Environmental viromes reveal global virosphere of deep-sea RNA viruses

Xiaobo Zhang (✉ zxb0812@zju.edu.cn)

Zhejiang University <https://orcid.org/0000-0003-4606-212X>

Biological Sciences - Article

Keywords:

Posted Date: February 10th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1294962/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **Title page**

2

3 **Environmental viromes reveal global virosphere of deep-sea RNA**
4 **viruses**

5

6 Xinyi Zhang¹, Haitao Wan¹, Min Jin², Liquan Huang¹, Xiaobo Zhang^{1*}

7 *1. College of Life Sciences, Laboratory for Marine Biology and Biotechnology of*
8 *Pilot National Laboratory for Marine Science and Technology (Qingdao) and*
9 *Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai),*
10 *Zhejiang University, Hangzhou 310058, People's Republic of China*

11 *2. Third Institute of Oceanography, Ministry of Natural Resources, Xiamen 361005,*
12 *People's Republic of China*

13

14

15 * Corresponding author: Prof. Xiaobo Zhang

16 Tel: 86-571-88981129

17 Fax: 86-571-88981151

18 Email: zxb0812@zju.edu.cn

19

20

21

22

Highlights

1. A global analysis of deep-sea RNA viruses revealed 85,059 viral operational taxonomic units (vOTUs), which is the largest number of RNA viruses known so far. Around 98.28% of the viruses were unclassified, indicating that the deep sea is a reservoir of novel RNA viruses.
2. The virome dataset of RNA viruses from 133 deep-sea sediments included 1,463 complete genomes, which considerably expands our understanding of the RNA viruses in deep-sea ecosystems.
3. The prokaryotic and eukaryotic RNA viruses respectively accounted for 7.09% and 65.81% of the classified deep-sea vOTUs.
4. The habitat had a greater impact on the deep-sea RNA viral communities compared to geographical location.

Abstract

On the earth, the most abundant and diverse life forms are viruses¹. Both DNA viruses and RNA viruses play important roles in marine ecosystems via regulating biogeochemical cycles². However, the virome of marine RNA viruses has not been explored so far. Here, we established the global virome dataset of deep-sea RNA viruses that were purified from 133 sediment samples collected from typical deep-sea ecosystems (hydrothermal vents, cold seeps, ocean basins and mid-ocean ridges) of the Pacific Ocean, Indian Ocean and Atlantic Ocean. A total of 85,059 viral operational taxonomic units (vOTUs) were identified, of which only 1.72% were hitherto known, indicating that the deep-sea sediment is a repository of novel RNA viruses. The vOTUs were classified into 20 viral families, including prokaryotic (7.09%) and eukaryotic (65.81%) RNA viruses, of which *Retroviridae* was the most abundant. Furthermore, circular genome analysis revealed 1,463 deep-sea RNA viruses with complete genomes, the majority of which were unclassified. The distribution and differentiation of deep-sea RNA viruses were mainly influenced by the habitat. Taken together, the deep sea is a reservoir of novel RNA viruses, and our study is the first to analyze the composition and diversity of RNA viruses in deep-sea ecosystems.

Main

Viruses are the most abundant and diverse life forms on the earth¹. The ocean, which covers 70% of the earth, is a habitat for many kinds of viruses³. Based on the analysis of a global ocean DNA virome dataset, a total of 195,728 marine viral populations were identified⁴. Another study showed that while most of marine viruses are novel, they are related to dominant, ecologically relevant microbial hosts⁵. Therefore these viruses likely influence biogeochemical cycles in marine ecosystems^{4,6} by infecting their hosts². Many virus-encoded genes are directly involved in sulfur and nitrogen cycling in the epipelagic zones of the oceans^{5,6}. In the deep-sea hydrothermal vents, the viruses compensate host metabolism by mediating branched metabolic pathways, which enable the hosts to survive in extreme environments⁷. Up to date, the marine DNA viruses have been well characterized. However, the information about marine RNA viruses, as well as deep-sea viruses, is limited.

Although most of the currently characterized marine viruses are double-stranded DNA (dsDNA) viruses⁸, RNA viruses may constitute as much as half of the marine viral community⁹. In fact, it is estimated that the marine RNA viruses may be as abundant or even exceed marine DNA viruses¹⁰⁻¹². Most marine RNA viruses are positive-sense, single-stranded RNA (+ssRNA) viruses, which are related to the current classification¹⁰. Very few double-stranded RNA (dsRNA) viruses are known and no RNA bacteriophages, negative-sense single-stranded RNA (−ssRNA) viruses or retroviruses have been identified in the ocean¹⁰. Almost all of the RNA viruses that

constitute the marine viroplankton can infect eukaryotic organisms, especially protists, such as diatoms, dinoflagellates, raphidophytes, prasinophytes and thraustochytrids¹⁰. Recently a novel RNA virus was isolated from deep-sea tubeworm *Osedax japonicus*, indicating that marine RNA viruses can infect animals apart from microorganisms¹³. Nevertheless, the deep-sea RNA viruses are largely unexplored.

The deep sea is one of the largest and the least explored ecosystems on the earth^{14,15}. Characterized by high pressure, low temperature, low nutrient levels, lack of sunlight and hypoxia, the deep sea is a challenge for inhabitants to survive^{16,17}. These extreme conditions put a selective pressure on the deep-sea organisms to adapt to the environment¹⁶. In this study, we analyzed the global virome of RNA viruses purified from 133 deep-sea sediment samples. The results showed that deep-sea RNA viruses are abundant and highly diverse, and most are hitherto unclassified. In addition, the distribution of these viruses is affected by the environmental factors.

Results

The global virome of deep-sea RNA viruses

A total of 133 deep-sea sediment samples were collected during the 26th, 30th, 34th, 39th, 40th and 45th cruises of the geomicrobiology cruise of China, which traveled more than 3,200,000 km from 2010 to 2018 (Fig 1A and Table S1). The samples were retrieved from four deep-sea habitats, including hydrothermal vents (71 samples), cold seeps (8 samples), seamounts (17 samples) and ocean basins (37 samples), from the Pacific Ocean, Atlantic Ocean and Indian Ocean (Fig 1A and Table S1). The depths of sampling stations ranged from 1,100 to 6,105 m, and the average

depth was 3,545.3 m (Table S1). The environmental types of sampling stations represented the typical environments of deep sea all over the world.

The viruses were purified from the sediments (Fig 1B), and RNAs extracted from the purified virions had no DNA contamination (Fig S1), which excluded the presence bacteria, archaea and eukaryotes. Furthermore, amplification of bacterial 16S rRNA fragments confirmed absence of bacteria in the extracted viral RNAs and cDNAs (Fig S2). These data indicated that the extracted viral genomic RNAs could be used for the sequencing analysis of the global virome of deep-sea RNA viruses.

To minimize sequencing errors, we established 400-bp fragment libraries of viral metagenomic DNAs obtained from the genomic RNAs. Each sample yielded more than 3.5 Gb clean data (Table S1), and the total clean data generated by all 133 samples was 673.721 Gb, which included 2,889,656,790 completely sequenced reads, representing the viral genomic sequences of the global virome of deep-sea RNA viruses.

Extreme diversity of deep-sea RNA viruses

To identify the community of global deep-sea RNA viruses, the sequenced reads of deep-sea RNA viruses were assembled, generating 261,857 contigs ($\geq 1.0\text{kb}$) (Fig 2A). Of these contigs, 86,517 contigs were assigned to viral contigs by VirSorter, VirFinder, VIBRANT and CAT analyses (Fig 2A and Table S2). Another 1,362 contigs included the signature genes of RNA viruses such as RNA-dependent RNA polymerase (RdRp) and reverse transcriptase (Fig 2A and Table S2), and therefore represented 1,362 RNA viruses. All 87,879 viral contigs have been uploaded into the

National Omics Data Encyclopedia database (accession number OEP002537). Based on nucmer analysis, 87,879 viral contigs were classified into 85,059 viral operational taxonomic units (vOTUs) (Fig 2A and Table S2).

Since most viral metagenomic sequences cannot be annotated and classified¹⁸, taxonomic classification of RNA viruses based solely on metagenomics is challenging. Therefore, we established a dataset of the genomic sequences of known RNA viruses from public databases, which included 1,963,257 sequences. Only 1.72% vOTUs matched the known sequences (Fig 2B and Table S3), indicating that the deep-sea sediment is a repository of novel RNA viruses.

The vOTUs were taxonomically assigned to 20 families of RNA viruses (Fig 2C and Table S3). *Retroviridae* was the most abundant family that accounted for 40.07% of all known vOTUs (Fig 2C). Depending on the hosts, 7.09% of the classified vOTUs were prokaryotic, 65.81% were eukaryotic, and 27.1% could not be assigned to a host (Fig 2D and Table S3). No archaeal RNA virus was detected, and the vertebrate-infecting viruses accounted for 42.12% of the known eukaryotic viruses (Fig 2D and Table S3).

Furthermore, 1,463 circular viral genomes (CVGs) were identified, representing 1,463 deep-sea RNA viruses with complete genomes (Fig 2E and Table S4). Only 0.75% of the CVGs matched the genomic sequences of known viruses, the remaining were indicative of novel viruses (Fig 2F and Table S4). The CVGs were classified into 6 families (Fig 2G and Table S4), of which *Metaviridae* was the most abundant viral family. In addition, 10 CVGs belonged to eukaryotic viruses and 1 CVG to

prokaryotic viruses (Fig 2H and Table S4).

Taken together, we established a dataset of 85,059 deep-sea RNA viruses, which is considerably larger than the viral RefSeq database of NCBI that includes only 2,094 RNA viruses (Fig 2I). Our findings indicate a highly diverse community of RNA viruses in deep-sea sediments, expanding our understanding of RNA viruses on the earth.

Global diversity signatures of deep-sea RNA viruses

To determine the diversity signatures of deep-sea RNA viruses on a global scale, we analyzed the distribution patterns of RNA viruses in deep-sea habitats spread across three oceans. Most RNA viruses were unique to each ocean except Atlantic Ocean, and 4,390 vOTUs were common to all three oceans and formed the core RNA virome (Fig 3A and Table S5). However, the abundance of the common RNA viruses was less than <1% (Fig 3B). In addition, we also detected vOTUs that were unique to each ecosystem, while 8,098 vOTUs were common to all habitats (Fig 3C and Table S6), representing the core RNA viruses in deep-sea ecosystems. The relative abundance of these core RNA viruses was also less than 1% (Fig 3D). The most abundant vOTUs in cold seep included vOTU-17680, vOTU-45543 and vOTU-9409, while vOTU-83477 and vOTU-41323 were abundant in hydrothermal vents (Fig 3D). In ocean basins, vOTU-82736, vOTU-81222 and vOTU-67630 were the most abundant, while 8 vOTUs were the most abundant in seamounts (Fig 3D). Thus, many deep-sea RNA viruses are habitat-specific, indicating that the characteristic features of these habitats influence the resident viral communities.

Fifteen of the 20 known families of deep-sea RNA viruses identified in this study formed the core viral families of all oceans and habitats (Fig 3E). Among the core viral families, *Totiviridae*, *Retroviridae*, *Cystoviridae* and *Leviviridae* were the dominant families globally. However, the relative abundances of these dominant families were significantly different between the three oceans as well as the four habitats (Fig 4F). These results were indicative of high intra-family diversity (vOTU level) and low inter-family diversity of global deep-sea RNA viruses, which suggested that the deep-sea RNA virus communities were influenced by environmental factors.

The viral diversity index was significantly different across the three oceans and four habitats ($p < 0.05$) (Fig 3G, 3H and Table S7). The Pacific Ocean had the most diverse RNA viruses, whereas the lowest diversity was observed in the Indian Ocean (Fig 3G). Among the different habitats, the cold seeps and hydrothermal vents had the highest and lowest diversity of RNA viruses, respectively (Fig 3H). Thus, the viral communities of deep-sea sediments are highly diverse and show spatial differences across the global scale.

Principal co-ordinates analysis (PCoA) was performed to further explore the global distribution and diversity of deep-sea RNA viruses, and the results showed that the vOTUs could be clustered into 3 groups (Fig 3I and 3J). All vOTUs of the Atlantic Ocean and most vOTUs of the Pacific Ocean and Indian Ocean were clustered into one group, while the remaining vOTUs of Pacific Ocean and Indian Ocean were aggregated into two groups (Fig 3I), suggesting that the deep-sea viral communities were similar in the three oceanic regions. Except for a few sediment samples collected

from cold seeps, ocean basins, seamounts and hydrothermal vents, the vOTUs of almost all samples from each habitat were clustered (Fig 3J), indicating that the unique features of these ecosystems rather than the geographical location influenced the deep-sea RNA virus communities.

Functional roles of virus-encoded genes in host metabolism

To determine the role of viral genes in host metabolism, the 85,059 vOTUs of deep-sea RNA viruses were further annotated using the Pfam, InterProScan, UniProt and Kyoto encyclopedia of genes and genomes (KEGG) databases. Based on the predicted open reading frames (ORFs), 167,959 putative RNA viral genes were identified, generating 60,216 annotated ORFs (Fig 4A). After clustering by CD-HIT, 33,021 protein clusters of deep-sea vOTUs were obtained (Fig 4A), which were searched against NCBI Viral RefSeq and IMG/VR v3 databases. Only 9.8% (n=3,245) protein clusters matched the proteins in NCBI Viral RefSeq database, while 30.5% (n=10,084) protein clusters were homologous to the known viral proteins in IMG/VR v3 database (Fig 4B). Thus, most proteins encoded by deep-sea RNA viruses were novel, further underscoring that the deep sea is a reservoir of novel RNA viruses. Furthermore, only 229 protein clusters were shared by the four deep-sea habitats, and 374 protein clusters were the core clusters for the three oceans (Fig 4C). Most protein clusters were unique to the local environment, indicating that the viral communities of the three oceanic regions or four habitats had low similarity at the protein level, which is consistent with the distribution pattern of vOTUs.

To further explore the role of deep-sea RNA viruses in host metabolism, we

identified the virus-encoded genes that potentially regulate metabolic pathways. Of the 17,140 viral genes annotated in the KEGG pathway database, 49.99% (n=8,568) were classified into the “Metabolism” category of KEGG database. These putative viral metabolic genes were further divided into 12 subgroups, of which “Carbohydrate metabolism” was most enriched (2037 metabolic genes), followed by “Amino acid metabolism” (1885), “Metabolism of cofactors and vitamins” (900) and “Energy metabolism” (804) (Fig 4D). Therefore, virus-encoded genes might affect the metabolic pathways in their hosts.

Given the crucial roles of sulfur metabolism in deep-sea ecosystems^{4,19}, we analyzed the type and abundance of virus-encoded genes that mediate sulfur metabolism in four deep-sea habitats. A total of 10 viral genes were identified that are involved in assimilatory sulfate reduction and thiosulfate oxidation, the most important processes in sulfur metabolism²⁰⁻²². The relative abundance of these 10 genes was significantly higher in hydrothermal vents compared to the other deep-sea habitats (Fig 4E), which is consistent with the high sulfur concentration in hydrothermal vent ecosystems. Taken together, the distribution of deep-sea RNA viruses is influenced by the ecological habitats.

Discussion

Marine viruses regulate biogeochemical cycles in the oceans by infecting hosts, and therefore have a significant impact on marine ecosystems². Although the majority of RNA viruses infect humans^{23,24}, little is known regarding the global diversity and distribution of marine RNA viruses. We identified 85,059 RNA viruses from global

deep-sea sediments, which is considerably higher than the 2,094 RNA viruses currently included in the viral RefSeq database of NCBI, and is indicative of the extremely diverse deep-sea viral communities. Furthermore, the viruses isolated from ocean sediments better represent the deep-sea communities compared to those isolated from ocean waters since the latter can be affected by currents²⁵. Only 1.72% of the deep-sea RNA viruses identified in our study were homologous to the known RNA viruses, and the majority of the genes encoded by the deep-sea RNA viruses are unknown, indicating that the deep sea is a reservoir of novel RNA viruses. Only a few RNA bacteriophages have been reported so far^{26,27}. We identified 104 RNA bacteriophages belonging to the *Cystoviridae*, *Leviviridae* and *Pseudoviridae* families, and no archaeal RNA viruses were detected. Therefore our findings provided novel knowledge for RNA viruses.

Studies show that the viral diversity and abundance differ significantly between marine ecosystems²⁸. The virions isolated from extreme thermal environments have few similar sequences compared to the published viruses, even though they are morphologically similar²⁹. In the Antarctic and bathypelagic regions, multi-zonal viral populations are predominant, whereas zone-specific regional viral populations dominate in temperate and tropical epipelagic regions and Arctic regions⁴. However, the relationship between the deep-sea habitats and viruses has not been explored so far. We analyzed the RNA viromes of typical global deep-sea habitats including hydrothermal vents, cold seeps, seamounts and ocean basins, and found that the deep-sea RNA virus communities are influenced by the environmental factors rather

than the geographical locations. Recent studies have shown that marine viruses encode metabolic genes that can modify host metabolism during virus infection^{5,7}. In this study, we identified 8,568 RNA virus-encoded genes involved in carbohydrate, amino acid and energy metabolism. The relative abundance of virus-encoded sulfur metabolism genes was highest in hydrothermal vent fields, which is consistent with the higher sulfur concentration and active sulfur-based chemosynthesis in these regions. Our findings provide new insights into deep-sea RNA virus communities and their roles in the local as well as global marine ecosystems.

References

1. Cobián Güemes, A. G. et al. Viruses as winners in the game of life. *Annu. Rev. Virol.* **3**, 197-214 (2016).
2. Chow, C. E. & Suttle, C. A. Biogeography of viruses in the sea. *Annu. Rev. Virol.* **2**, 41-66 (2015).
3. Zhang, X. *Virus Infection and Tumorigenesis-Hints From Marine Hosts' Stress Responses*. (Springer, 2019).
4. Gregory, A. C. et al. Marine DNA viral macro- and microdiversity from pole to pole. *Cell* **177**, 1109-1123 (2019).
5. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693 (2016).
6. York, A. Marine microbiology: algal virus boosts nitrogen uptake in the ocean. *Nat. Rev. Microbiol.* **15**, 573 (2017).
7. He, T., Li, H. & Zhang, X. Deep-sea hydrothermal vent viruses compensate for

microbial metabolism in virus-host interactions. *mBio* **8**, e00893-17 (2017).

8. Brum, J. R. et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).

9. Steward, G. F. et al. Are we missing half of the viruses in the ocean? *ISME J.* **7**, 672-679 (2013).

10. Culley, A. I. et al. The characterization of RNA viruses in tropical seawater using targeted PCR and metagenomics. *mBio* **5**, e01210-14 (2014).

11. Vlok, M., Lang, A. S. & Suttle, C. A. Marine RNA virus quasispecies are distributed throughout the oceans. *mSphere* **4**, e00157-19 (2019).

12. Urayama, S. I. et al. Unveiling the RNA virosphere associated with marine microorganisms. *Mol. Ecol. Resour.* **18**, 1444-1455 (2018).

13. Urayama, S. I., Takaki, Y., Nunoura, T. & Miyamoto, N. Complete genome sequence of a novel RNA virus identified from a deep-sea animal, *Osedax japonicus*. *Microbes Environ.* **33**, 446-449 (2018).

14. Sebastián, M. et al. Deep ocean prokaryotic communities are remarkably malleable when facing long-term starvation. *Environ. Microbiol.* **20**, 713-723 (2018).

15. Tortorella, E. et al. Antibiotics from deep-sea microorganisms: current discoveries and perspectives. *Mar. Drugs* **16**, 355 (2018).

16. Danovaro, R., Corinaldesi, C., Dell'Anno, A. & Snelgrove, P. V. R. The deep-sea under global change. *Curr. Biol.* **27**, 461-465 (2017).

17. Pilkington, L. I. A chemometric analysis of deep-sea natural products. *Molecules* **24**, 3942 (2019).

18. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833-844 (2017).
19. Anantharaman, K., Breier, J. A. & Dick, G. J. Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. *ISME J.* **10**, 225-239 (2016).
20. Cech, G. M., Szalewska-Pałasz, A., Potrykus, K. & Kloska, A. Virus-host interaction gets curiouser and curiouser. PART II: functional transcriptomics of the *E. coli* DksA-deficient cell upon phage P1_{vir} infection. *Int. J. Mol. Sci.* **22**, 6159 (2021).
21. Kieft, K. et al. Virus-associated organosulfur metabolism in human and environmental systems. *Cell Rep.* **36**, 109471 (2021).
22. McKay, L. J. et al. Sulfur cycling and host-virus interactions in Aquificales-dominated biofilms from Yellowstone's hottest ecosystems. *ISME J.* doi: 10.1038/s41396-021-01132-4 (2021).
23. Wolf, Y. I. et al. Origins and evolution of the global RNA virome. *mBio* **9**, e02329-18 (2018).
24. Luo, M., Terrell, J. R. & Mcmanus, S. A. Nucleocapsid structure of negative strand RNA virus. *Viruses* **12**, 835 (2020).
25. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356-361 (2005).
26. Krishnamurthy, S. R., Janowski, A. B., Zhao, G., Barouch, D. & Wang, D. Hyperexpansion of RNA bacteriophage diversity. *PLoS Biol.* **14**, e1002409 (2016).
27. Harb, L. et al. ssRNA phage penetration triggers detachment of the F-pilus. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25751–25758 (2020).

331 28. Suttle, C. A. Marine viruses-major players in the global ecosystem. *Nat. Rev.*
332 *Microbiol.* **5**, 801-812 (2007).

333 29. Snyder, J. C. & Young, M. J. Advances in understanding archaea-virus
334 interactions in controlled and natural environments. *Curr. Opin. Microbiol.* **14**,
335 497-503 (2011).

Fig 1

A

B

Fig 1. The global virome of deep-sea RNA viruses. (A) The distribution of global deep-sea sediment stations. The spots represent the locations of deep-sea stations. (B) The representative images of viruses purified from 133 deep-sea sediments. The viruses were observed by transmission electron microscopy. Scale bar, 100 nm or 200 nm.

Fig 2

A

B

D

C

E

F

G

H

I

Fig 2. Extreme diversity of deep-sea RNA viruses. (A) Identification of viral

operational taxonomic units (vOTUs). After assembly of total reads, 261,857 contigs

(≥ 1 kb) were obtained. The subsequent analyses by VirSorter, VirFinder, VIBRANT, CAT, RNA-dependent RNA polymerase (RdRp) and reverse transcriptase (or RNA-dependent DNA polymerase) analyses generated 87,879 viral contigs, the completeness of which was determined by CheckV. Based on the analysis using nucmer pipeline of MUMmer 4.0, the viral contigs were assigned into 85,059 vOTUs. (B) The relative proportion of unknown or known vOTU in the entire dataset. (C) The families of classified vOTUs and their abundance. (D) The proportion of hosts of the vOTUs matching the known RNA viruses. (E) The length and proportion of RNA viruses with circular viral genomes (CVGs). (F) The known and unknown RNA viruses with CVGs. (G) The number of the known RNA viruses with complete CVGs at the family level. (H) The hosts of RNA viruses with circular genomes at the host domain level. (I) The number of deep-sea RNA virus database in our study and viral RefSeq database in NCBI.

Fig 3

Fig 3. Global diversity signatures of deep-sea RNA viruses. (A) The number of vOTUs in three oceans. The core RNAs viruses existing in all oceans consisted of 4,390 vOTUs. (B) The relative abundance of vOTUs in three oceans. “Others” indicate the vOTUs with relative abundance less than 1%. (C) The number of vOTUs in four deep-sea ecosystems. (D) The relative abundance of vOTUs in hydrothermal vents, cold seeps, seamounts and ocean basins. “Others” represent the vOTUs with relative abundance less than 1%. (E) The distribution of the known viral families of RNA viruses in three oceans (Atlantic Ocean, Indian Ocean and Pacific Ocean) or four deep-sea habitats (hydrothermal vent, cold seep, seamount and ocean basin). Fifteen viral families were the core families of three oceans or four habitats. The number of the known viral families is indicated. (F) The relative abundance of the known viral families in Atlantic Ocean, Indian Ocean and Pacific Ocean or in hydrothermal vents, cold seeps, seamounts and ocean basins. “Others” show the viral families with relative abundance less than 1%. (G) Boxplots indicating the diversity index of communities of deep-sea RNA viruses in three oceans. The top, medium and bottom part of each box correspond to the highest, median and lowest diversity index of single viral community. The letter “n” represents the number of samples in each ocean. (H) The diversity index of RNA viruses of four deep-sea ecosystems. The letter “n” represented the number of samples in each ecological habitat. (I) Principal co-ordinates analysis (PCoA) of the vOTUs from Atlantic Ocean, Indian Ocean and Pacific Ocean. (J) PCoA of the vOTUS from hydrothermal vents, cold seeps, seamounts and ocean basins.

Fig 4

A

B

C

D

E

Fig 4. Functional roles of virus-encoded genes in host metabolism. (A) Diagram showing the identification of protein clusters from vOTUs of global deep-sea RNA viruses. vOTUs, viral operational taxonomic units; ORFs, open reading frames. (B) Proportion of protein clusters of deep-sea vOTUs homologous to the known viral proteins in NCBI Viral RefSeq and IMG/VR v3 databases. (C) The distribution of protein clusters of deep-sea RNA viruses in three oceans and four habitats. The numbers indicate the quantity of protein clusters of RNA viruses. (D) The metabolic pathways mediated by putative viral genes of deep-sea RNA viruses based on the

KEGG pathway database. (E) The relative abundance of sulfur metabolism genes encoded by deep-sea RNA viruses in different habitats. The genes encoded sulfate/thiosulfate transport system ATP-binding protein, adenylylsulfate kinase, sulfate adenylyltransferase subunit 2, phosphoadenosine phosphosulfate reductase, sulfite reductase (NADPH) hemoprotein beta-component, sulfate adenylyltransferase subunit 1, 5'-bisphosphate nucleotidase, sulfate/thiosulfate transport system permease protein, sulfate/thiosulfate transport system permease protein and sulfur-oxidizing protein.

Supplementary Information is available in the online version of the paper

Acknowledgements

This work was financially supported by China Ocean Mineral Resources R & D Association (DY135-B-04) and the National Key Research and Development Program of China (2018YFC0310703). The bioinformation analysis of this study was supported by Mingke Biotechnology Co., Ltd. (Hangzhou, China).

Author contributions

X. Y. Z. and X. B. Z. designed the study. M. J. contributed extensively sample collection. X. Y. Z. and H. W. performed the experiments. X. Y. Z., L. H. and X. B. Z. wrote and edited the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary information

Methods

Collection of deep-sea sediments

The deep-sea sediment samples, derived from oceanic vessel No.1 (Dayang No. 1) geomicrobiology cruises of China, were collected during the 26th, 30th, 34th, 39th, 40th and 45th cruises in the Pacific Ocean, the Atlantic Ocean and Indian Ocean from 2012 to 2018 (Table S1). The samples came from various deep-sea habitats including hydrothermal vents, cold seeps, seamounts, ocean basins and mid-ocean ridges with depth ranging from 1,154 to 6,105 m. The samples were stored at -80°C before experiments.

Purification of virions from deep-sea sediment samples

A deep-sea sediment (20 g), resuspended using 10 ml prefiltered (0.015-μm pore size) Milli-Q water and glass beads, was incubated with shaking for 30 min at 4°C. Subsequently the mixture was centrifuged at 5,000×g for 20 min (4°C) to collect the supernatant. After repeating these steps for six times, all the supernatants were collected and centrifuged at 5,000×g for 10 min at 4°C. The supernatant was filtered through a 0.22-μm filter. The filtrate was added with PEG6000 at a final concentration of 10% and then incubated at 4°C overnight. After ultracentrifugation at 200,000×g for 2 h, the pellet was collected. The purified virions were dissolved in SM buffer (400mmol/L NaCl, 20mmol/L MgSO₄ · 7H₂O, 50mmol/L Tris-HCl, pH7.5) and observed under a transmission electron microscope.

RNA extraction, reverse transcription and amplification of deep-sea RNA viruses

To exclude the contamination of exogenous DNA and RNA, the purified deep-sea virions were treated with DNase and RNase at 37°C for 1h. Subsequently the mixture was incubated in 20μL 1M ethylene diamine tetraacetic acid (EDTA) (pH8.0) at 65°C for 10 min, followed by the extraction of viral genomic RNAs with a RNA purification kit (NorgenBiotek Corp, Thorold, Canada). The extracted RNAs were treated with RNase-free DNase at 37°C for 30 min. Then the reverse transcription was performed using HiScript II 1st strand cDNA synthesis kit (+gDNA wiper) (Vazyme, Nanjing, China) according to the manufactory's protocol. The synthesis of cDNA was conducted with second strand cDNA synthesis kit (Beyotime Biotechnology, Shanghai, China). Subsequently the cDNA was subjected to isothermal amplification using GenomiPhi™ V2 DNA amplification kit (GE Healthcare Life Sciences, Buckinghamshire, UK). The amplification condition was 95°C for 3 min, 30°C for 3h and 65°C for 10 min.

Detection of bacterial 16S rRNA gene for the prepared samples

To exclude the bacterial contamination, the viral genomic RNAs extracted from the purified deep-sea virions of each of 133 deep-sea sediments, the first-strand cDNAs, the second-strand cDNAs and the products of isothermal amplification were subjected to PCR using bacterial 16S rRNA gene-specific primers (515F, 5'-GTGCCAGCMGCCGCGG-3'; 907R, 5'-CCGTCAA TTCMTTTRAGTTT-3') (M=A/C; R=A/G). The PCR was conducted at 95°C for 5min, followed by 30 cycles of 95°C for 30s, 55°C for 30s, 72°C for 30s, and 72°C for 10min.

Sequencing of RNA viral metagenome

The amplified product of RNA virus with isothermal amplification was checked using a 1% agarose gel, purified and then quantified with the QuantiFluor-ST fluorescence quantitative system (Promega, CA, USA). After the treatment with an M220 focused ultrasonicator (Covaris Inc., Woburn, MA, USA), DNA was sheared and 400-bp fragments were excised and extracted. The paired-end library was prepared with the TruSeq DNA sample prep kit (Illumina Inc., San Diego, CA, USA). Subsequently the paired-end sequencing (2×250 bp) was conducted on an IlluminaHiSeq 2500 system (Illumina Inc., San Diego, CA, USA). The sequencing was cooperated with Mingke Biotechnology Co., Ltd. (Hangzhou, China). After removing the reads of adapters and duplicate reads, the raw sequences were trimmed to get the clean data.

Assembly of contigs and identification of viral contigs

To assemble the contigs, the reads were trimmed with Trimmomatic (version:0.33, default parameters)³⁰. Subsequently the reads of each sample were assembled to contigs using metaSPAdes 3.12.0³¹.

To obtain the viral contigs, three kinds of softwares, including VirSorter³², VirFinder³³ and VIBRANT³⁴, were used. The contigs whose length $\geq 1.0\text{kb}$ were selected and then were screened using the following algorithms: (1) VirSorter categories 1 and 2, (2) VirFinder score ≥ 0.9 and $p < 0.05$, (3) both VirSorter (categories 1-6) and VirFinder (score ≥ 0.7 and $p < 0.05$) or (4) all the contigs screened by VIBRANT. The contigs obtained by the VirSorter, VirFinder and VIBRANT algorithms were identified as viral contigs. To obtain more viral contigs, the

remaining contigs were further screened using the CAT algorithm³⁵. In this algorithm, the contigs with <40% of the genomes classified as bacteria, archaea or eukaryotes were considered as viral contigs⁴. All the viral contigs identified by Virsorter, VirFinder, VIBRANT and CAT were pooled together and the duplications were removed. To get more viral contigs, the contigs ($\geq 1.0\text{kb}$) that were not identified by VirSorter, VirFinder, VIBRANT and CAT were searched for RNA-dependent RNA polymerase (RdRp) and reverse transcriptase (or RNA-dependent DNA polymerase) based on viral RefSeq database (release 201)³⁶, NCBI non-redundant protein sequence database and IMG/VR database. The contigs contained at least one of RdRp and reverse transcriptase (or RNA-dependent DNA polymerase) were considered to be viral contigs. The viral contigs represented the viral genomes.

To evaluate the completeness of viral genomes, the sequences of viral contigs were subjected to CheckV v7.0 analysis³⁷.

Identification of viral operational taxonomic units

Nucmer pipeline of MUMmer 4.0³⁸ was used to classify the viral contigs into viral operational taxonomic units (vOTUs). The viral contigs were classified into vOTUs if the coverage between viral contigs was larger than 80% of the shortest contig, which shared $\geq 95\%$ mummer-based average nucleotide identity⁴.

Establishment of the known RNA virus genomic sequences database

All of the genomic sequences of the known RNA viruses were collected from the public databases to establish the known RNA virus genomic sequences database. The public databases included NCBI (National Center for Biotechnology Information)

viral RefSeq database (<https://www.ncbi.nlm.nih.gov/refseq/>), NCBI non-redundant protein sequence database (<https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/>), GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) and IMG/VR database (<https://img.jgi.doe.gov/cgi-bin/vr/main.cgi>). The established known RNA virus genomic sequences database was used to perform taxonomy annotation of vOTUs.

Identification of viral taxonomy

To identify viral taxonomy, Prodigal was utilized to predict open reading frames (ORFs) of each vOTU³⁹. Based on protein sequences of the predicted ORFs, viral taxonomy of vOTUs were identified using blastp (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/magicblast/LATEST>). The vOTUs were aligned with our self-built database (the known RNA virus genomic sequences database). In the meanwhile, the RNA-dependent RNA polymerase (RdRp) and reverse transcriptase of vOTUs were also aligned with our self-built database. The vOTUs would be classified into a known virus if >50% ORFs were aligned to that kind of virus and the blastp bit score was $\geq 50^4$.

Analysis of RNA viruses with circular genomes

To obtain the RNA viruses with circular genomes, the sequences of vOTUs of deep-sea RNA viruses were loaded into the Cenote-Taker software (<https://github.com/mtisza1/Cenote-Taker>)⁴⁰. Based on overlapping ends, the viruses with circular genomes were identified.

Analysis of the relative abundance of vOTUs

To reveal the relative abundance of vOTUs, the vOTUs of every sample were

mapped to the reads using the software bowtie2 v2.4.4⁴¹. After removal of the reads with low-quality mapping by BamM v1.7.3 (<https://github.com/Ecogenomics/BamM>) with parameters --percentage_id 0.95 --percentage_aln 0.75, the reads per kilobase per million mapped reads (RPKM) values were obtained based on the analysis using CoverM v0.3.1 (<https://github.com/wwood/CoverM>) (--percentage_id 0.95 --percentage_aln 0.75 -rpkm)⁴².

Diversity analysis of deep-sea RNA viruses

To characterize the diversity of vOTUs across three oceans (Pacific Ocean, Atlantic Ocean and Indian Ocean) or four ecological habitats (hydrothermal vent, cold seep, seamount and ocean basin), α -diversity indices (Shannon, Simpson, Chao and Ace) were calculated using vegan in R⁴³. Boxplots were drawn by GraphPad Prism 8.0 (<https://www.graphpad.com/>) to show the changes of four indices in different oceans or ecological habitats. At the same time, the β diversity of deep-sea RNA viruses was characterized using principal coordinate analysis (PCoA) with vegan package in R⁴³.

Principal co-ordinate analysis (PCoA) of deep-sea RNA viruses

PCoA was carried out to reveal the diversity of deep-sea RNA viruses using vegan package in R⁴³. Bray-Curtis dissimilarity matrices were generated from both the subsampled and total reads of global deep-sea RNAs viruses by vegdist (method=bray) after a cube root transformation by function nthroot (n=3)⁴.

Identification and comparison of the proteins encoded by vOTUs of deep-sea RNA viruses

To identify the proteins encoded by deep-sea RNA viruses, the ORFs predicted

from deep-sea vOTUs in this study were annotated by (1) blast hit analysis against the Kyoto encyclopedia of genes and genomes (KEGG) database⁴⁴, (2) blast against the UniProt Reference Clusters database⁴⁵, (3) searching for matches against the InterPro protein signature database using InterProScan⁴⁶ and (4) HMM searches against Pfams⁴⁷. After removal of duplications of the annotated viral proteins using 4 algorithms, the unique viral proteins were clustered by CD-HIT at 60% identity, 80% coverage and “-g 1 -n 4 -d 0”⁴⁸. The resulting protein clusters were compared with the Refseq Virus database and IMG/VR v3 by blastp in DIAMOND with an *e*-value threshold of $1e^{-5}$, identity of 30% and coverage of 50%³⁶.

Calculation of the relative abundance of viral open reading frames (ORFs)

To evaluate the relative abundance of viral ORFs of each sample, bowtie2 v2.4.4 was used to map the reads of each sample to viral ORFs⁴¹. After removal of the reads with low-quality mappings by BamM v1.7.3 (<https://github.com/Ecogenomics/BamM>) (--percentage_id 0.95 --percentage_aln 0.75), the reads per kilobase per million mapped reads (RPKM) values were generated using CoverM v0.3.1 (<https://github.com/wwood/CoverM>) (--percentage_id 0.95 --percentage_aln 0.75 -rpkm)⁴².

Phylogenetic analysis of deep-sea RNA viruses with circular viral genomes

To reveal the relationship between the classified deep-sea RNA viruses with circular viral genomes and the known RNA viruses, the phylogenetic analysis was carried out using MEGA (version 7.0.26)⁴⁹. The amino acid sequences of gag protein, protease and reverse transcriptase of deep-sea RNA viruses with circular viral genomes were aligned with the homologous sequences in NCBI nr database

(<https://www.ncbi.nlm.nih.gov/protein/?term=>) through the MUSCLE algorithm (v3.8)⁴⁹. The phylogenetic trees were produced by neighbor-joining method⁵⁰. The confidence coefficient was tested by bootstrap test (500 replicates).

Data availability All viral contigs are available in the National Omics Data Encyclopedia database (accession number OEP002537).

References

30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
31. Nurk, S. D., Meleshko, A. K. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824-834 (2017).
32. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
33. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).
34. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
35. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).

646 36. Buchfink, B., Xie, C. & Huson, D. Fast and sensitive protein alignment using
647 DIAMOND. *Nat. Methods* **12**, 59-60 (2015).

648 37. Nayfach, S. et al. CheckV assesses the quality and completeness of
649 metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578-585 (2021).

650 38. Marcais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS*
651 *Comput. Biol.* **14**, e1005944 (2018).

652 39. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation
653 site identification. *BMC Bioinformatics* **11**, 119 (2010).

654 40. Tisza, M. J. et al. Discovery of several thousand highly diverse circular DNA
655 viruses. *eLife* **9**, e51971 (2020).

656 41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat.*
657 *Methods* **9**, 357-359 (2012).

658 42. Li, Z. et al. Deep sea sediments associated with cold seeps are a subsurface
659 reservoir of viral diversity. *ISME J.* **15**, 2366-2378 (2021).

660 43. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.*
661 **14**, 927-930 (2003).

662 44. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes.
663 *Nucleic Acids Res.* **28**, 27-30 (2000).

664 45. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters:
665 a comprehensive and scalable alternative for improving sequence similarity searches.
666 *Bioinformatics* **31**, 926-932 (2015).

667 46. Zdobnov, E. M. & Apweiler, R. InterProScan--an integration platform for the

signature-recognition methods in InterPro. *Bioinformatics* **17**, 847-848 (2001).

47. Bateman, A. et al. The Pfam protein families database. *Nucleic Acids Res.* **32**, D138-141 (2004).

48. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

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

50. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641-1650 (2009).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS2.xlsx](#)
- [TableS7.xlsx](#)
- [TableS5.xlsx](#)
- [TableS4.xlsx](#)
- [TableS3.xlsx](#)
- [TableS1.xlsx](#)
- [deepseaRNAvirusExtendeddatafigure.pdf](#)
- [TableS6.xlsx](#)