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

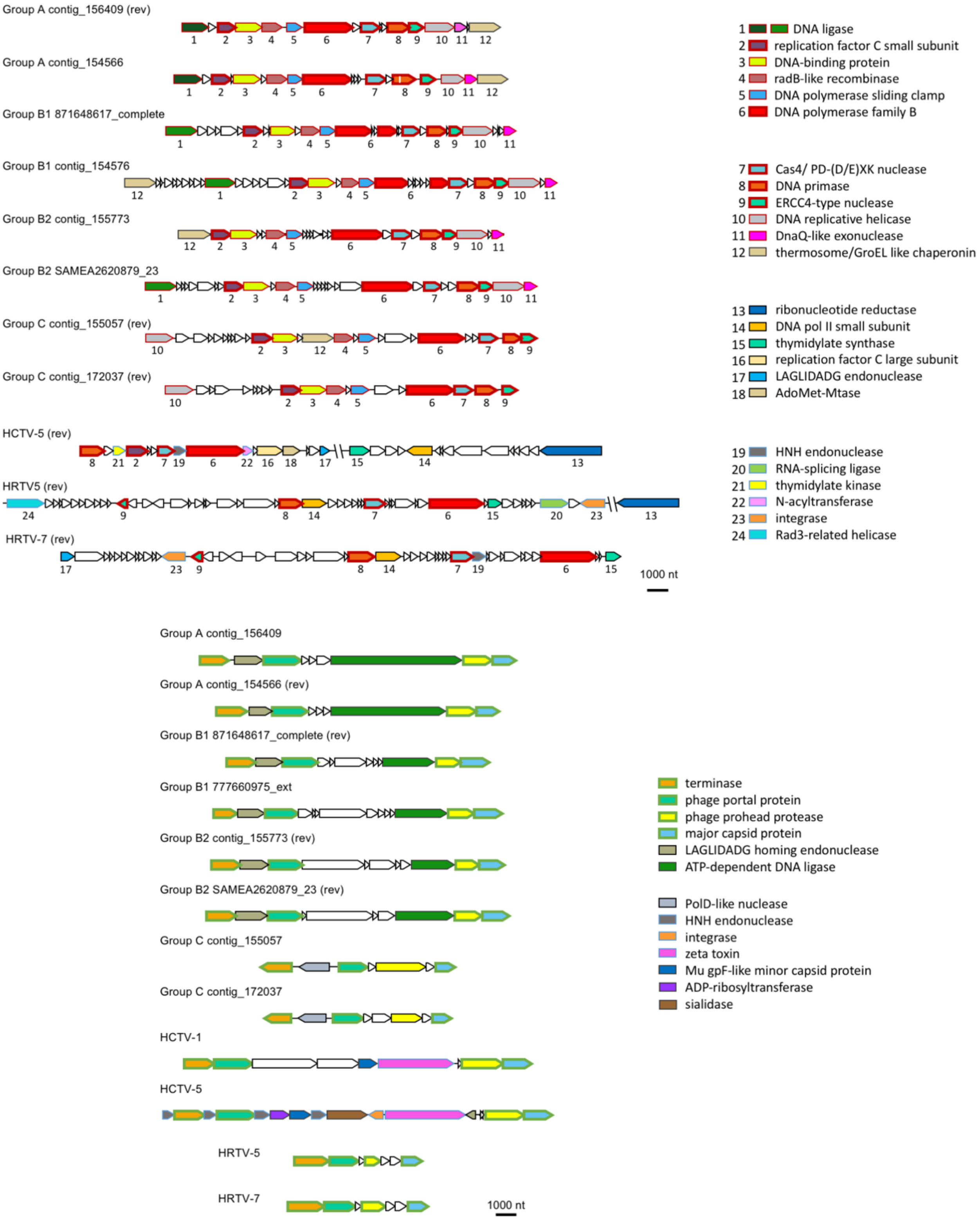
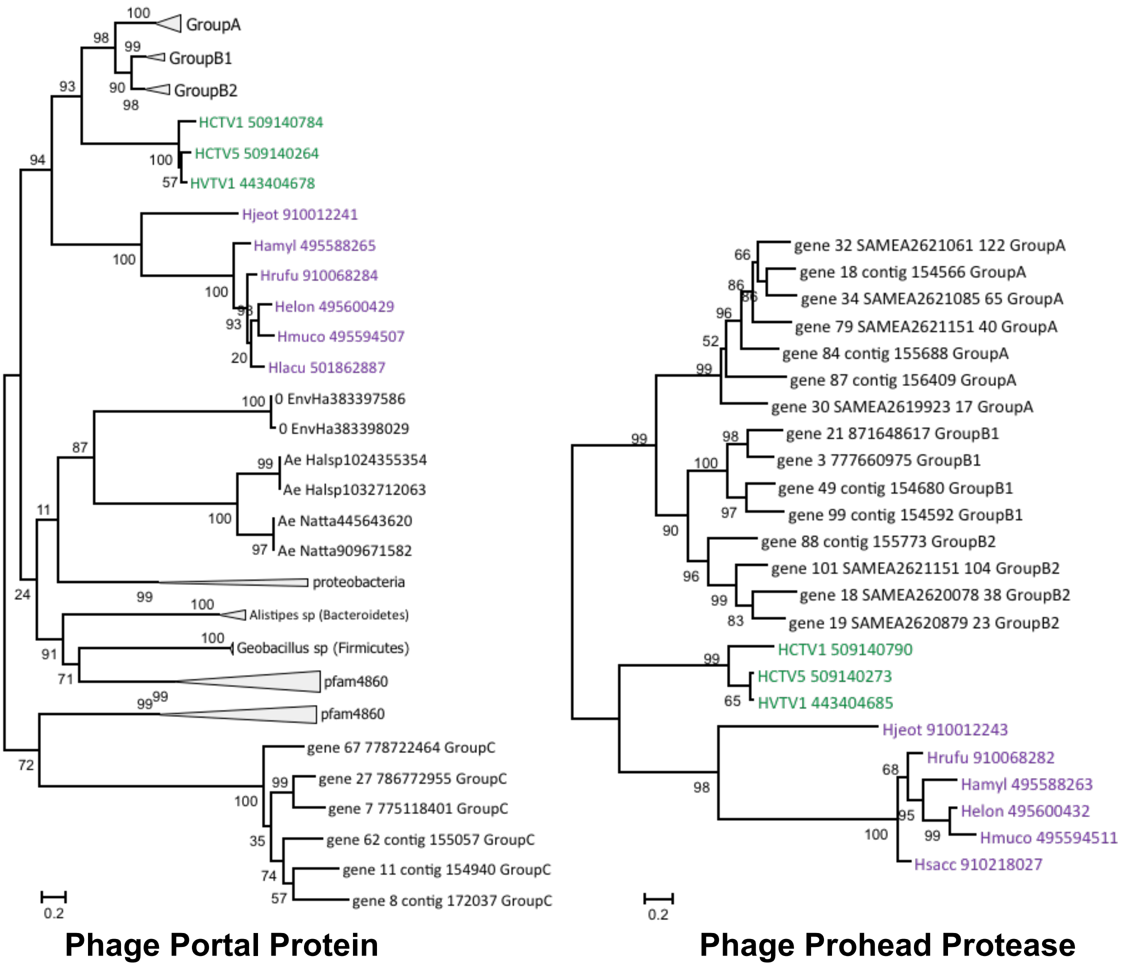


Figure S1: Detailed genomic maps of Magroviruses and haloviruses (related to Figure 2)

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.

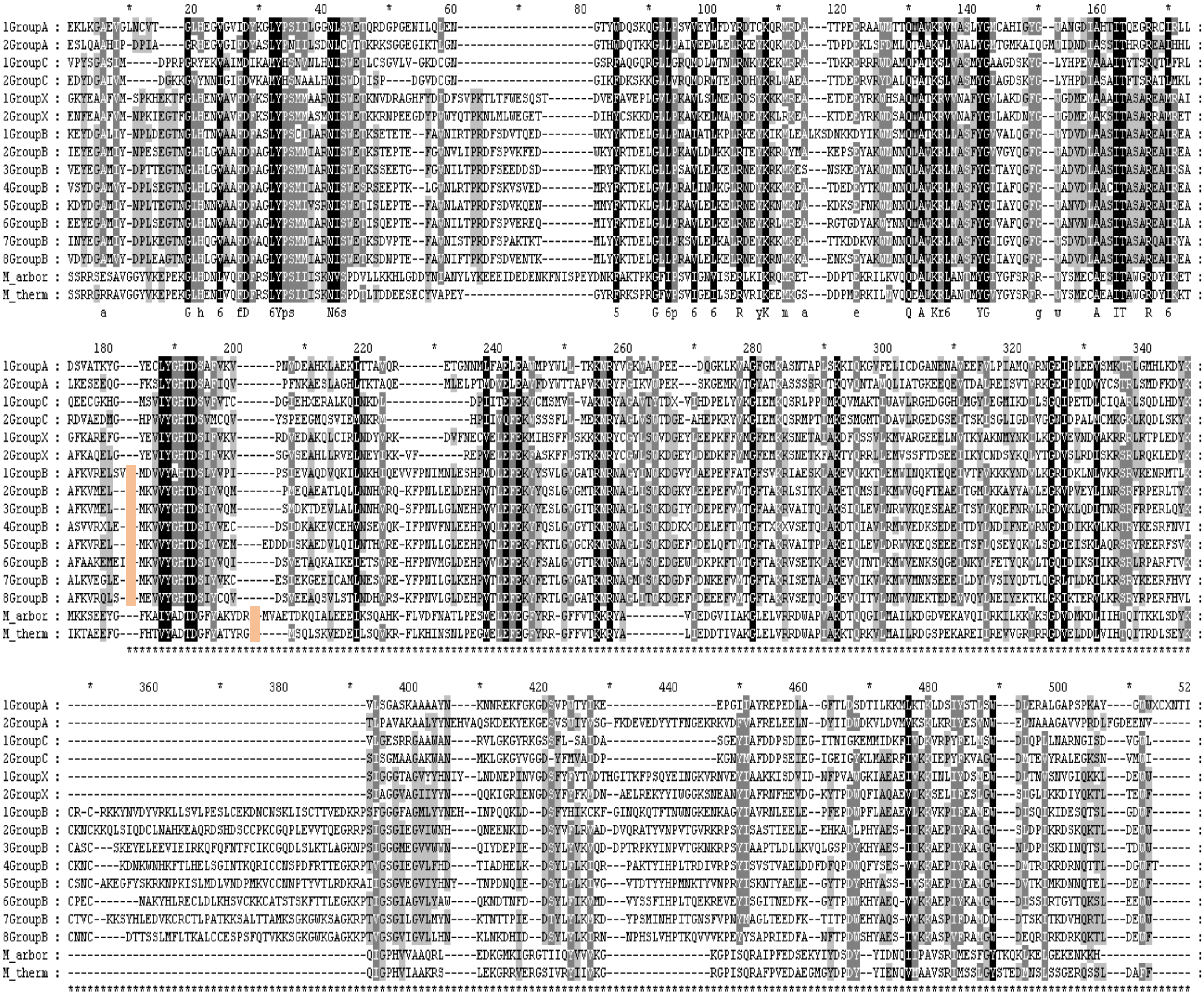


Figure S3: Multiple Sequence alignment of the DNA polymerase family B catalytic domain (related to Figure 1B)

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.