Supplementary Information for

Diversity and distribution of viruses inhabiting the deepest ocean on Earth

Huahua Jian[1,2†], Yi Yi[1†], Jiahua Wang[1†], Yali Hao[1], Mujie Zhang[1], Siyuan Wang[1], Canxing Meng[1], Yue Zhang[1], Hongmei Jing[3], Yinzhao Wang[1,2], Xiang Xiao[1,2*]

[1]State Key Laboratory of Microbial Metabolism, Joint International Research Laboratory of Metabolic & Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China

[2]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, China

[3]CAS Key Laboratory for Experimental Study under Deep-sea Extreme Conditions, Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences, Sanya, China

*Xiang Xiao, Email: zjxiao2018@sjtu.edu.cn

[†]These authors contributed equally to this work.

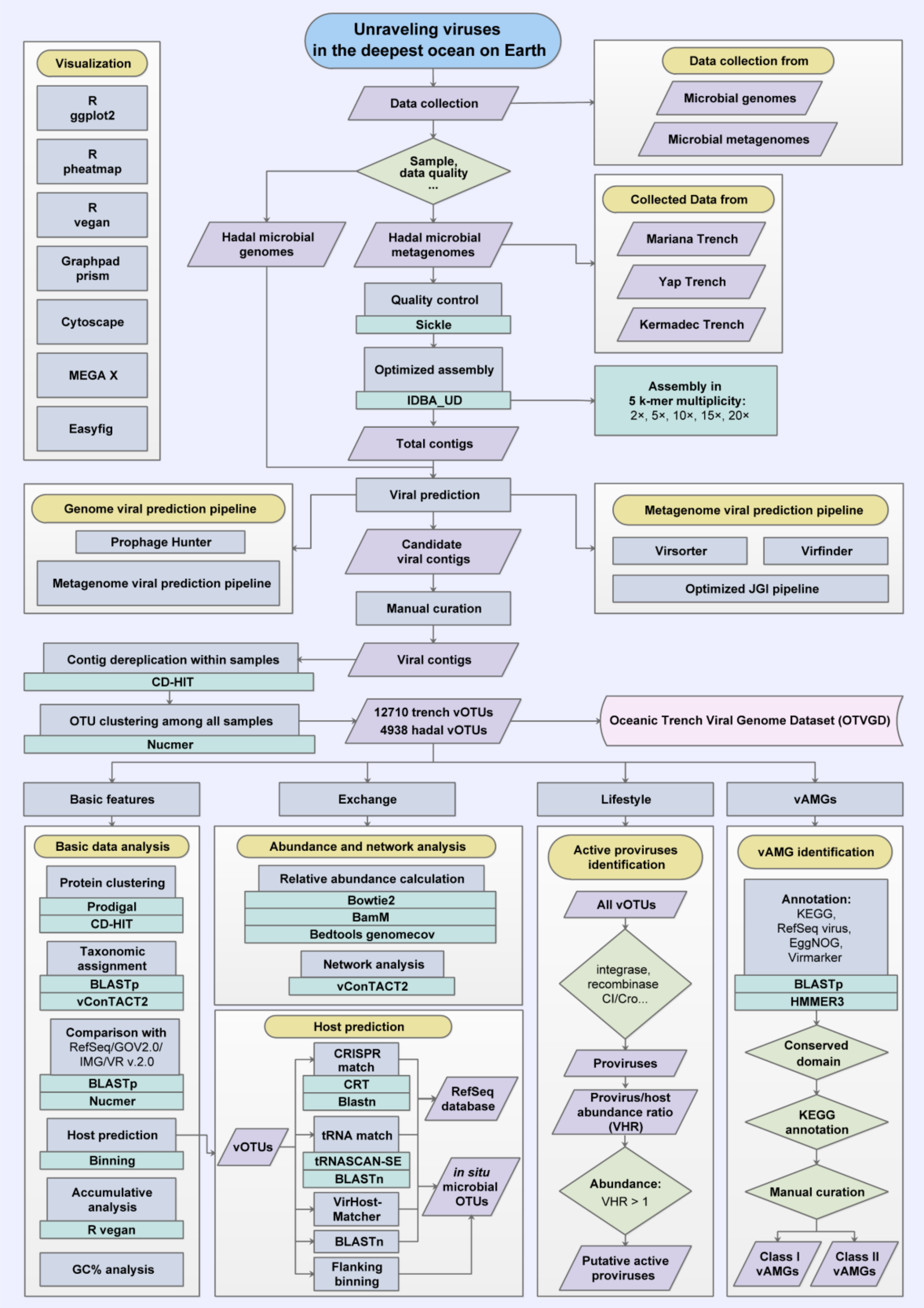


Fig. S1. Overview of the bioinformatic workflow. See the Methods for detailed information.

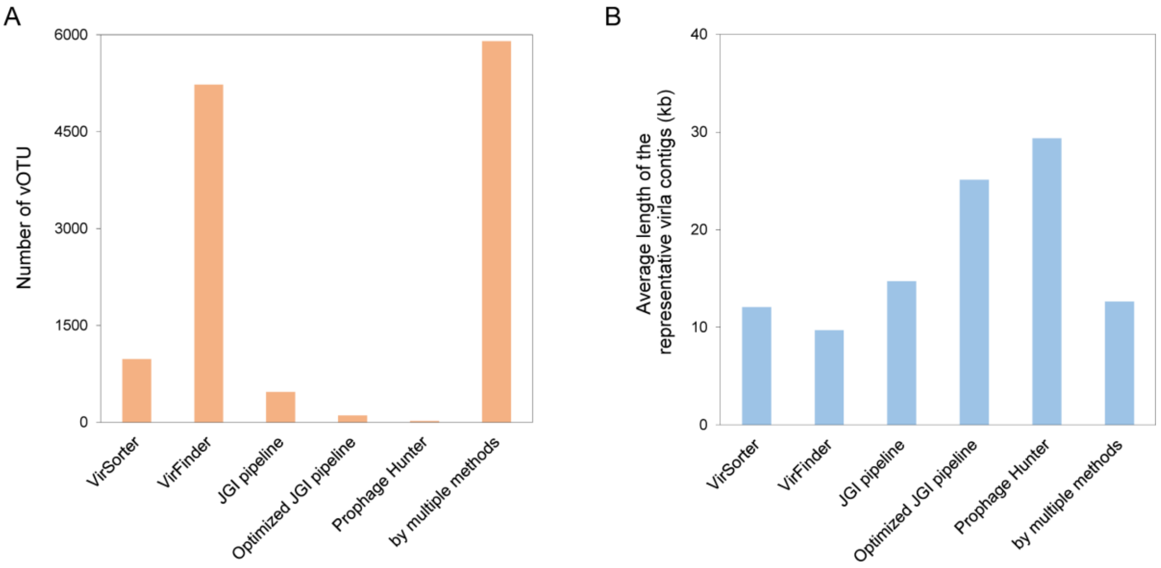


Fig. S2. Evaluation of viral prediction approaches. The number of vOTUs (A) and the average length of the representative viral contigs (B) that were specifically recovered by different prediction approaches.

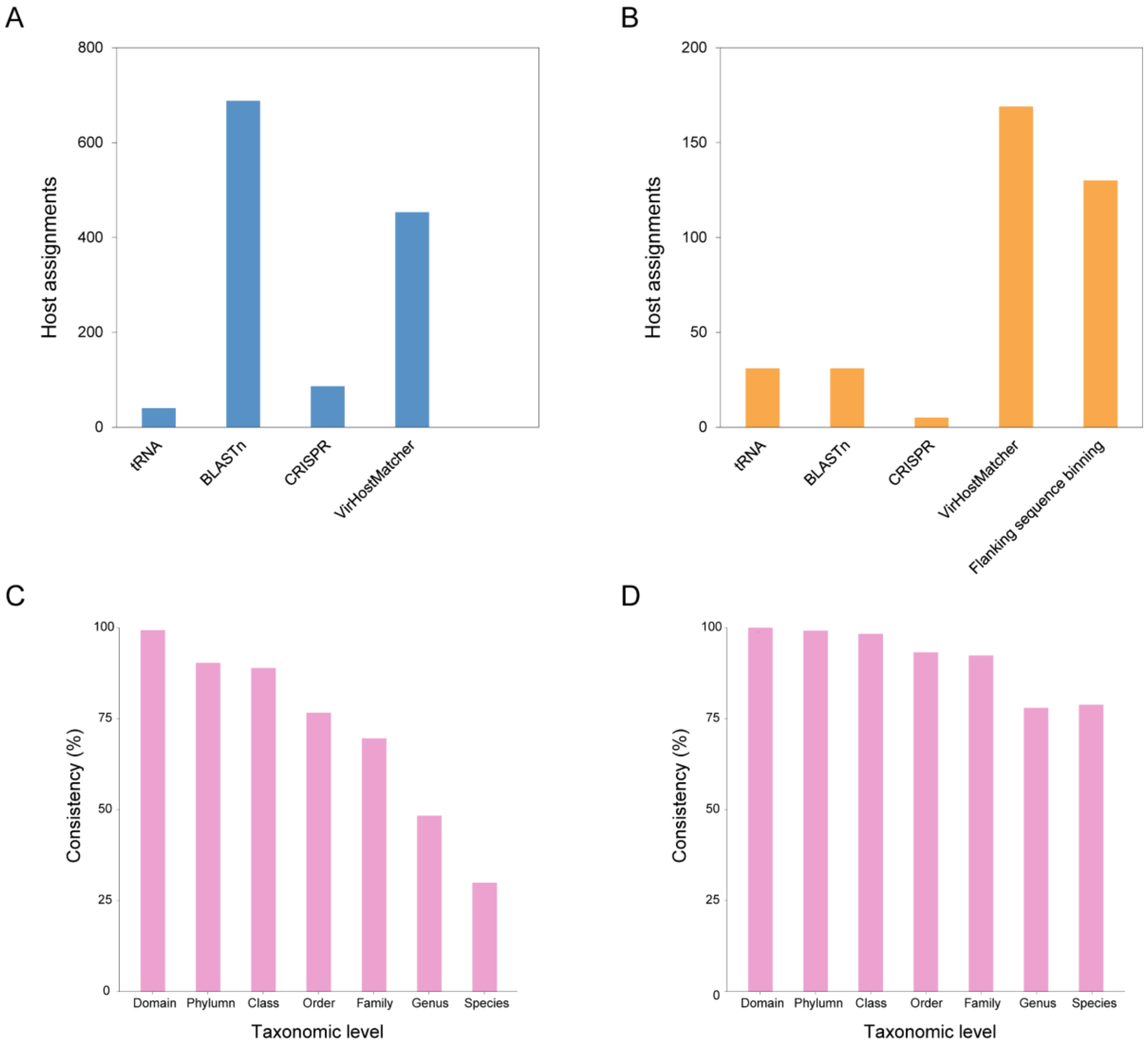


Fig. S3. Evaluation of viral host predictions. Numbers of host assignments according to the "*ex situ*" (A) and "*in situ*" (B) pipelines by different prediction approaches. (C) Consistency between different approaches at different taxonomic levels for vOTUs with host assignments by multiple prediction methods (*n*=424). (D) Consistency between different approaches at different taxonomic levels for vOTUs with the "*in situ*" host assignments by multiple prediction methods (*n*=118). For each taxonomic level, only if 100% of the assigned host taxa were identical, the assignment was regarded as consistent.

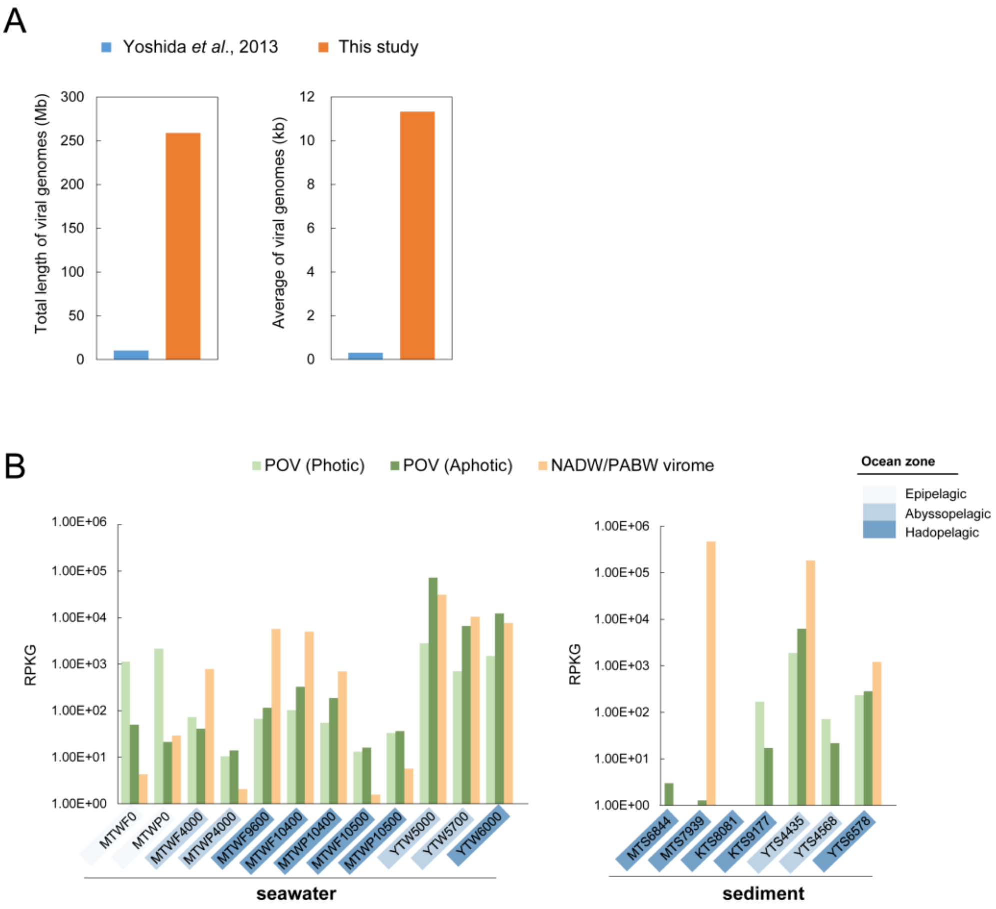


Fig. S4. (A) Total and average lengths of viral genomes of OTVGD compared with those of the previously published oceanic trench virome. (B) Read recruitment of OTVGD against previously published viromes from the Pacific Ocean (1, 2). Reads hits with an e-value < 10^{-5}, identity > 95%, and length > 50 bp were used to calculate reads recruited per kb of genome per Gb of metagenome (RPKG).

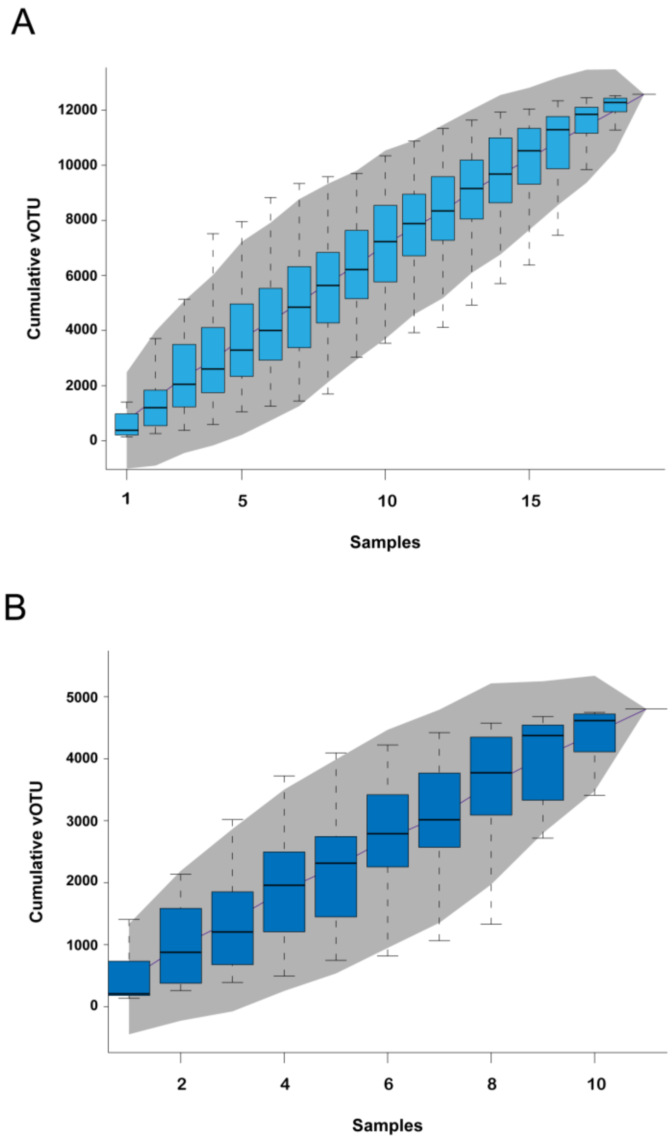


Fig. S5. Viral OTU accumulation curves for the OTVGD (A) and hadal trench viral genome dataset (HTVGD) (B). The black lines within the box indicate the median value of the vOTU number, and the ranges of the error bars represent 100 random replicates.

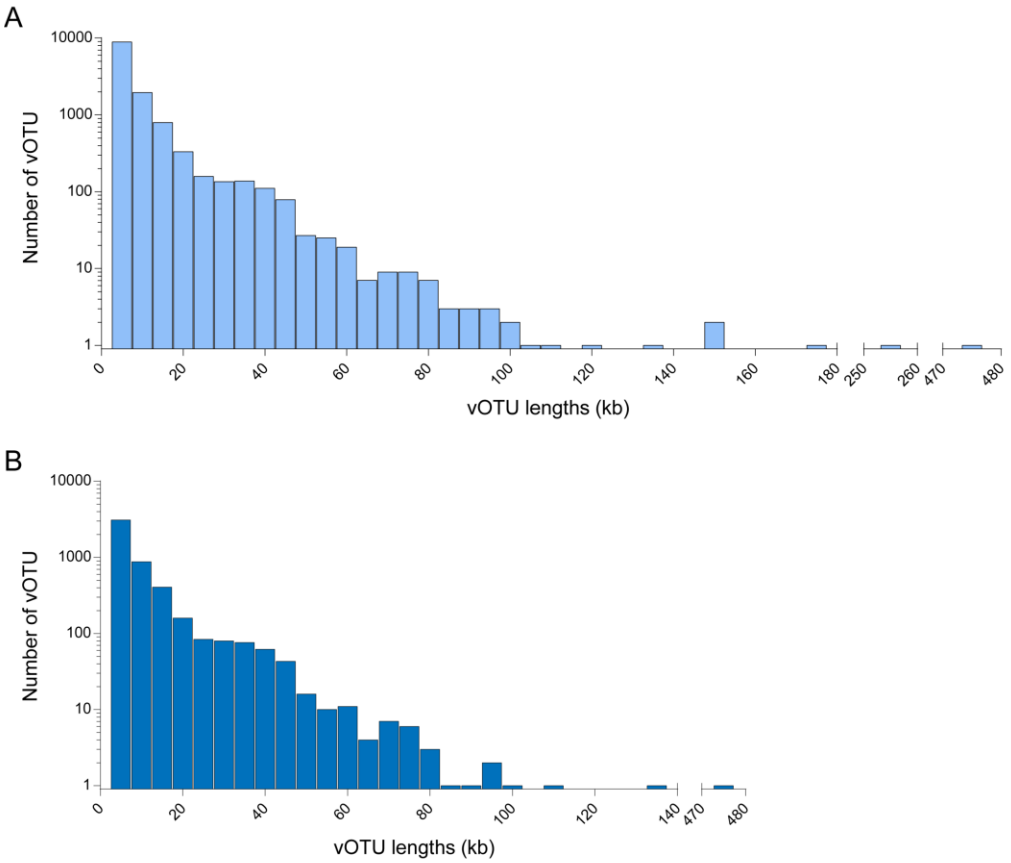


Fig. S6. Genome size distribution in the vOTUs of OTVGD (A) and HTVGD (B). The x-axis was truncated to increase the clarity of the graphs.

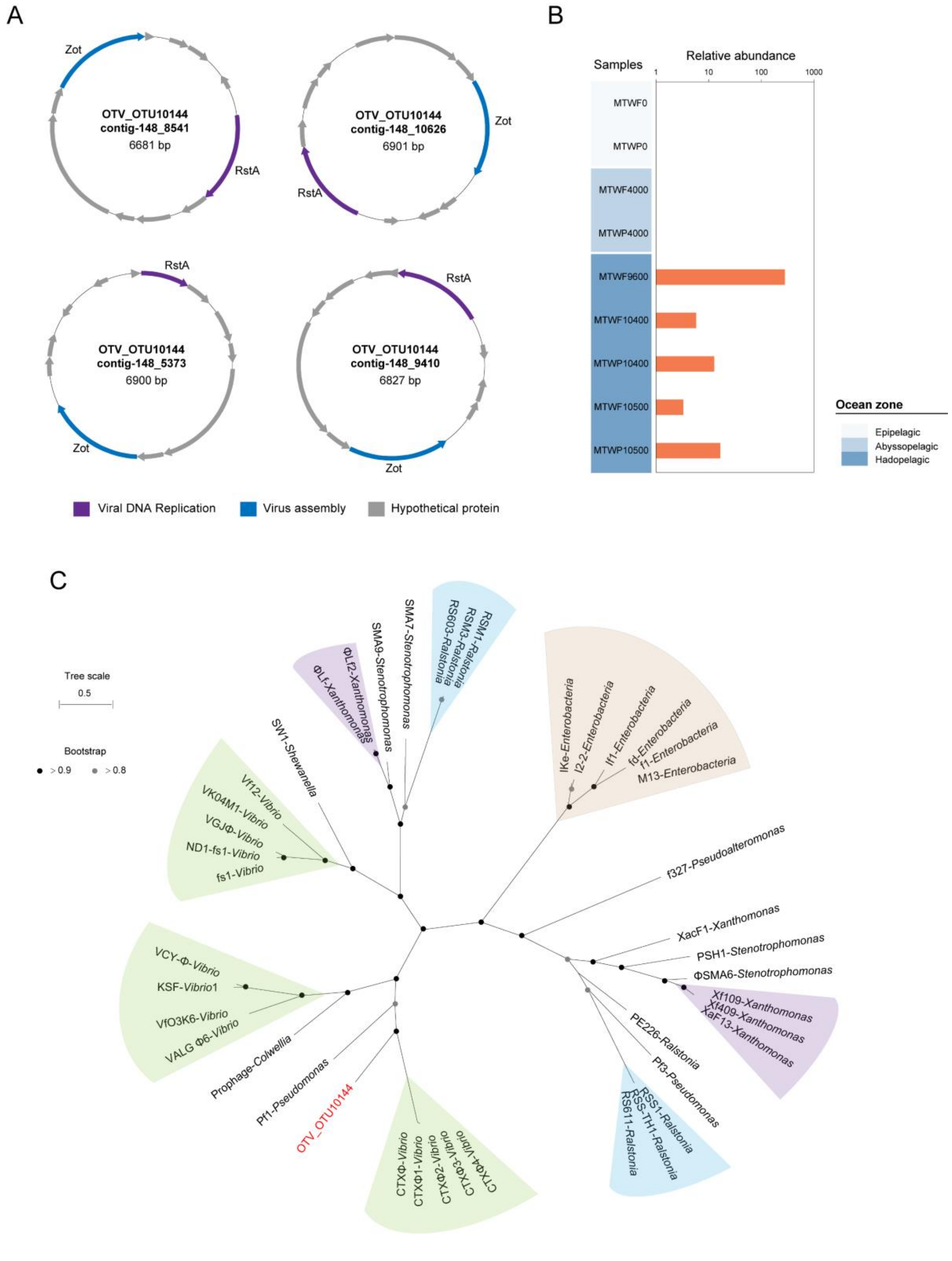


Fig. S7. Hadal vOTUs belonging to inoviruses. (A) Genomic maps depicting predicted proteins encoded by representative hadal inoviruses. The arrows depict the location and direction of predicted proteins in the viral genomes, and the fill colours indicate different functional categories

of genes, as indicated in the legend. (B) Relative abundance of hadal inoviruses in different ocean zones of the Mariana Trench. The relative abundance of one vOTU in a sample was based on the recruitment of reads to the representative vOTU contig and was considered only if more than 75% of the reference contig was covered. (C) Phylogenetic tree of inoviruses. An unrooted phylogenetic tree was built from the conserved Zot-like proteins of known inoviruses using the maximum-likelihood method with 1000 bootstraps. Nodes with bootstrap support values greater than 0.9 and 0.8 are marked with black and grey circles, respectively. Clades of Zot proteins from the same bacteria genus are assigned the same background colours. The Zot protein of hadal inovirus obtained from this study is highlighted in red.

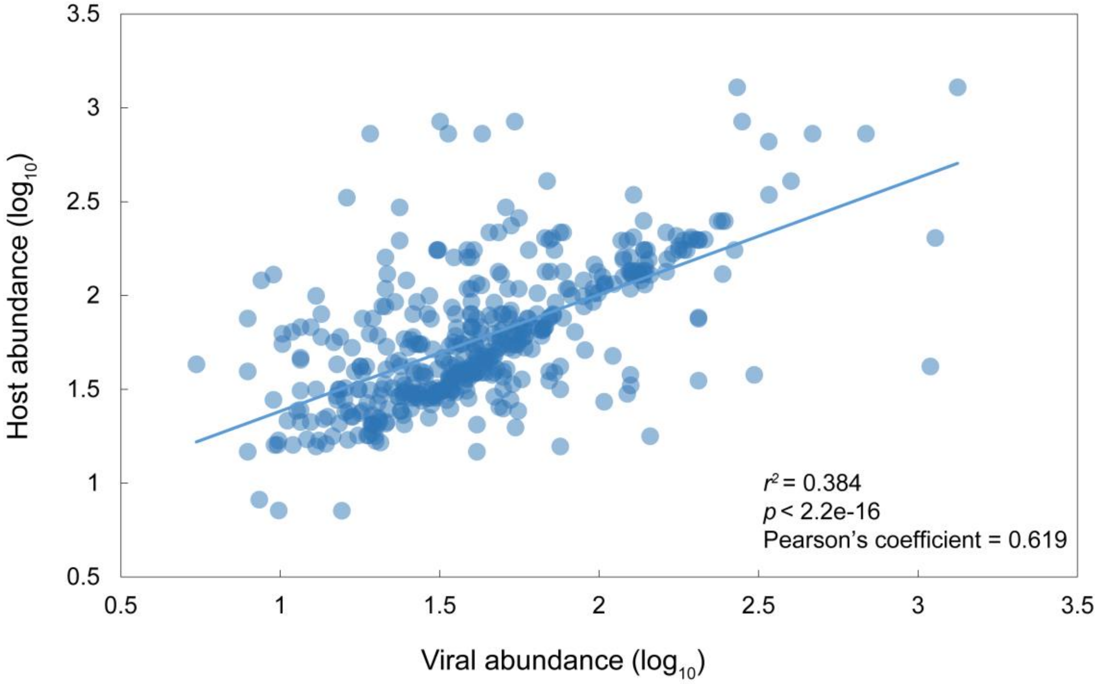


Fig. S8. Virus–host abundance patterns in the OTVGD. Viral abundance and predicted host abundance (*n*=457) (both calculated as the mean coverage depth from metagenomic read mapping, normalized by the number of reads in the sample) are shown. Based on linear regression analysis, best-fit lines and adjusted r^2 values are presented. The *p* value of the F-test indicates whether the use of an interaction term in the linear regression models yielded a significantly different result from no interaction term. Pearson's correlation coefficient is also presented. The statistical analysis was performed by using R 4.0.0.

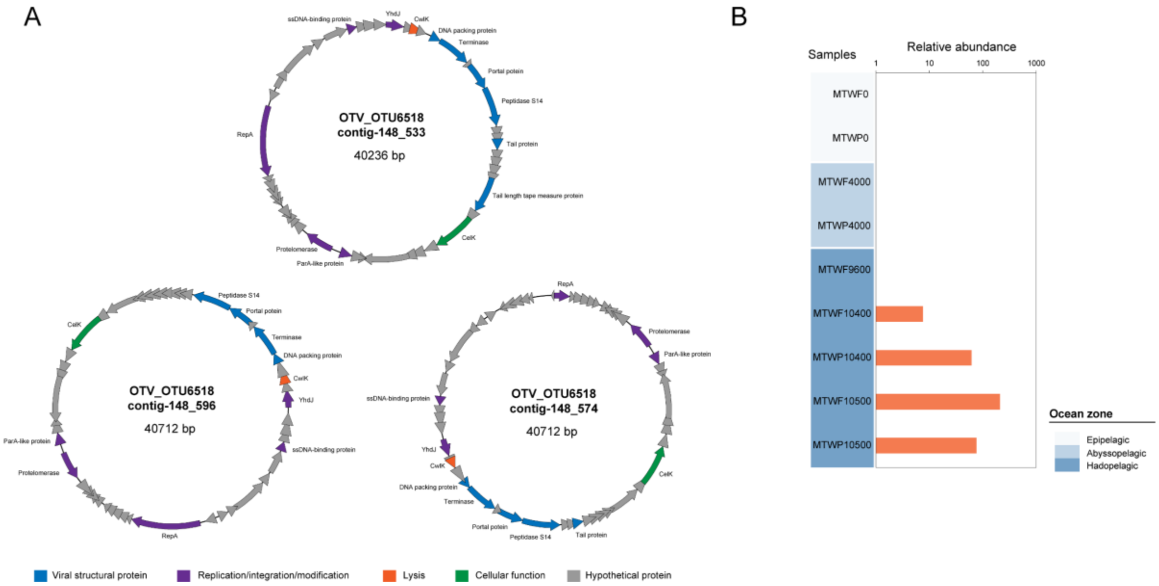


Fig. S9. Putative *Oleibacter* viruses in the hadal seawater of the Mariana Trench. (A) Genomic maps depicting predicted proteins encoded on representative *Oleibacter* vOTUs. The arrows depict the location and direction of predicted proteins on the viral genomes, and the fill colours indicate different functional categories of genes, as indicated in the legend. (B) Relative abundances of *Oleibacter* vOTUs in different ocean zones of the Mariana Trench. The relative abundance of one vOTU in a sample was based on the recruitment of reads to the representative vOTU contig and was considered only if more than 75% of the reference contig was covered.

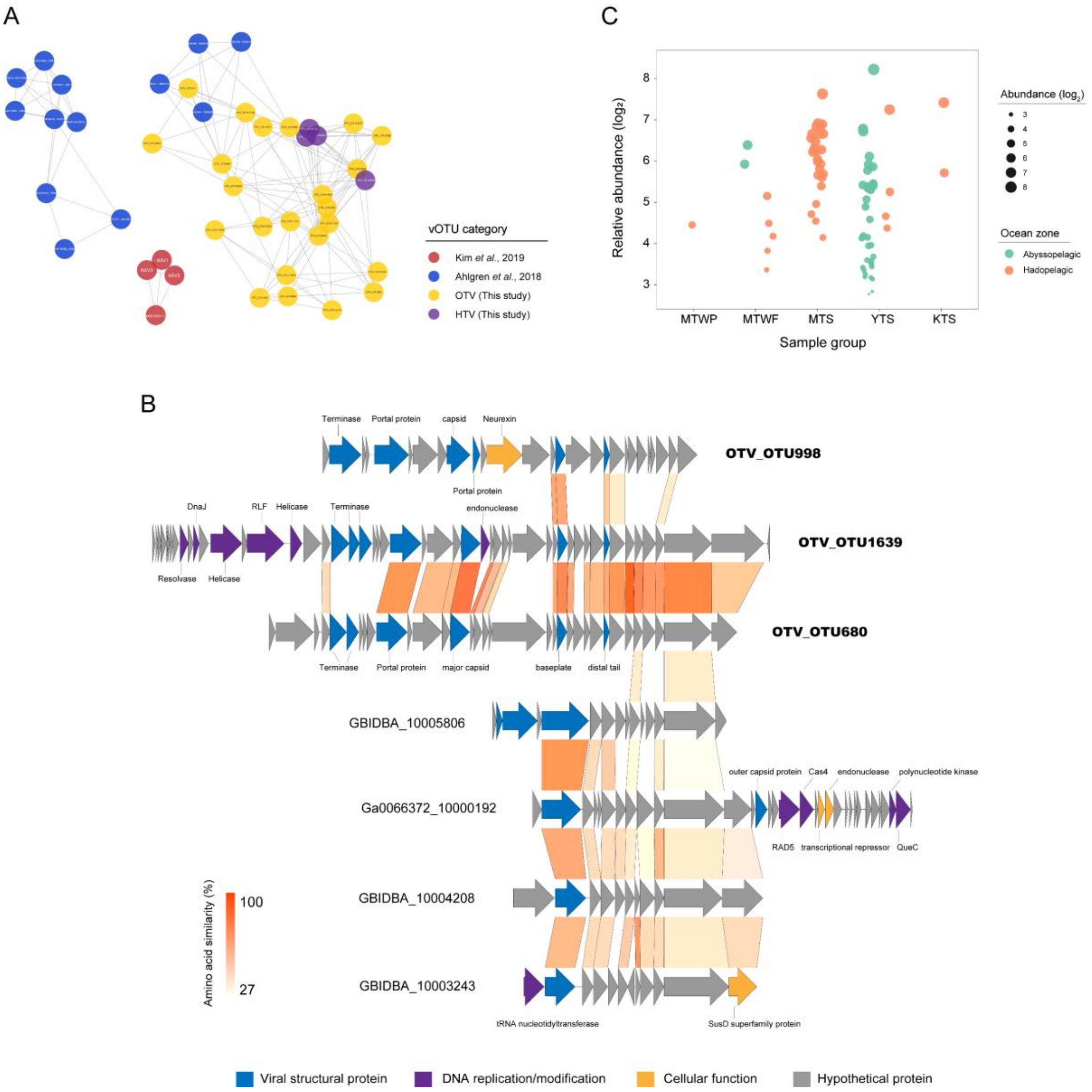


Fig. S10. Putative Thaumarchaeota viruses in the OTVGD. (**a**) Protein-sharing network of vOTUs belonging to Thaumarchaeota viruses using vConTACT v2.0. The nodes represent vOTUs, and the connecting edges indicate significant protein sharing among them. Each node is depicted in a different colour, representing vOTUs from different studies, as indicated in the legend. The four isolated viruses that infect *Nitrosopumilus* are depicted in red. (**b**) Genomic maps depicting predicted proteins encoded on representative thaumarchaeal vOTUs. The arrows depict the location and direction of predicted proteins on the viral genomes, and the fill colours indicate different gene functional categories, as indicated in the legend. The annotations were based on searches against NCBI's nr protein database or HHpred analyses using standard settings, and only significant results (e-value < 1e^{-5}) were considered. The names of vOTUs identified in this study are indicated in bold black text. (**c**) Relative abundance of thaumarchaeal vOTUs in

different samples from the abyssal and hadal zones. The relative abundance of one vOTU in a sample was based on the recruitment of reads to the representative vOTU contig and was considered only if more than 75% of the reference contig was covered.

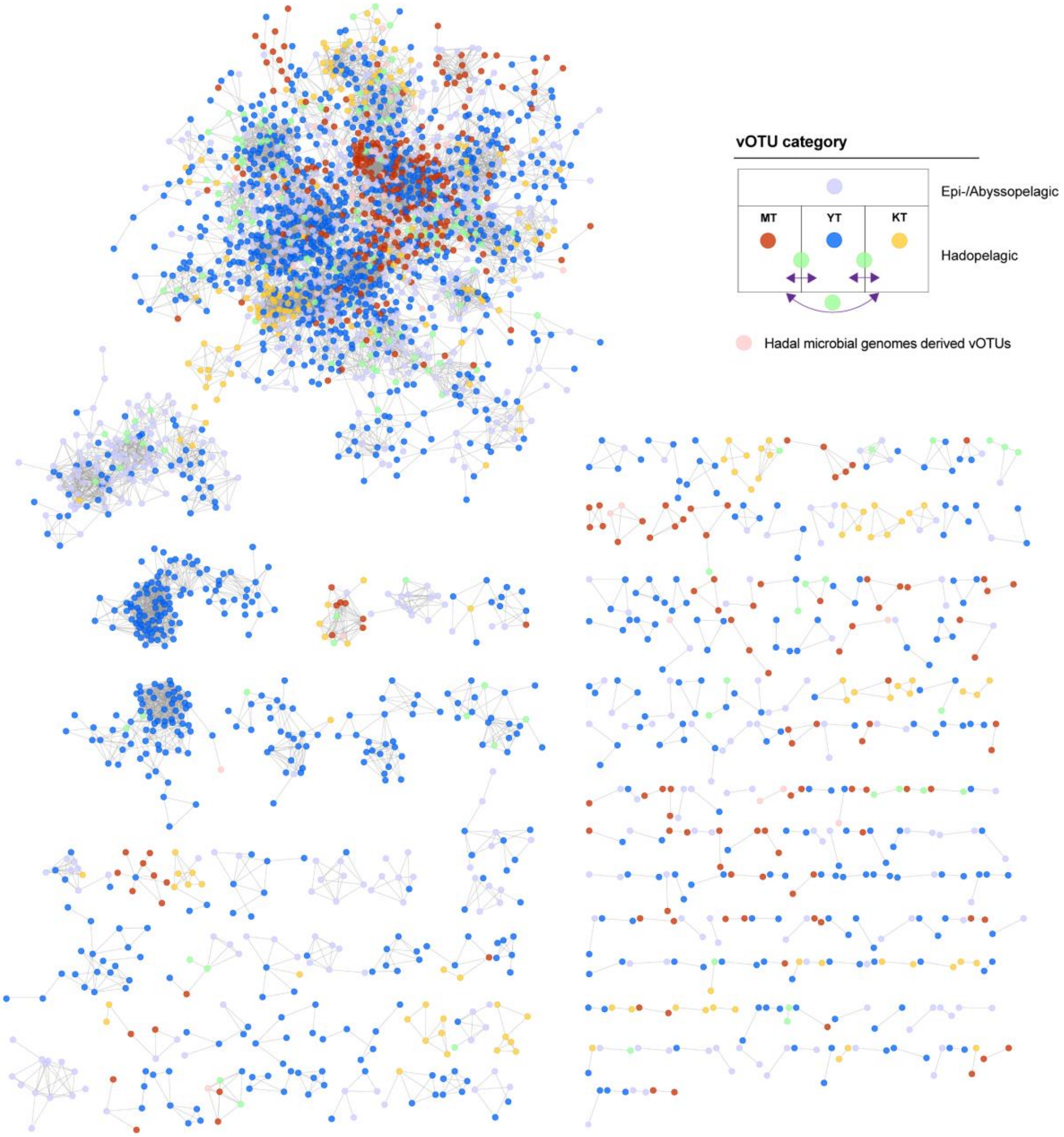


Fig. S11. The whole protein-sharing network of the vOTUs in OTVGD using vConTACT v2.0. The nodes and the connecting edges represent vOTUs their shared proteins, respectively. Nodes are depicted in the color representing vOTUs from the hadopelagic samples of the Mariana Trench (MT, red), Yap Trench (YT, blue), Kermadec Trench (KT, yellow), and epi-/abyssopelagic samples from these trenches (light blue). vOTUs present in multiple trenches, which suggest exchange of hadal viruses, are indicated by green nodes. Pink nodes correspond to the vOTUs that were derived from hadal microbial genomes.

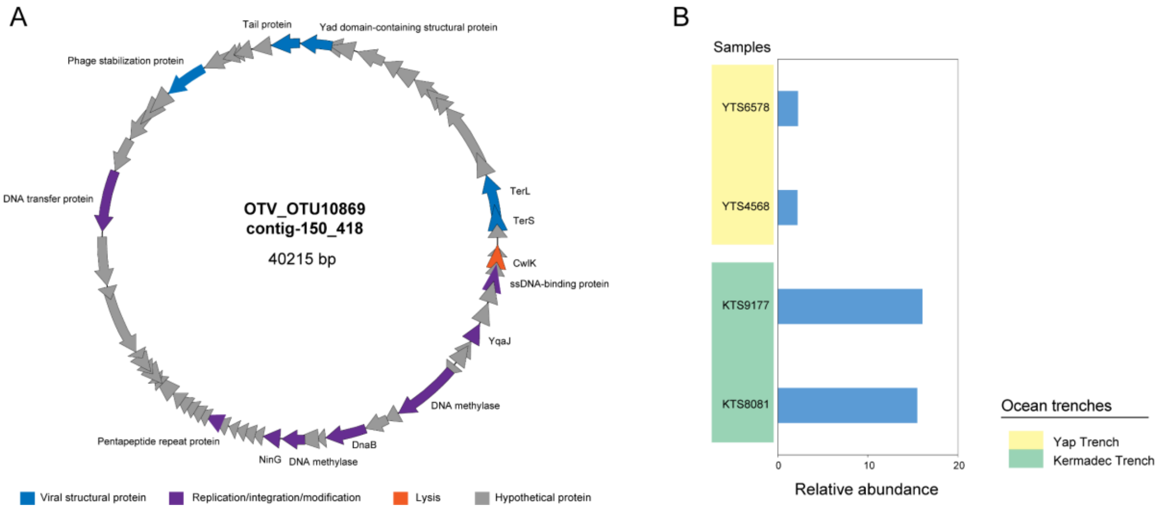


Fig. S12. Hadal vOTUs present in two different ocean trenches. (A) Genomic maps depicting predicted proteins encoded by representative vOTUs. The arrows depict the location and direction of the predicted proteins on the viral genomes, and the fill colours indicate different gene functional categories, as indicated in the legend. (B) Relative abundances of vOTUs in different trenches. The relative abundance of one vOTU in a sample was based on the recruitment of reads to the vOTU representative contig and was considered only if more than 75% of the reference contig was covered.

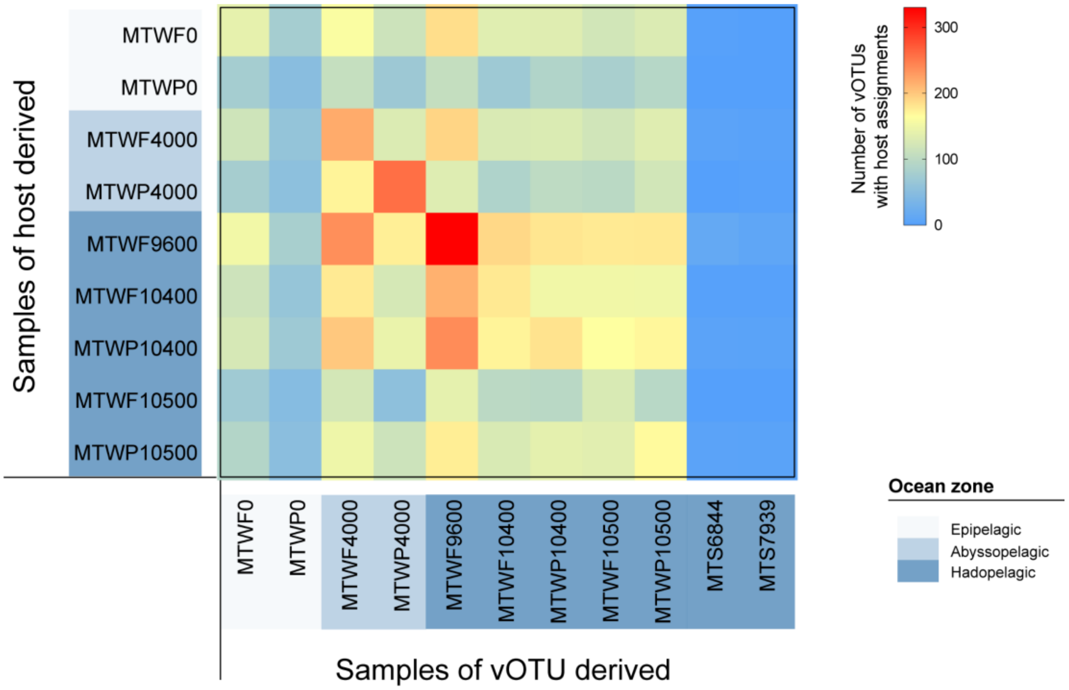


Fig. S13. Across-sample hosts of viruses in the Mariana Trench. The heatmap displays the number of vOTUs with host assignments from corresponding samples.

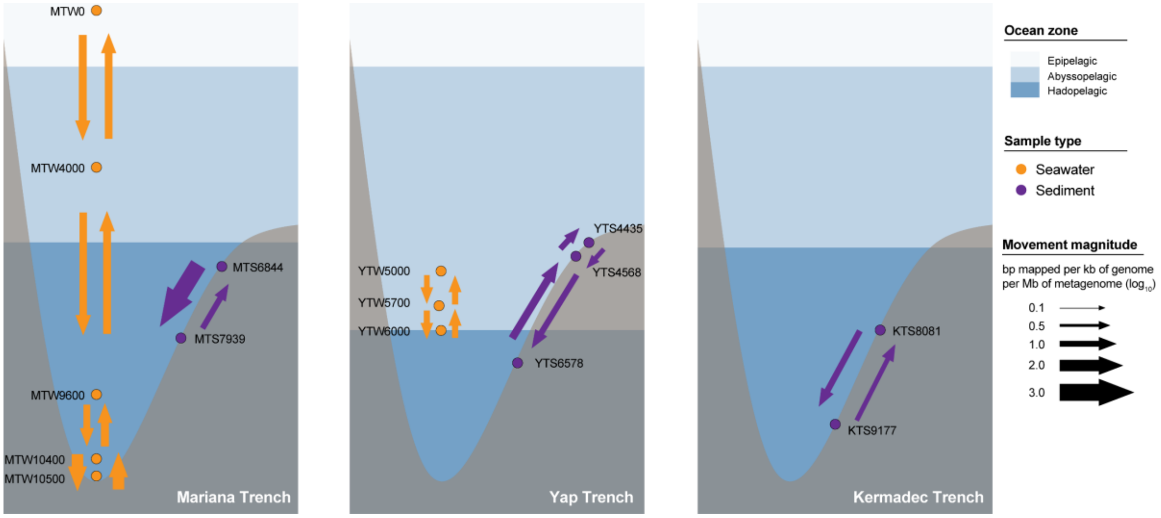


Fig. S14. Quantitative movement of vOTUs within oceanic trenches. Calculations were based on reciprocal comparison of vOTU abundances between neighbouring samples according to previously described methods (3). For each sample pair, the relative abundances of the vOTUs in one sample originating from a neighbouring sample, and *vice versa*, were calculated and are exhibited. The movement direction and magnitude are depicted with arrowheads and line widths, respectively.

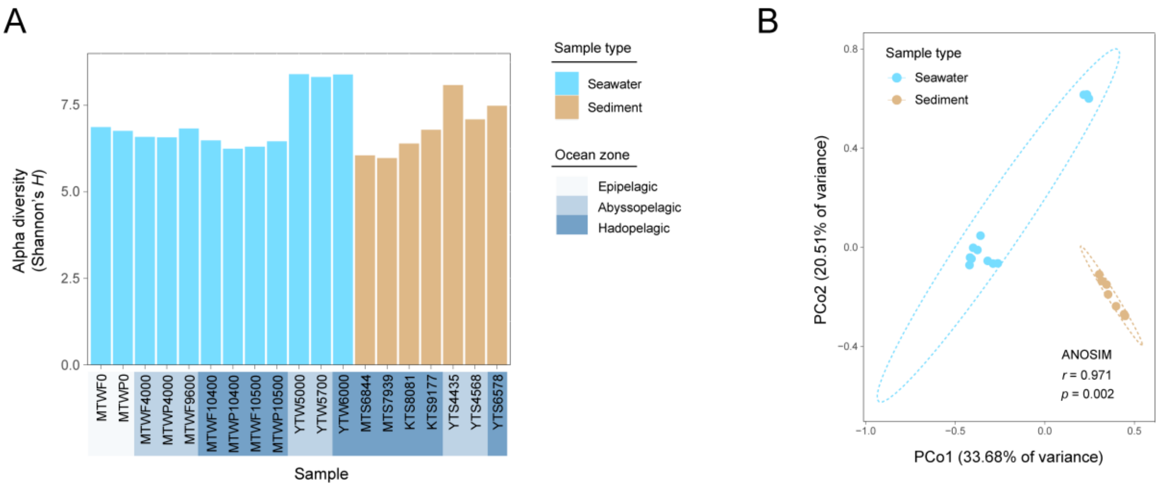


Fig. S15. Alpha and beta diversity analyses of oceanic viruses. (A) Shannon's *H* index of 19 trench samples. (B) Principal coordinates analysis (PCoA) based on a Bray-Curtis dissimilarity matrix calculation from relative abundances of all vOTUs of 19 trench samples. Samples were grouped by sample types in ANOSIM.

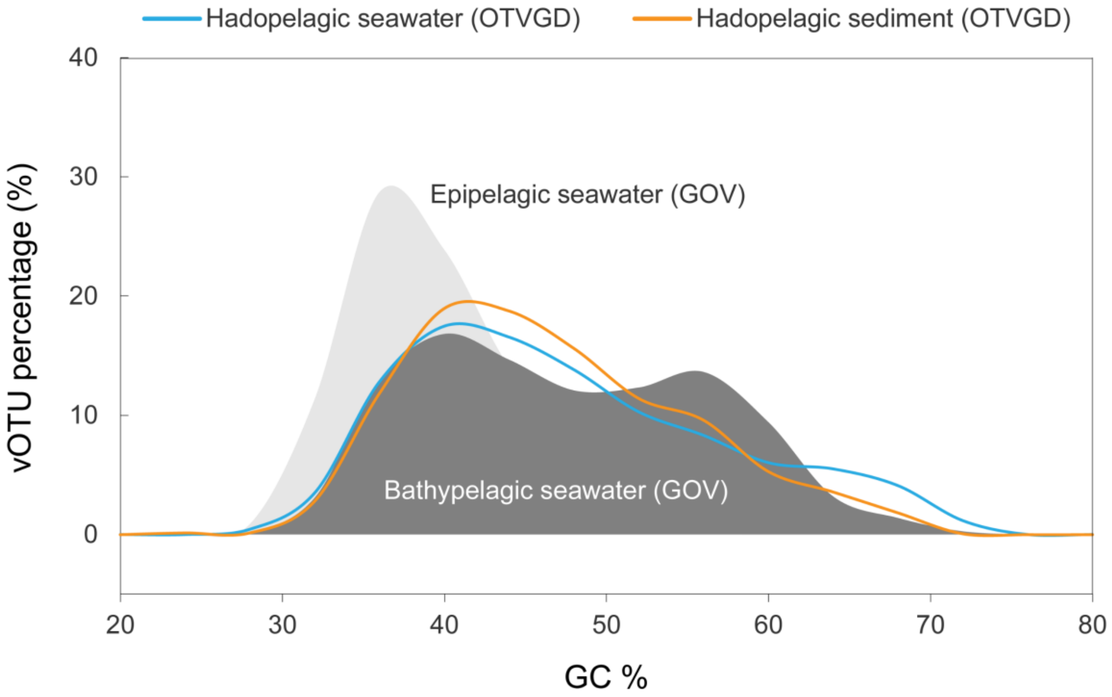


Fig. S16. GC plot of viral communities in different ocean zones. The blue and orange lines represent the GC plot of vOTUs derived from the hadopelagic seawater and sediment in this study, respectively, and the light and dark grey shading refer to the GC content patterns of vOTUs of GOV 2.0 from the epipelagic and bathypelagic zones, respectively (4).

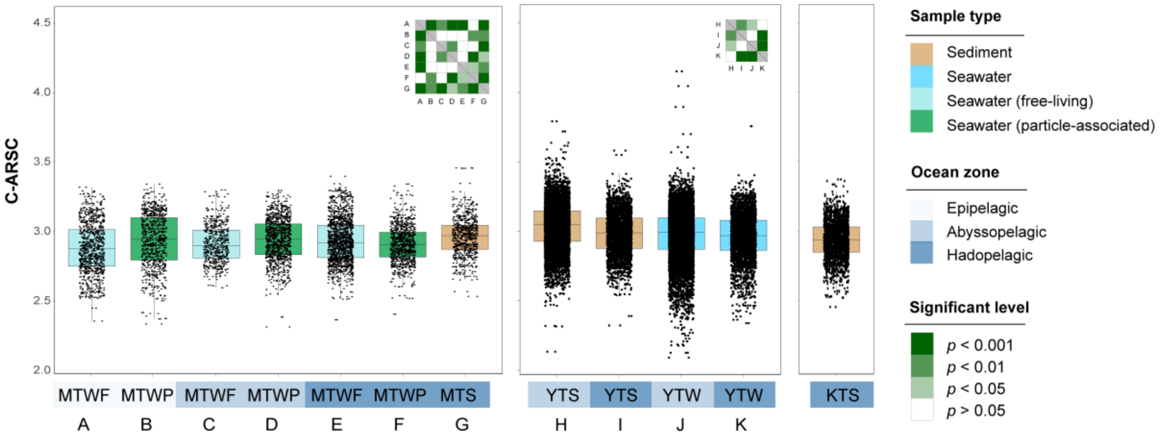


Fig. S17. Carbon atoms per residue side chain (C-ARSC) analysis of viral genomes in OTVGD. All representative viral contigs in each trench sample were used for the calculation. Heatmaps in the top right corner of frames show significance levels of differences between all pairs of sample groups calculated by the two-tailed Student's t test. The bottom letters representing sample groups correspond to coordinates of significance heatmaps.

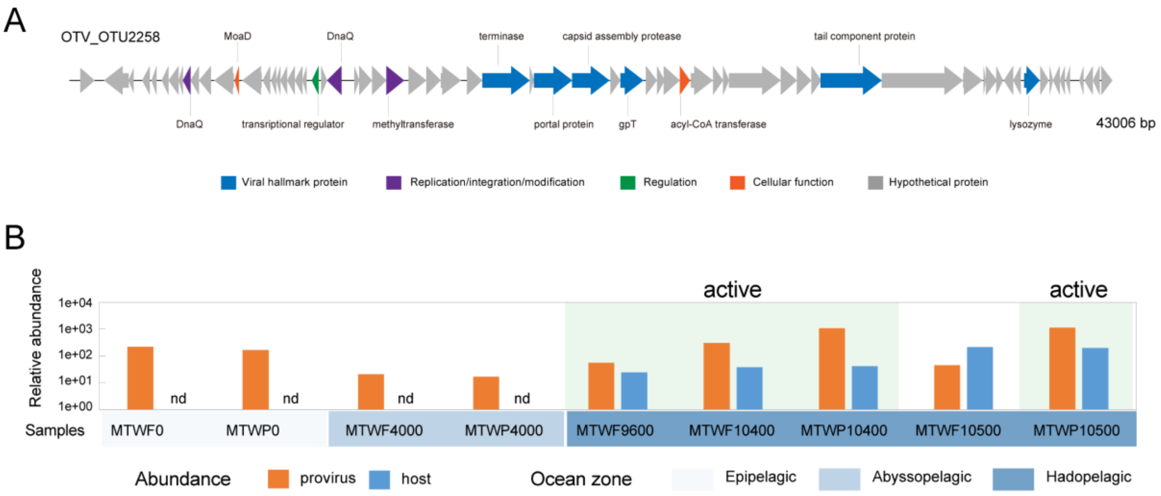


Fig. S18. Detailed gene contents (A) and abundance patterns (B) of the mostly active hadal proviruses. The arrows depict the location and direction of the predicted proteins in the viral genomes, and the fill colours indicate different gene functional categories, as indicated in the legend. Grey genes correspond to those with hypothetical or unknown functions. The provirus and host relative abundances across the Mariana Trench at different water depths, based on read mapping, are shown. The “active” and “latent” categories were defined by higher and lower abundances of each provirus than those of its host, respectively. nd, not determined.

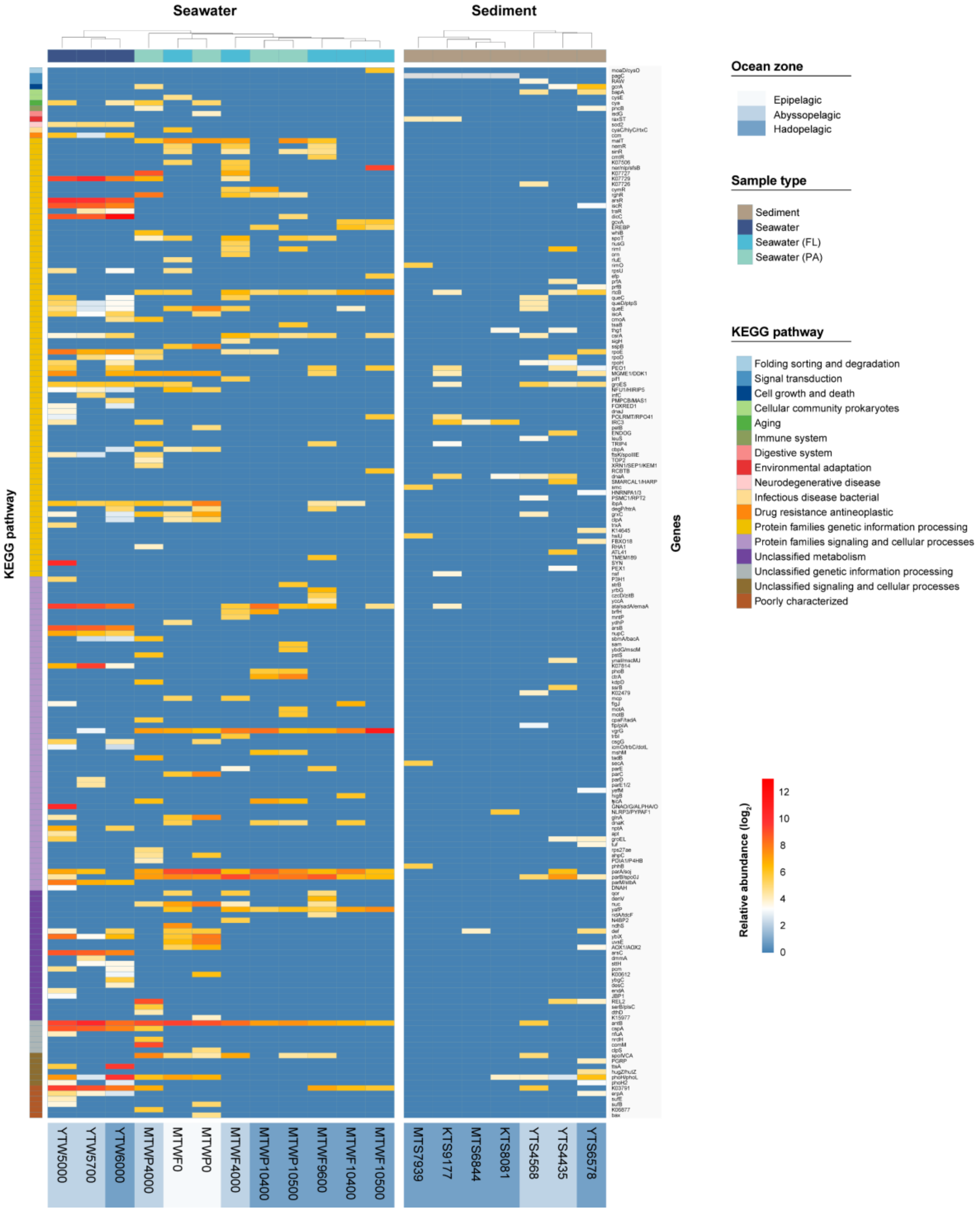
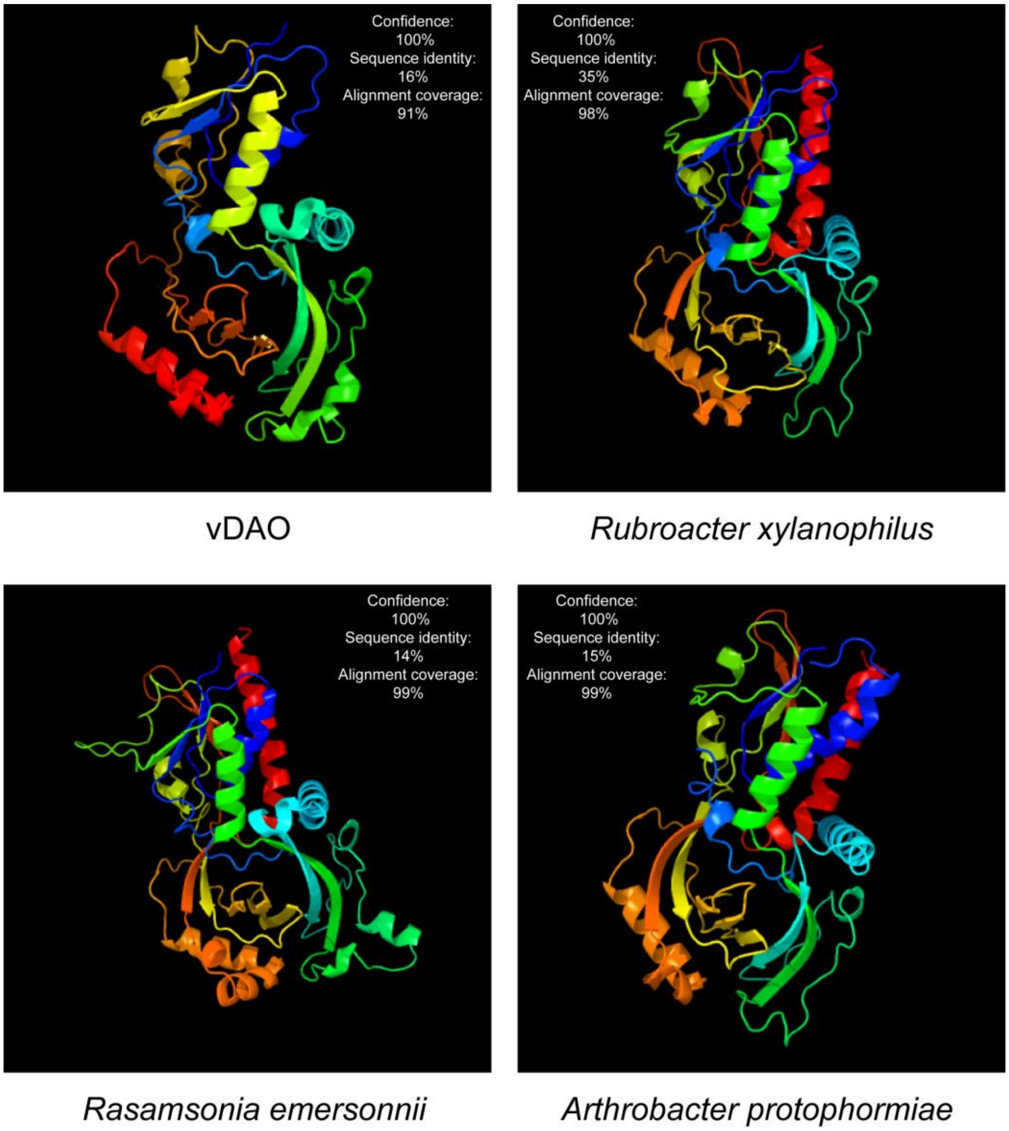


Fig. S19. Abundance patterns of the class II AMGs in the OTVGD by sample. The heatmap displays the relative abundance of each viral AMG (y-axis) in each sample (x-axis). The per-base per-contig coverage of mapped reads to vOTUs harbouring the corresponding AMGs is depicted as a heatmap of relative abundances on the $\log_2$ scale. The AMGs are clustered by sample and abundance (average linkage, Spearman rank correlation). The bars on the top and bottom of the figure indicate the type and ocean zone of each sample, respectively.



vDAO

Rubroacter xylanophilus

Rasamsonia emersonnii

Arthrobacter protophormiae

Fig. S20. Predicted 3D structures of vDAO and microbial DAOs with experimentally verified functions. The structures were modelled by PHYRE2 (www.sbg.bio.ic.ac.uk/phyre2/html/) with default parameters. All the structures have a confidence of 100%.

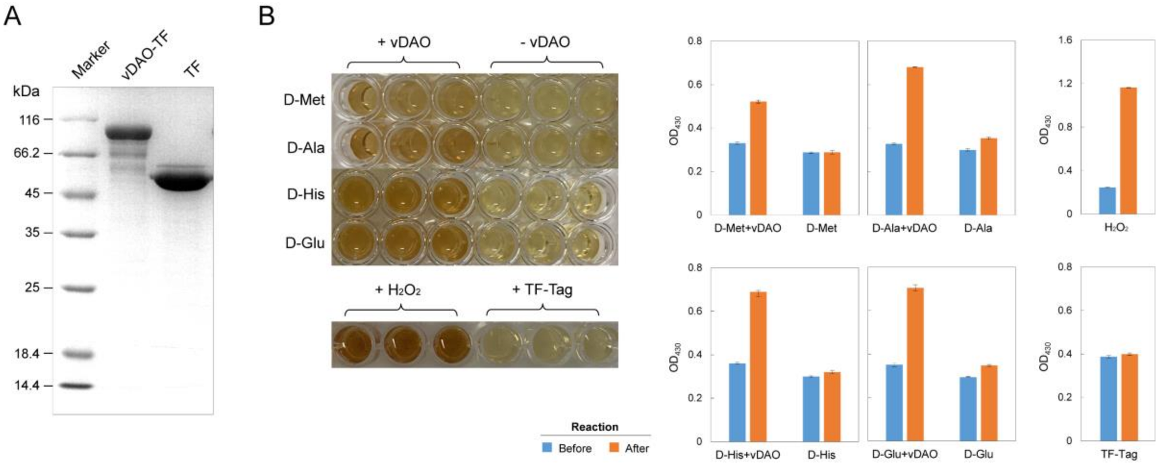


Fig. S21. vDAO expression, purification, and activity assay. (A) SDS-PAGE of purified vDAO with the TF-Tag protein. (B) D-amino acid oxidation assays. Enzyme activity was shown by the production of a brown product and an increase in the absorbance at 430 nm upon addition of vDAO. H_2O_2 and the TF-Tag protein were used as the positive and negative controls, respectively.

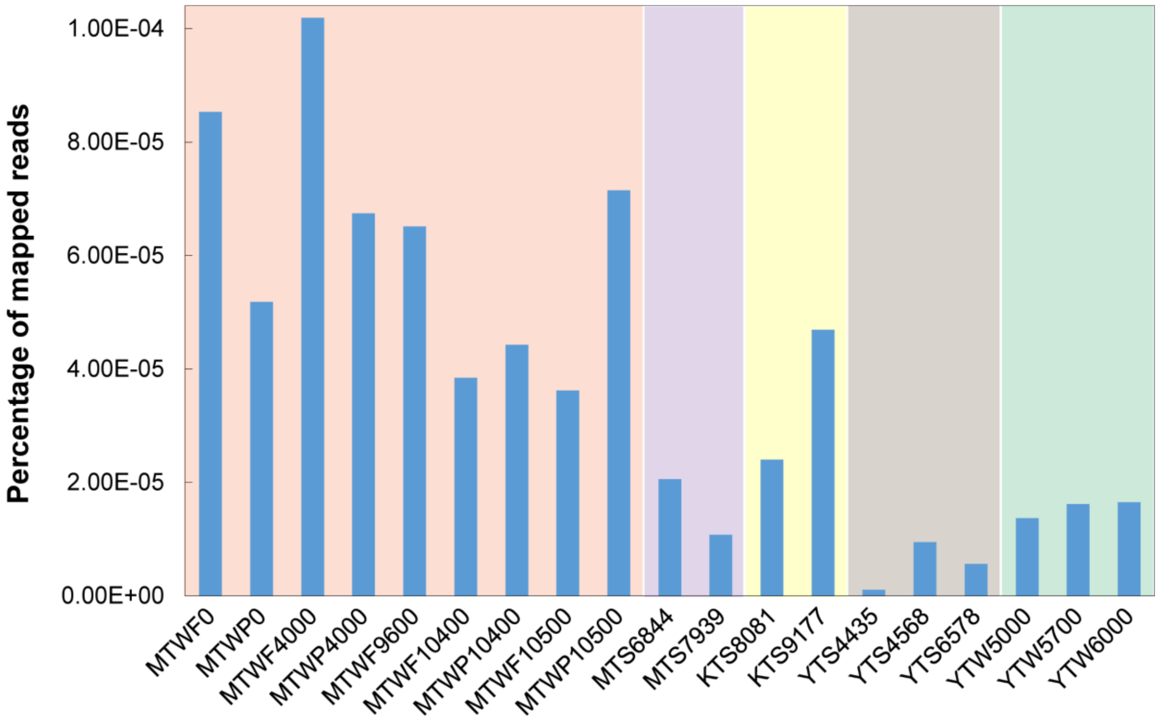


Fig. S22. Relative abundances of DAO-encoding genes in oceanic trenches. The abundances were calculated by the percentage of reads mapped to DAO-encoding genes among the total reads in each sample.

Supplementary Tables S1-S11

Table S1. List of metagenome datasets used in this study.

Table S2. List of hadal microbial genomes used in this study.

Table S3. Viral contig recovery by different assembly methods.

Table S4. Summary of manual curation for viral contigs.

Table S5. List and information of OTVGD vOTUs.

Table S6. Identification of NCLDVs in OTVGD by ViralRecall.

Table S7. Abundance and distribution of OTVGD vOTUs.

Table S8. List of microbial OTUs and metagenomic bins recovered in this study.

Table S9. List of host predictions for OTVGD vOTUs.

Table S10. Distribution and abundance of class I viral AMGs in the OTVGD.

Table S11. Distribution and abundance of class II viral AMGs in the OTVGD.

References

1. D. D. Corte *et al.*, Viral Communities in the Global Deep Ocean Conveyor Belt Assessed by Targeted Viromics. *Frontiers in Microbiology* **10** (2019).
2. B. L. Hurwitz, M. B. Sullivan, The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8**, e57355 (2013).
3. J. R. Brum *et al.*, Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
4. A. C. Gregory *et al.*, Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1-15 (2019).