*, 661–667.
8. Ottesen, E.A., Young, C.R., Eppley, J.M., Ryan, J.P., Chavez, F.P., Scholin, C.A., and DeLong, E.F. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. Proc. Natl. Acad. Sci. USA *110*, E488–E497.
9. Aylward, F.O., Eppley, J.M., Smith, J.M., Chavez, F.P., Scholin, C.A., and DeLong, E.F. (2015). Microbial community transcriptional networks are conserved in three domains at ocean basin scales. Proc. Natl. Acad. Sci. USA *112*, 5443–5448.
10. Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., and Armbrust, E.V. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science *335*, 587–590.
11. Martin-Cuadrado, A.-B., Garcia-Heredia, I., Moltó, A.G., López-Úbeda, R., Kimes, N., López-García, P., Moreira, D., and Rodriguez-Valera, F. (2015). A new class of marine Euryarchaeota group II from the Mediterranean deep chlorophyll maximum. ISME J. *9*, 1619–1634.
12. Prangishvili, D. (2013). The wonderful world of archaeal viruses. Annu. Rev. Microbiol. *67*, 565–585.
13. Krupovic, M., Prangishvili, D., Hendrix, R.W., and Bamford, D.H. (2011). Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol. Mol. Biol. Rev. *75*, 610–635.
14. Porter, K., Russ, B.E., and Dyall-Smith, M.L. (2007). Virus-host interactions in salt lakes. Curr. Opin. Microbiol. *10*, 418–424.

15. Pina, M., Bize, A., Forterre, P., and Prangishvili, D. (2011). The archeoviruses. FEMS Microbiol. Rev. *35*, 1035–1054.

16. Atanasova, N.S., Roine, E., Oren, A., Bamford, D.H., and Oksanen, H.M. (2012). Global network of specific virus-host interactions in hypersaline environments. Environ. Microbiol. *14*, 426–440.

17. Pietilä, M.K., Laurinmäki, P., Russell, D.A., Ko, C.C., Jacobs-Sera, D., Butcher, S.J., Bamford, D.H., and Hendrix, R.W. (2013). Insights into head-tailed viruses infecting extremely halophilic archaea. J. Virol. *87*, 3248–3260.

18. Brum, J.R., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., et al.; Tara Oceans Coordinators (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science *348*, 1261498.

19. Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425–430.

20. Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., et al.; Tara Oceans Consortium Coordinators (2015). Open science resources for the discovery and analysis of Tara Oceans data. Sci. Data *2*, 150023.

21. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al.; Tara Oceans coordinators (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science *348*, 1261359.

22. Iranzo, J., Koonin, E.V., Prangishvili, D., and Krupovic, M. (2016). Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. J. Virol. *90*, 11043–11055.

23. Makarova, K.S., and Koonin, E.V. (2013). Archaeology of eukaryotic DNA replication. Cold Spring Harb. Perspect. Biol. *5*, a012963.

24. Kazlauskas, D., Krupovic, M., and Venclovas, Č. (2016). The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. Nucleic Acids Res. *44*, 4551–4564.

25. Breitbart, M., Thompson, L.R., Suttle, C.A., and Sullivan, M.B. (2007). Exploring the vast diversity of marine viruses. Oceanography (Wash. D.C.) *20*, 135–139.

26. Luk, A.W., Williams, T.J., Erdmann, S., Papke, R.T., and Cavicchioli, R. (2014). Viruses of haloarchaea. Life (Basel) *4*, 681–715.

27. Makarova, K.S., Krupovic, M., and Koonin, E.V. (2014). Evolution of replicative DNA polymerases in archaea and their contributions to the eukaryotic replication machinery. Front. Microbiol. *5*, 354.

28. Takemura, M., Yokobori, S., and Ogata, H. (2015). Evolution of eukaryotic DNA polymerases via interaction between cells and large DNA viruses. J. Mol. Evol. *81*, 24–33.

29. 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*, e144.

30. Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., et al. (1997). Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J. Bacteriol. *179*, 7135–7155.

31. Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., et al. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature *390*, 364–370.

32. Jokela, M., Eskelinen, A., Pospiech, H., Rouvinen, J., and Syväoja, J.E. (2004). Characterization of the 3′ exonuclease subunit DP1 of *Methanococcus jannaschii* replicative DNA polymerase D. Nucleic Acids Res. *32*, 2430–2440.

33. Hertveldt, K., Lavigne, R., Pleteneva, E., Sernova, N., Kurochkina, L., Korchevskii, R., Robben, J., Mesyanzhinov, V., Krylov, V.N., and Volckaert, G. (2005). Genome comparison of Pseudomonas aeruginosa large phages. J. Mol. Biol. *354*, 536–545.

34. Cornelissen, A., Hardies, S.C., Shaburova, O.V., Krylov, V.N., Mattheus, W., Kropinski, A.M., and Lavigne, R. (2012). Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse ΦKZ-related phages. J. Virol. *86*, 1844–1852.

35. Semenyuk, P.I., Orlov, V.N., Sokolova, O.S., and Kurochkina, L.P. (2016). New GroEL-like chaperonin of bacteriophage OBP *Pseudomonas fluorescens* suppresses thermal protein aggregation in an ATP-dependent manner. Biochem. J. *473*, 2383–2393.

36. Zhao, Y., Temperton, B., Thrash, J.C., Schwalbach, M.S., Vergin, K.L., Landry, Z.C., Ellisman, M., Deerinck, T., Sullivan, M.B., and Giovannoni, S.J. (2013). Abundant SAR11 viruses in the ocean. Nature *494*, 357–360.

37. Kang, I., Oh, H.-M., Kang, D., and Cho, J.-C. (2013). Genome of a SAR116 bacteriophage shows the prevalence of this phage type in the oceans. Proc. Natl. Acad. Sci. USA *110*, 12343–12348.

38. Haro-Moreno, J.M., Rodriguez-Valera, F., López-García, P., Moreira, D., and Martin-Cuadrado, A.-B. (2017). New insights into marine group III Euryarchaeota, from dark to light. ISME J. http://dx.doi.org/10.1038/ismej.2016.188.

39. Simmonds, P., Adams, M.J., Benkő, M., Breitbart, M., Brister, J.R., Carstens, E.B., Davison, A.J., Delwart, E., Gorbalenya, A.E., Harrach, B., et al. (2017). Consensus statement: virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. *15*, 161–168.

40. Danovaro, R., Dell'Anno, A., Corinaldesi, C., Rastelli, E., Cavicchioli, R., Krupovic, M., Noble, R.T., Nunoura, T., and Prangishvili, D. (2016). Virus-mediated archaeal hecatomb in the deep seafloor. Sci. Adv. *2*, e1600492.

41. Nishimura, Y., Watai, H., Honda, T., Mihara, T., Omae, K., Roux, S., Blanc-Mathieu, R., Yamamoto, K., Hingamp, P., Sako, Y., et al. (2017). Environmental viral genomes shed new light on virus-host interactions in the ocean. mSphere *2*, e00359–e16.

Current Biology, Volume *27*

Supplemental Information

Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota

Alon Philosof, Natalya Yutin, José Flores-Uribe, Itai Sharon, Eugene V. Koonin, and Oded Béjà

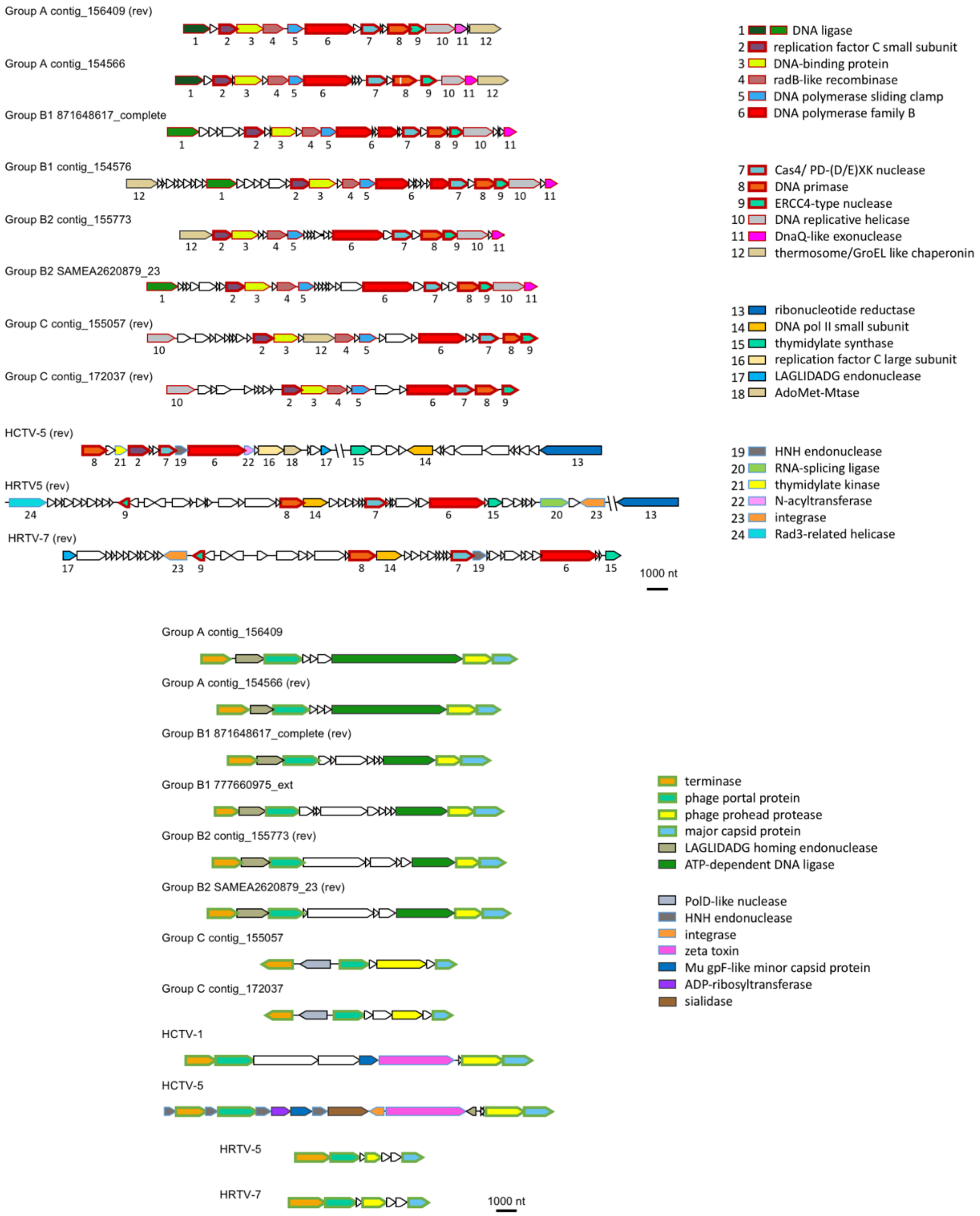
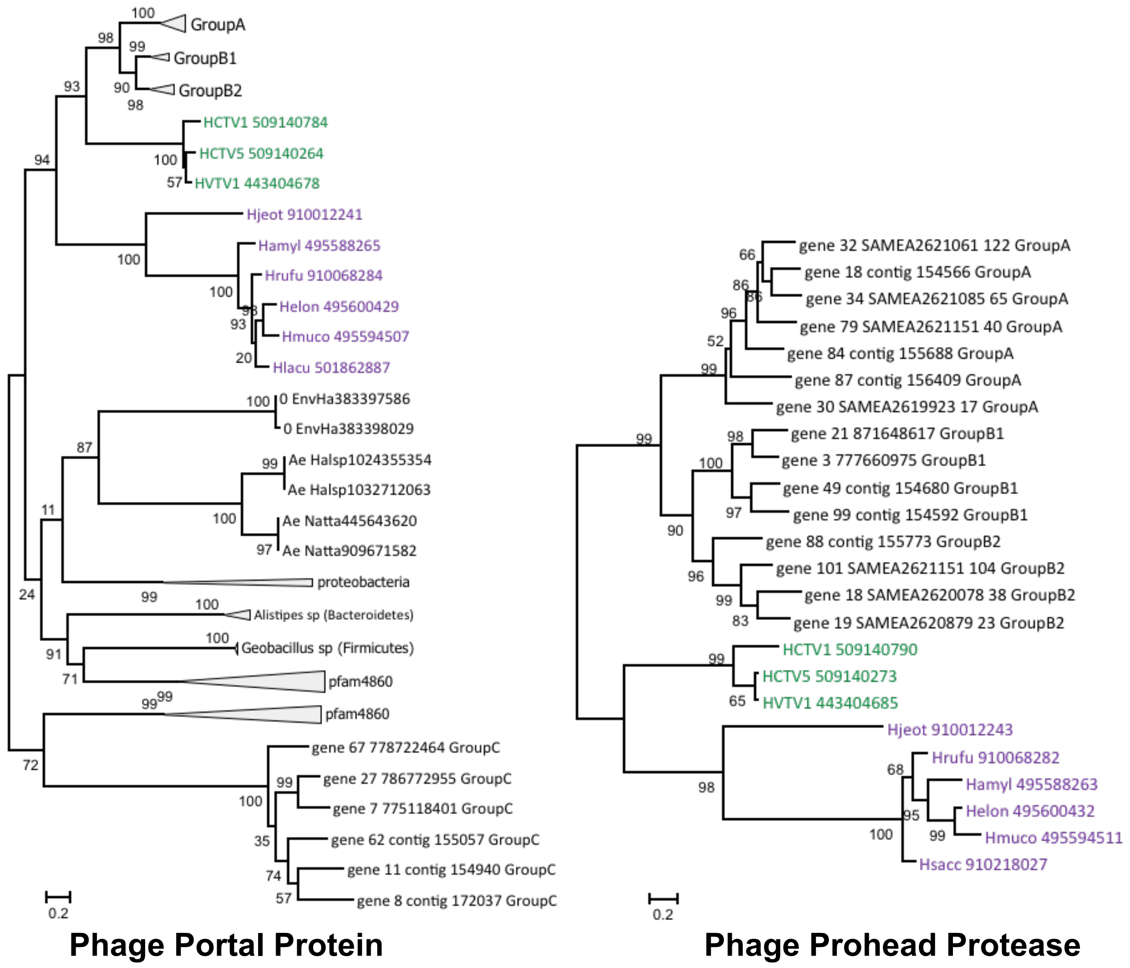


Figure S1: Detailed genomic maps of Magroviruses and haloviruses (related to Figure 2)

Genomic Maps of the replication gene cluster (upper panel) and viral structural gene cluster (lower panel) in selected Magroviruses and the haloviruses HCTV-5, HRTV-5, HRTV7, HCTV-1.



Phage Portal Protein

Phage Prohead Protease

Figure S2: Maximum likelihood phylogenetic trees for phage-portal-protein and phage-prohead-protease protein (related to Figure 1)

Metagenome-assembled genomes (MAGs) from this study are marked with the group name they belong to. Purple sequences are from Archaeal proviruses. Green sequences are from Haloviruses.

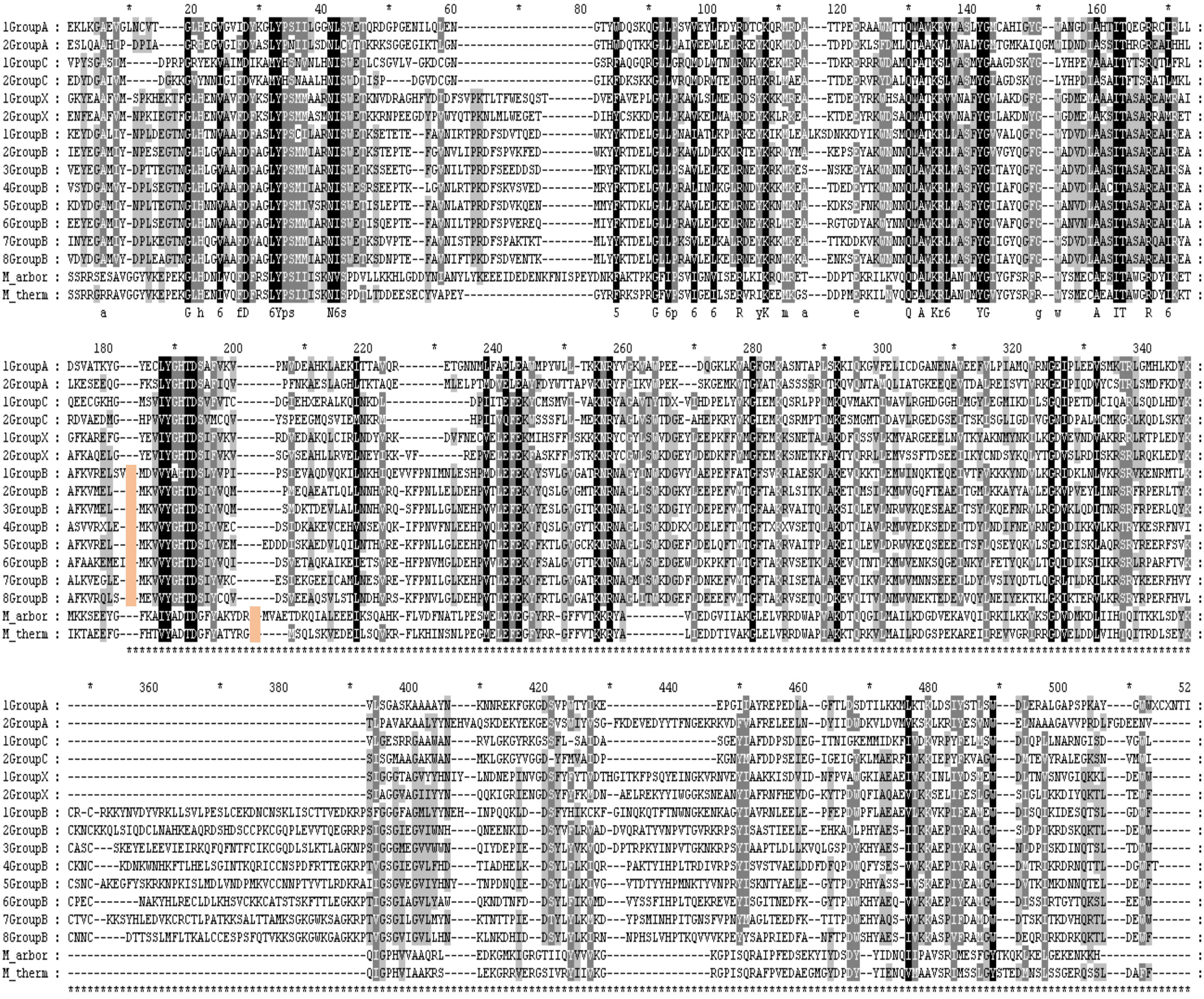


Figure S3: Multiple Sequence alignment of the DNA polymerase family B catalytic domain (related to Figure 1B)

Orange boxes mark splits. Gene abbreviations: 1GroupA, gene26_contig_156409; 2GroupA, gene_66_SAMEA2621061_122; 1GroupC, gene_32_contig_172037; 2GroupC, gene_88_contig_155057; 1GroupX, gene_114_SAMEA2620879_23; 2GroupX, gene_50_contig_155773; 1GroupB, gene_19_17_786803042; 2GroupB, gene_38_37_777660975; 3GroupB, gene_104_107_871648617; 4GroupB, gene59_62_contig_154592; 5GroupB, gene_90_87_contig_154680; 6GroupB, gene56_62_contig_154676; 7GroupB, gene_16_15_772041692; 8GroupB, gene_43_45_837327169; M_arbor, *Methanobrevibacter arboriphilus* gi|757148798 and gi|757150045; M_therm, *Methanothermobacter thermautotrophicus* gi|499179292 and gi|499178307

contig name	**contig length, nt**	***Group***	***notes***	**Extended contig name**	**ext. contig length**	**genome**	**%GC**
SAMEA2621061_122	106637	GroupA	no ext in wgs				39.29
contig_156409	118049	GroupA				complete (circular)	51.1
gi|871648617	78671	GroupB	extended by gi|870181751, gi|871648619, gi|870131030, gi|868862040	871648617_complete	100277	complete (circular)	32.84
contig_154592	96096	GroupB1	no ext in wgs				43.78
contig_154676	53712	GroupB1	no ext in wgs. tRNA^{Leu} and tRNA^{Arg}				34.24
contig_155057	63764	GroupC				complete (circular)	46.5
contig_172037	61133	GroupC				complete (circular)	44.55
SAMEA2620879_23	93879	GroupB2				complete (circular)	35.36
contig_155773	84959	GroupB2	no ext in wgs				35..89
SAMEA2619923_17	102781	GroupA	no ext in wgs				43.94
SAMEA2621085_65	102641	GroupA	no ext in wgs				44.74
SAMEA2621151_40	108353	GroupA	no ext in wgs				52.47
contig_154566	105939	GroupA				complete (circular)	48.06
contig_155688	125638	GroupA				complete (circular)	45.17
gi|772041692	36978	GroupB1	no ext in wgs				32.53
gi|777660975	64966	GroupB1	extended up to 97754 nt by gi|774121447, gi|774778739, gi|774839755, gi|775082829, gi|775114504, gi|775117949, gi|776721548, gi|776810293, gi|776811288, gi| 777672392	777660975_ext	97754	nearly complete	36.64
gi|786803042	22546	GroupB1	no ext in wgs				33.34
gi|837327169	50367	GroupB1	extended a bit (~1200 nt) by gi|765915719, gi|837063603	837327169_ext	51368		43.24
contig_154680	102054	GroupB1	250 nt "insert"			complete (circular)	37.88
gi|766270197	50189	GroupC	extended by gi|789589345	766270197_ext	57681		41.06
gi|775118401	60429	GroupC	extended by gi|777632841, gi|776868392	775118401_ext	71480		44.46
gi|778722464	66608	GroupC	circular, suported by gi|786717538			complete (circular)	43.11
gi|786772955	70925	GroupC	circular, suported by gi|789713218			complete (circular)	43.76
contig_154940	60318	GroupC	no ext in wgs				47.03
SAMEA2620078_38	94102	GroupB2	no ext in wgs				30.14
SAMEA2621151_104	92290	GroupB2	completed by 7 gi|96655877	SAMEA2621151_104_ext	99445	complete (circular)	32.84

Table S1: Metadata on genomes reported in this study (related to Figure 2)

Supplemental Experimental Procedures

The scripts and analyses files used in our data analyses are available on github: https://github.com/BejaLab/**Magrovirus**

Sample Collection and processing

Sampling was performed on October 12th, 2012 at the Gulf of Aqaba, Station A (exact location). Four time points were collected during this single day: 06:00, 12:00, 18:00 and 24:00. At each time point three fractions were collected after 2.8µm pre-filtration: metagenomic (gDNA, >0.22 µm), transcriptomic (cDNA, >0.22 µm) and viral (vDNA, <0.22 µm). DNA was extracted using alkaline-lysis protocol. For the viral fraction, two sampling methods were used: 1) 60L of the flow-through the 0.22 µm filters were concentrated with a TFF filter (Millipore); 2) Iron-Chloride ($FeCl_3$) precipitation of viral particles as described by [S1]. The precipitate was collected on a 0.22 µm Durapore filter (Millipore) and washed with 10mL of the following solution: 0.1M Na_2EDTA, 0.1M $MgCl_2$, 0.125 M Tris, 0.125 M Oxalate at pH ~6. The washed precipitate was further concentrated using a Centricon filter (30 kDa cutoff). Both the TFF and $FeCl_3$ samples were further purified on a CsCl gradient followed by DNase treatment and DNA extraction as described in [S2] with the exception of not using phi29 for whole genome amplification. The transcriptomic fraction was collected onto multiple 0.22 µm durapore filters (Millipore) using a four head peristaltic pump. After collection, the filters were immediately transferred to a screw cap tube containing 1 mL of RNAlater (Ambion) and frozen in liquid nitrogen. Total handling time was less than 15 min, total RNA extraction was done using mirVana RNA isolation kit (Ambion), followed by DNA removal with Turbo DNase (Ambion) and cleanup using RNeasy MinElute Kit (QIAGEN). 10 filters from each time point were extracted separately and then pooled together for sequencing.

The vDNA and gDNA samples were sheared using Covaris E220 with the following parameters: 10% Duty factor, 45 sec duration time, 200 cycles per burst, 175W peak incident power and a temperature of 6 °C. Libraries were constructed with 50 ng of DNA per sample using Illumina's TrueSeq Nano library construction kit according to protocol. Eight PCR cycles were performed in the construction of the gDNA libraries and 15 cycles in the vDNA samples.

The three sample types were loaded on three separate lanes of Illumina HiSeq. Both vDNA sample types (TFF and FeCl3) per time point were barcoded separately. All samples were paired-end (PE) sequenced, the gDNA and vDNA at 150bp from each end and the cDNA 100bp from each end. Sequencing was performed by the Technion Sequencing center on Illumina Hiseq 2500.

Bioinformatic Analyses

The files and scripts used in our data analysis are available in the Github repository found at: https://github.com/BejaLab/**Magrovirus**

Quality Control of Raw Reads

Illumina adapters were first removed from the Raw reads using Trimmomatic [S3] 0.32 (TruSeq3PE.fa; 2:30:10, LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15). The paired-ends were then interleaved and short sequences and orphans were removed with Biopieces (www.biopieces.org). Fastx (http://hannonlab.cshl.edu/fastx_toolkit/index.html) was used to remove low quality bases from the ends of the reads.

Following QC, digital normalization and graph partition using the Khmer package [S4-S6], were applied on each sample. Briefly, the interleaved files were normalized by median coverage (normalize-by-median.py -C 20 -k 21 -x 64e9 -N 4 -p) and afterwards filtered by abundance (filter-abund.py -C 100 -x 64e9 -N 4 -V) and then paired reads were extracted (extract-paired-reads.py). The resulting files were assembled as described later. In addition, these files were partitioned with Khmer [S6] into disconnected de Bruijn graph components in order to assemble each partition separately. The partitioning was done as follows: first the graph was loaded (load-graph.py -k 32 –N 4 –x 60e9) then partitioned (partition-graph.py -s 10e6) merged (merge-partitions.py --ksize 32) and annotated (annotate-partitions.py). Finally, the partitions were extracted (extract-partitions.py) and assembled as described later.

Each sample, post QC, was analyzed with KRAKEN [S7] using the genome databases refseq viral and refseq microbial (http://www.ncbi.nlm.nih.gov/refseq/) at kmer size of 21 to find candidate genomes to map against and to remove vector and other contaminations (such as human sequences and reagent contamination).

Assembly of Red Sea samples

Following QC, contaminations were removed from the samples and the samples were diginormed and partitioned as described above. At first, each sample was assembled separately with IDBA-ud [S8] (Kmers 20-120, steps of 4). In addition, the diginormed and partitioned files were assembled in the same manner with Velvet [S9] (Kmers 21-99, steps of 4). Then, the entire fraction (gDNA, vDNA or RNA) was concatenated (e.g. all four vDNA samples) and then assembled with IDBA-ud (Kmers 20-120, steps of 4). All the assemblies from each fraction were then pooled together and de-replicated using *derep_fulllength* option in VSEARCH (https://github.com/torognes/vsearch). The entire de-replicated assembly of each fraction was then used as an input to CONCOCT [S10]. CONCOCT performs unsupervised binning of metagenomic contigs using a Gaussian mixture model incorporating both coverage and nucleotide composition from multiple samples along with linkage data from the paired end reads. The samples from each fraction were mapped back to the assembled contigs using Bowtie2 and mean coverage was calculated using *genomeCoverageBed* from the BedTools package [S11]. Linkage information was then generated and served as an input along with the mean coverage data to the CONCOCT program. The resulting bin clusters were visualized in an interactive PCA plot (using a modification of *ClusterPlot.R* from the CONCOCT package). Each bin was taxonomically annotated using BLASTn against NCBI nr. The interactive PCA plot was then inspected manually and fastq files of close taxonomical groups were extracted and assembled separately with IDBA-ud. These assemblies were later pooled together with the reset of the assemblies and were de-replicated with VSEARCH. After using this assembly process for each fraction (vDNA, gDNA and cDNA) and for a pool of all fractions, the assemblies were de-replicated with VSEARCH and merged with the AMOS suite program minimus2 [S12].

Reads from each sample were mapped back to the final contigs with BWA-MEM [S13] and then binned using MetaBAT [S14], with the “superspecific” option. The resulting bins were then analyzed with CheckM [S15]. Bins were assigned taxonomy with Diamond [S16] and blast2lca from the MEGAN5 package [S17] [https://ab.inf.uni-tuebingen.de/software/megan5]. Bins were then inspected manually.

Re-assembly of *Tara* Oceans data

The *Tara* Oceans microbiome [S18] and virome [S19] data sets were assembled using the IDBA-UD [S8] assembler providing higher quantity of longer scaffolds than previously reported [S18]. Errors in the assembly were corrected using two read-mapping based in-house tools (Sharon *et al*., *in prep*). The initial fragment from group D was extended using mini-assembly technique as described in [S20]. This process lead to the recruitment of other fragments of the same genome until no further elongation could be done.

Mapping of reads back to assembled contigs

Following assembly, the reads from each sample were mapped back to assembled contigs using bowtie2 [S21]. PE Reads were counted as mapped (+1) if both end mapped perfectly to the same contig or if only one of the ends mapped to a contigs. For visualization, BEDTOOLS [S11] was used to generate bedgraph files from raw BAM files and plots of read coverage depth were created with the R package SUSHI [S22].

Abundance of Magroviruses in environmental metagenomic samples

A reference collection composed by 257 FASTA sequences from magroviruses, MG-II, archaeal, cyanobacterial, SAR11, and SAR116 hosts and viral genomes or contigs (available on github) was used to estimate and compare the distribution and abundance of the magroviruses.

Paired-end Illumina raw reads from the Red-Sea samples (this report), *Tara* Oceans expedition microbiome [S18] and virome [S19] data sets (Accession: EBI-ENA: PRJEB402), were mapped on each one of the reference sequences using Bowtie2 [S21] version 2.2.6 with settings “-a --very-sensitive-local”. The alignments were converted to BAM format using SAMTools [S23] and stored.

The BAM files were processed using custom python scripts based on HTSeq-count [S24] namely a parser (bamParser.py) that filters the alignments according to the percent-identity of each read to the sequence used as reference, with a cutoff value of 90% identity over an alignment length of at least 80 nt, and readCounter.py which counts how many reads mapped to a reference sequence by comparing the alignments' percent-identity similarity and Bowtie2 alignment score of the mapped alignments.

The counts were normalized by calculating genome fragments per kilobase reference sequence per million library reads (GFPM) as described in the Jupyter Notebooks deposited in the Github repository.

Metagenomic database screening

The MCP and DNAp sequences of the Magrovirus MAGs detected in Red Sea metagenomes were used as queries in a PSI-BLAST search [S25] of the Whole Genome Shotgun database (WGS) at the NCBI [S26]. The collected metagenomic sequences were extended whenever possible using alternating cycles of BLASTN searches and assembly with Geneious ((www.geneious.com).

Protein sequence analysis

All MAGs, both from Red Sea and WGS, were translated by MetaGeneMark [S27]. The resulting protein sequences were used as queries to search the nr database using PSI-BLAST, the Conserved Domain Database (CDD) using RPS-BLAST [S28], and the CDD and Pfam databases using HHpred [S29].

Phylogenetic analysis

Magrovirus protein sequences were pooled with their respective best hits from the nr and/or CDD databases. Protein sequences were aligned using MUSCLE [S30]. Gapped columns (more than 30% of gaps) and columns with low information content were removed from the alignment [S31]. A preliminary tree was constructed using the FastTree program with default parameters [S32]. The final maximum likelihood tree was calculated using the PhyML program [S33], the latest version of which (http://www.atgc-montpellier.fr/phyml-sms/) includes automatic selection of the best-fit substitution model for a given alignment.

Supplemental References

S1. John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K., Kern, S., Brum, J., Polz, M.F., Boyle, E.A., and Sullivan, M.B. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ Microbiol Rep 3, 195-202.

S2. Thurber, R.V., Haynes, M., Breitbart, M., Wegley, L., and Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. Nature protocols 4, 470-483.

S3. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

S4. Crusoe, M.R., Alameldin, H.F., Awad, S., Boucher, E., Caldwell, A., Cartwright, R., Charbonneau, A., Constantinides, B., Edvenson, G., Fay, S., et al. (2015). The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res 4, 900.

S5. Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., and Brom, T.H. (2012). A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv preprint arXiv:1203.4802.

S6. Pell, J., Hintze, A., Canino-Koning, R., Howe, A., Tiedje, J.M., and Brown, C.T. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci U S A 109, 13272-13277.

S7. Wood, D.E., and Salzberg, S.L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome biology 15, R46.

S8. Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420-1428.

S9. Zerbino, D.R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 18, 821-829.

S10. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nature methods 11, 1144-1146.

S11. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.

S12. Treangen, T.J., Sommer, D.D., Angly, F.E., Koren, S., and Pop, M. (2011). Next generation sequence assembly with AMOS. Current protocols in bioinformatics Chapter 11, Unit 11 18.

S13. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997v2.

S14. Kang, D.D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165.

S15. 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 research 25, 1043-1055.

S16. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature methods 12, 59-60.

S17. Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome research 17, 377-386.

S18. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. Science 348, 1261359.

S19. Brum, J., Ignacio-Espinoza, J.C., Roux, S., Doulcier, G., Acinas, S.G., Alberti, A., Chaffron, S., Cruaud, C., de Vargas, C., Gasol, J.M., et al. (2015). Patterns and ecological drivers of ocean viral communities. Science 348, 1261498.

S20. Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome research 23, 111-120.

S21. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods 9, 357-359.

S22. Phanstiel, D.H., Boyle, A.P., Araya, C.L., and Snyder, M.P. (2014). Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics 30, 2808-2810.

S23. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

S24. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31, 166-169.

S25. 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 research 25, 3389-3402.

S26. Coordinators, N.R. (2015). Database resources of the National Center for Biotechnology Information. Nucleic acids research 43, D6-17.

S27. Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic acids research 38, e132.

S28. Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., et al. (2013). CDD: conserved domains and protein three-dimensional structure. Nucleic acids research 41, D348-352.

S29. Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.

S30. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32, 1792-1797.

S31. Yutin, N., Makarova, K.S., Mekhedov, S.L., Wolf, Y.I., and Koonin, E.V. (2008). The deep archaeal roots of eukaryotes. Molecular biology and evolution 25, 1619-1630.

S32. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS ONE 5, e9490.

S33. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate Maximum-Likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59, 307-321.