ARTICLE OPEN

Viruses in deep-sea cold seep sediments harbor diverse survival mechanisms and remain genetically conserved within species

Yongyi Peng[1,2], Zijian Lu[3], Donald Pan[4], Ling-Dong Shi[5], Zhao Zhao[2,6], Qing Liu[2], Chuwen Zhang[1], Kuntong Jia[2], Jiwei Li[7], Casey R. J. Hubert[8] and Xiyang Dong[1,6]✉

Deep sea cold seep sediments have been discovered to harbor novel, abundant, and diverse bacterial and archaeal viruses. However, little is known about viral genetic features and evolutionary patterns in these environments. Here, we examined the evolutionary ecology of viruses across active and extinct seep stages in the area of Haima cold seeps in the South China Sea. A total of 338 viral operational taxonomic units are identified and linked to 36 bacterial and archaeal phyla. The dynamics of host-virus interactions are informed by diverse antiviral defense systems across 43 families found in 487 microbial genomes. Cold seep viruses are predicted to harbor diverse adaptive strategies to persist in this environment, including counter-defense systems, auxiliary metabolic genes, reverse transcriptases, and alternative genetic code assignments. Extremely low nucleotide diversity is observed in cold seep viral populations, being influenced by factors including microbial host, sediment depth, and cold seep stage. Most cold seep viral genes are under strong purifying selection with trajectories that differ depending on whether cold seeps are active or extinct. This work sheds light on the understanding of environmental adaptation mechanisms and evolutionary patterns of viruses in the sub-seafloor biosphere.

The ISME Journal (2023) 17:1774–1784; https://doi.org/10.1038/s41396-023-01491-0

INTRODUCTION

Cold seeps are deep-sea environments where hydrocarbon fluids and gas seepage occur at the continental margins worldwide. The continuous seepage of gaseous and liquid hydrocarbons boosts local biodiversity and microbial activity, featuring prevalent archaeal anaerobic methanotrophs (ANME) and sulfate-reducing bacteria (SRB) [1, 2]. Compared to the rich knowledge of cold seep bacterial and archaeal communities, viruses remain largely underexplored in spite of their significant roles in impacting microbes and corresponding biogeochemical cycles [3, 4]. Virus studies using enumeration or cultivation have shown that cold seep sediments are hotspots of viral production with high virus-prokaryote ratios [5, 6]. A recent survey of metagenomes from seven cold seeps demonstrates that these sediments harbor diverse and novel viruses, hinting at their potential impact on hydrocarbon biodegradation and other local metabolisms catalyzed by cold seep microbiomes [7]. However, cold seep viral diversity and distribution patterns, virus-microbe interactions, adaptive mechanisms to environmental factors, and viral genetic diversity are still relatively unexplored.

Viruses have a genetic toolbox of diverse mechanisms to adapt to the environment and co-evolve with hosts. As foreign mobile genetic elements, viruses face a wide repertoire of antiviral defense systems, including restriction-modification (RM) and CRISPR-Cas [8]. In line with antagonistic co-evolution of viruses and their hosts [9, 10], viruses have developed efficient and robust counter-defense systems, such as anti-restriction, anti-CRISPR, and other counter-defense proteins [11, 12]. Diversity-generating retroelements (DGRs) containing reverse transcriptase (RT) are another important diversification mechanism for driving sustained amino acid-level diversification of their target domains [13, 14]. Viruses also encode DGRs to produce many mutations in specific regions of host target genes through error-prone reverse transcription [15–17]. To replicate more efficiently, viruses can alter their hosts' metabolic potential through the expression of auxiliary metabolic genes (AMGs) to modulate host cell metabolism during infection [18]. In addition to these gene inventories, viruses can use alternative genetic codes different from those of their host, potentially increasing viral adaptability (e.g., in regulating translation of lytic genes) [19, 20]. Whether or not cold seep viruses incorporate these strategies into their repertoire of mechanisms for mediating host-virus interactions and environmental adaptation in these harsh deep-sea subseafloor environments requires further investigation.

Intra-population genetic variations (microdiversity) can also improve virus adaptation to their environment by driving phenotypic variation [21, 22]. For example, depth-dependent evolutionary strategies of viruses were observed in the Mediterranean Sea [9] and grassland soil in northern California [10]. Large viral microdiversity was observed for perhaps the most abundant

[1]Key Laboratory of Marine Genetic Resources, Third Institute of Oceanography, Ministry of Natural Resources, Xiamen 361005, China. [2]School of Marine Sciences, Sun Yat-Sen University, Zhuhai 519082, China. [3]South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China. [4]School of Oceanography, Shanghai Jiao Tong University, Shanghai 200030, China. [5]College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China. [6]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519000, China. [7]Institute of Deep-Sea Science and Engineering, Chinese Academy of Sciences, Sanya 572000, China. [8]Department of Biological Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada. ✉email: dongxiyang@tio.org.cn

Received: 22 March 2023 Revised: 1 August 2023 Accepted: 1 August 2023
Published online: 12 August 2023

ocean virus in temperate and tropical waters infecting *Pelagibacter* [23], whereas viruses were under significantly low evolutionary pressures in stable subzero Arctic brines [24]. The principles governing the viral evolution likely differ depending on environmental conditions, such as host dynamics, physicochemical properties, and population sizes [25–27]. Examining 39 abundant microbial species identified in sediment layers below the sea floor and across six cold seep sites, we previously reported that their evolutionary trajectories were depth-dependent and differed across phylogenetic clades [1]. However, it remains to be answered if cold-seep viruses are undergoing similar evolutionary patterns and selection pressures.

To understand adaptive survival mechanisms and genetic microdiversity of cold seep viruses, we extracted viral genomes from 16 sediment core samples in the area of Haima cold seeps in the South China Sea (Supplementary Figure 1 and Supplementary Table 1). Cores were collected from two active seeps with dense and living bivalves, as well as from one extinct seep covered with many dead clams [28]. We explored viral diversity patterns at both the community-level (macrodiversity) and population-level (microdiversity), and the viral functional gene repertoire related to arms race between viruses and their prokaryotic hosts. This study expands the knowledge of ecological and evolutionary patterns of viruses inhabiting cold seep subsurface ecosystems.

RESULTS AND DISCUSSION

Diverse antiviral strategies in cold seep microbial genomes

In total, 16 metagenomic data sets were derived from depth-discrete sediment core samples obtained from two active ($n = 5$ for Active1; $n = 6$ for Active2) and one extinct ($n = 5$) cold seeps (Supplementary Figure 1 and Supplementary Table 1), at depths ranging from 0 to 20 cm below the sea floor (cmbsf) [28]. Bacterial and archaeal community structures varied between different depth layers at the three sites (Supplementary Fig. 2 and Supplementary Table 2). Active seep sediments were dominated by taxa affiliated with *Halobacteriota* and *Desulfobacterota*, whereas the members of *Desulfobacterota* and *Chloroflexota* were the major microbial lineages in extinct seep sediments. After assembly, 487 species-level metagenome-assembled genomes (MAGs) were reconstructed at an average nucleotide identity (ANI) threshold of 95% (Supplementary Figure 3 and Supplementary Table 3), spanning 53 bacterial and 10 archaeal phyla, with the majority affiliated with *Proteobacteria* ($n = 59$), *Desulfobacterota* ($n = 56$), *Chloroflexota* ($n = 49$), *Bacteroidota* ($n = 38$), and *Thermoplasmatota* ($n = 24$).

Bacteria and archaea possess diverse antiviral strategies to defend against infection by their viruses [29–31]. A total of 2,145 antiviral genes were detected in 63% of cold seep microbial genomes, and could be assigned to 43 families of antiviral systems [8, 32]; these include restriction-modification (RM) systems that target specific sequences on the invading DNA elements, and CRISPR-Cas systems that use RNA-guided nucleases to cleave foreign sequences [33] (Fig. 1a and Supplementary Table 4). On average, the cold seep microbial genomes encode two antiviral systems per genome and the number of antiviral systems is positively correlated with the genome size for each MAG (linear regression; $R^2 = 0.27$, $p = 4.73 \times 10^{-5}$; Fig. 1b), similar to previous observations on the importance of genome size for encoding accessory systems in prokaryotes or ocean microbiomes [8, 34]. The number of antiviral systems per genome varies from zero (179 genomes) to 32 in a genome belonging to the phylum *Fermentibacterota* (classified as JAFGKV01 at the family-level; Supplementary Table 4), followed by 30 in a *Gammaproteobacteria* genome and 27 in a *Bacteroidia* genome. On average, the bacterial genomes encode more antiviral systems per genome than those in archaeal genomes (3.9 vs 2.4). The most abundant species in the metagenomic dataset (18% of the microbial community) is the putative anaerobic methanotroph ANME-1 SY_S15_40 that encodes two RM type II and one RM Type IIG systems (Supplementary Tables 3 and 4). Based on surveying large datasets of sequenced genomes, RM and CRISPR-Cas systems were reported to be present in ~75% and ~40% of microbial genomes, respectively [29, 35]. Relatively fewer cold seep microbial genomes appear to encode RM (50.8%) and CRISPR-Cas systems (22.7%), yet feature higher frequencies of AbiEii (44%; one antiviral system of Abortive infection [36]) and SoFIC (38%) that can modulate various target protein activities [32] (Fig. 1c and Supplementary Table 4). Diverse antiviral systems were also found in microbial communities from Mediterranean sponge species [37], epipelagic and mesopelagic layers in the Pacific Ocean [38], a deep-sea hydrothermal microbial mat in the Guaymas Basin [39]. In general, they have different distribution patterns of antiviral systems from cold seep sediments. Overall, these data reveal diverse antiviral strategies throughout the Haima cold seep microbiome with specific enrichment in some antiviral systems that govern the dynamics of host-virus interactions.

Novel viral genomes linked to 36 microbial phyla

Cold seep samples contained highly abundant viruses with densities up to 7.6×10^{11} per gram sediments, with viral abundances being associated with sediment depth (Supplementary Table 5). From the 16 metagenomic data set, 488 single-contig viral genomes with ≥50% estimated completeness (based on CheckV [40]) were recovered using multiple virus identification tools (Fig. 2a and Supplementary Figure 4). Viral genomes were clustered into 338 species-level viral operational taxonomic units (vOTUs) [41], belonging to 83 viral clusters (VCs; roughly equivalent to an ICTV genus) utilizing whole genome gene-sharing profiles [42] (Supplementary Fig. 5 and Supplementary Table 6). Similar to observations in prokaryotic communities [1, 2, 43], alpha and beta diversity analyses of 338 vOTUs suggest that sampling site, sediment depth in relation to redox conditions [28], and the geological state of cold seeps (active or extinct) shape the structure of viral communities (Supplementary Fig. 6 and Supplementary Table 5).

Among the 338 vOTUs, 291 could be taxonomically assigned revealing that 288 are affiliated with the class *Caudoviricetes* (Fig. 2a and Supplementary Table 6), which encompasses tailed phages that are the most prevalent viral taxon across ecosystems [44]. Only ten vOTUs could be annotated at the order level, confirming a large knowledge gap in the taxonomy of deep-sea cold seep viruses [7]. With respect to viral lifestyles, 48 and 22 vOTUs were predicted to be lytic and lysogenic, respectively, with others being unclassified (Fig. 2a). Host predictions of these vOTUs revealed that virus-infected hosts were detected in 36 bacterial and archaeal phyla (Fig. 2b, c and Supplementary Table 7). From the 475 host-virus linkages, the most common phylum among predicted hosts was *Chloroflexota* ($n = 80$), followed by *Halobacteriota* ($n = 31$), *Asgardarchaeota* ($n = 30$), and *Desulfobacterota* ($n = 29$). This is consistent with our previous observation that a significant portion of viruses targeted archaea in cold seep sediments, and such a host-virus pattern has not been reported in other deep-sea ecosystems [7, 45, 46]. Ten viruses were predicted to infect ANME-1 and ANME-2 groups that perform anaerobic methane oxidation. Viruses infecting *Methanosarcinales* and *Gammaproteobacteria* were highly abundant in the extinct and active cold seep samples, respectively.

Cold seep viruses harbor diverse strategies for environmental adaptation

To protect against antiviral systems of their microbial hosts, cold seep viruses encode an extensive repertoire of counter-defense systems, including anti-CRISPR (Acr) proteins, methyltransferases, and antitoxins (Fig. 3a–c and Supplementary Table 8). A total of

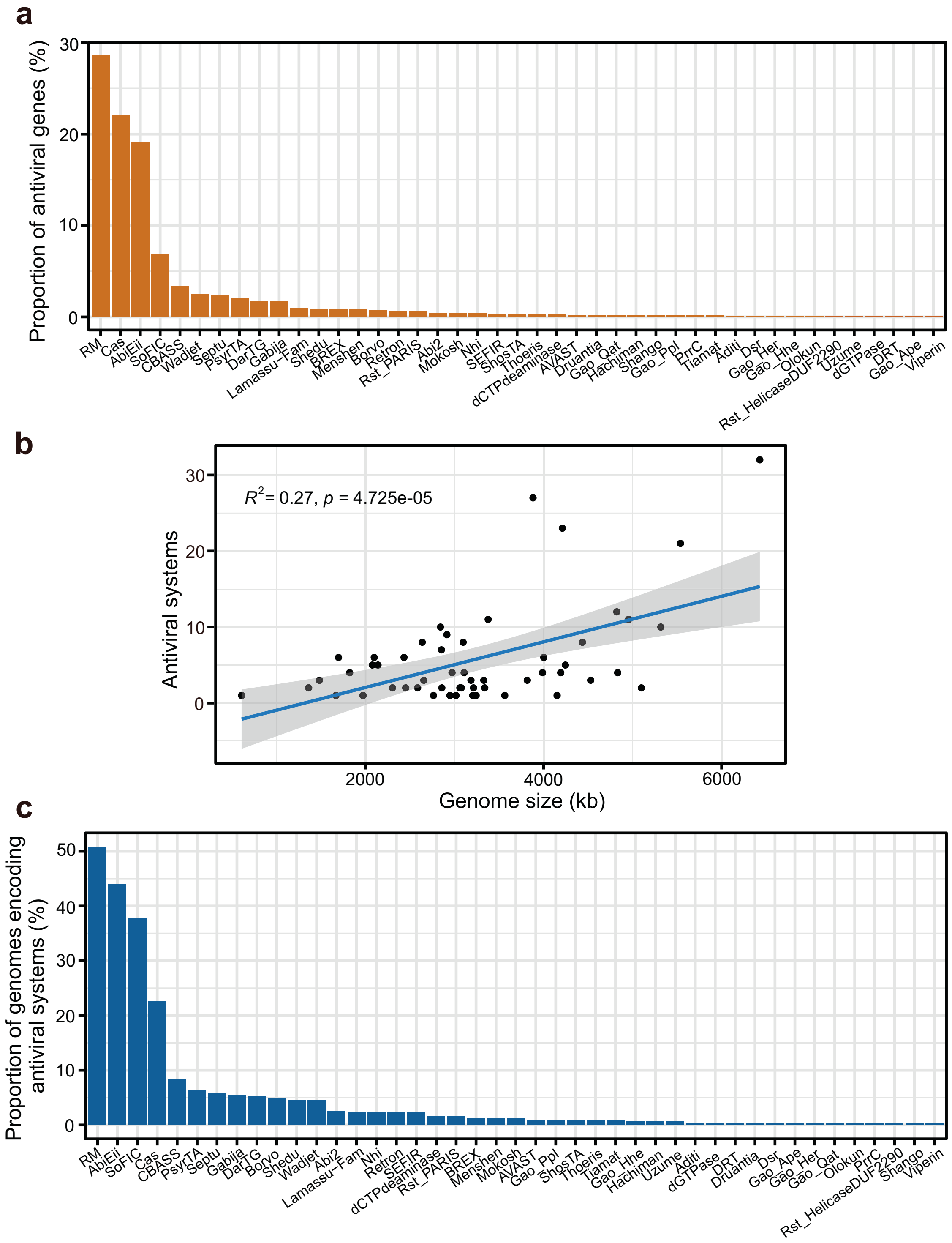

Fig. 1 Diversity of antiviral systems found in cold seep bacterial and archaeal genomes. **a** Proportion of antiviral genes from each type of antiviral systems in all the identified antiviral genes. **b** Relationship between antiviral system numbers per prokaryotic genome and their genome sizes. The correlation analysis was conducted with the completeness-filtered dataset (>90% genome completeness) to reduce the potential bias caused by the genome incompleteness. **c** Frequency of antiviral systems detected in microbial genomes. Detailed statistics for antiviral systems of microbial genomes are provided in Supplementary Table 4.

75 type II DNA methyltransferases without counterpart restriction enzymes were detected in 55 viral genomes (16% of all viruses), encoding diverse DNA modification enzymes (e.g., adenine- and cytosine-specific methyltransferases, and adenine methylase) [34]. The *acr-aca* operon (anti-CRISPR gene *acr* and *acr*-associated gene *aca*) [47] was identified in ten viral genomes (3%), which may inhibit the CRISPR-Cas immunity of the host to allow viruses to propagate [48]. Accordingly, one *Poribacteria* genome SY_Active_Co137 infected by a virus with the *acr-aca* operon has nine *cas* genes (Supplementary Tables 4 and 8). Interference modules of the antitoxin genes (e.g., *vapBC*, *relBE*, *hicBA*) were found in 63 viruses (19%) and belonged to the type II Toxin-antitoxin (TA)

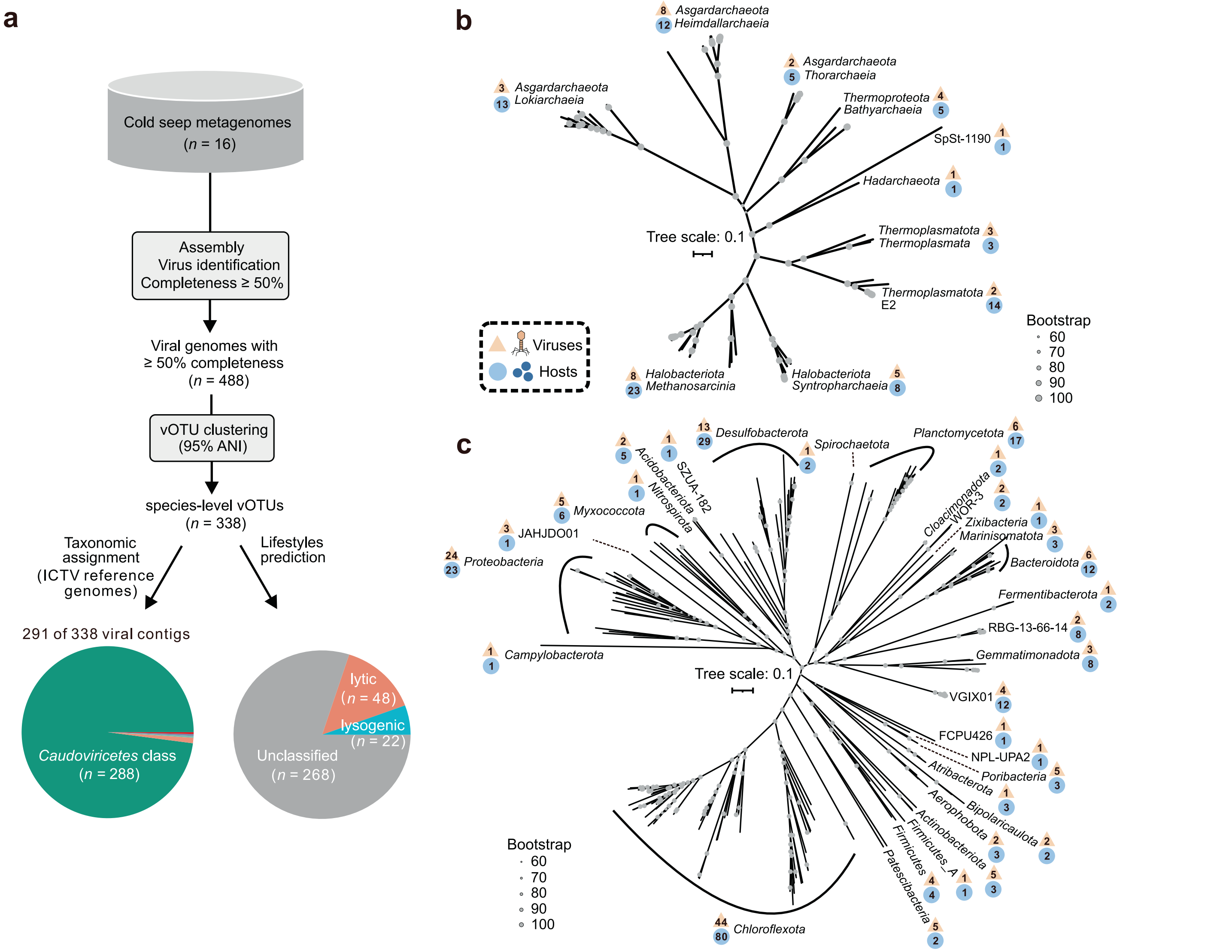

Fig. 2 Ecological features of cold seep viruses. a Workflow for identification, taxonomic assignment, and lifestyle prediction of viruses. Phylogenomic trees of predicted (**b**) archaeal and (**c**) bacterial hosts based on concatenated alignments of single-copy marker genes predicted by GTDB-Tk. Scale bars indicate the average number of substitutions per site. The orange triangle shows the number of viruses predicted to infect hosts in each clade, and the blue circle shows the number of microbial genomes in each clade with predicted viruses. Detailed statistics for taxonomy, lifestyles, and host-virus linkages are provided in Supplementary Tables 6 and 7.

system [49]. Additionally, a total of 17 viruses were found to encode two or more types of counter-defense systems.

As an important mechanism in adaptation to the environment, viruses can acquire new functional genes via transduction, namely auxiliary metabolic genes (AMGs) that contribute to host and/or viral fitness [4, 45]. Ten AMGs were identified in seven viral genomes (Fig. 3d, Supplementary Fig. 7 and Supplementary Table 9), related to four different types of functions. Two AMGs encoded GTP cyclohydrolase I (FolE), and six belonging to Que super family (QueC and QueD) may contribute to synthesizing GTP to 7-Cyano-7-deazaguanine (preQ$_0$) for genome modifications and translational efficiency [50]. The preQ$_0$ is the key intermediate in Q and G$^+$ pathways, which can be further modified for protecting viral DNA from host restriction enzymes [51]. AMGs encoding S-adenosylmethionine (SAM) decarboxylase (SpeD) and Dehydrogenase E1 component were also identified, and are involved in biosynthesis of amines or polyamines and the tricarboxylic acid cycle, respectively. SAM is the methyl donor for methyltransferases that modify DNA, RNA, histones, and other proteins; decarboxylation of SAM to S-adenosylmethioninamine might reduce the SAM required for methylation by host enzymes [52]. These AMGs have been also reported to be encoded by viruses in other deep-sea settings, including seawater and sediments of oceanic trenches, and free-living and particle-attached fractions from the bathypelagic ocean [45, 53–55], suggesting their importance roles in increasing viral adaptability in deep oceans.

Different classes of reverse transcriptases (RTs) were also found in 22 viruses, including diversity-generating retroelements (DGRs), retrons, UG26, and UG28 (Fig. 3e and Supplementary Fig. 8). Among them, RTs associated with DGRs were detected in five viruses; this mechanism can introduce variations in the target gene and facilitating the evolution of their hosts [17]. Retrons were found in three viruses, also possibly involved in defense systems for foreign DNA elements [49, 56]. Other RTs systems were identified with their roles and mechanisms remaining unknown.

Diverse lineages of viruses from different habitats have been seen to be self-beneficially employ alternative genetic codes to reassign one or more codons [20, 57–59]. In the dataset from the Haima cold seeps, 16 viral genomes are predicted to use genetic

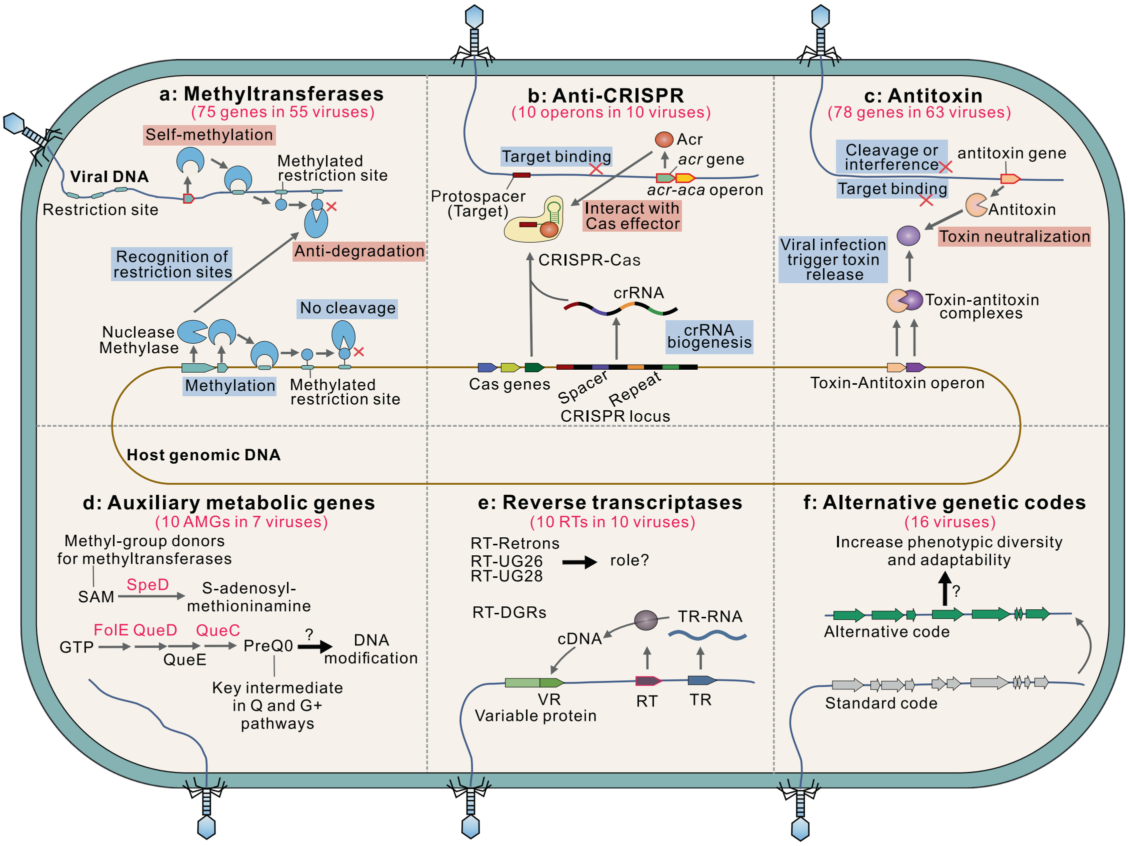

Fig. 3 Diverse strategies for environmental adaptation in cold seep viruses. a Viruses encode methylases that can modify their DNA to prevent its recognition by host restriction-modification systems and cleavage by certain restriction endonucleases. **b** Anti-CRISPR genes in viruses can inhibit CRISPR-Cas activities when it is targeted by the CRISPR-Cas system of the host. **c** Viruses encode antitoxins that can neutralize host toxin-antitoxin systems. **d** Potential functions of auxiliary metabolic genes. SAM: S-adenosylmethionine. preQ$_0$: 7-cyano-7-deazaguanine. **e** Reverse transcriptases (RTs) in cold seep viruses including diversity generating retroelements (DGRs), retrons, UG26, and UG28. For DGR, RT mediates exchange between two repeats: one serves as a donor template (TR) and the other as a recipient of variable sequence information (VR). **f** Alternative genetic codes found in some cold seep viral genomes. Related genes identified in cold seep viruses are marked in red (gene name) or with red border (gene arrow). Detailed statistics for diverse strategies for environmental adaptation in viruses are provided in Supplementary Tables 8–10.

codes characterized by reassignments of the *ochre* (TAA; $n = 620$ recoding events of genes), *amber* (TAG; $n = 182$) or *opal* (TGA; $n = 3$) stop codons (Fig. 3f, Supplementary Fig. 9a and Supplementary Table 10). These viruses are associated with hosts in multiple phyla (e.g., *Desulfobacterota* and *Acidobacteriota*). Genome sizes of these viruses range from 5.2 kb to 179.7 kb, with larger genomes having more recoding events of genes (linear regression; $R^2 = 0.58$, $p = 0.0004$). Recoded genes were mostly associated with replication, recombination and repair functions, followed by unknown functions (Supplementary Fig. 9b), suggesting adaptive recoding in controlling viral replication and regulation.

Cold seep viruses are genetically conserved and under strong purifying selection

Nucleotide diversity (π), single nucleotide polymorphisms (SNPs) and fixation indices (F_{ST}) were calculated to track viral microdiversity (Supplementary Tables 11 and 12). Nucleotide diversity of cold seep viral populations ranged from zero to 3.06×10^{-3}, and were on-average 1.29×10^{-4} (median 3.38×10^{-5}) for viruses detected in both active and extinct cold seep sediments (Fig. 4a). This viral nucleotide diversity is significantly lower than that observed for viral populations in seawater sampled from throughout the world's oceans (on-average 3.78×10^{-4}) [22] and in soils having various land uses (on-average 6.54×10^{-3}) [60]. Low SNP frequencies were also observed in Haima cold seep viral populations (0.33 SNP per 1000 bp on average, median 0.076; Fig. 4b), e.g., compared to those detected in the SARS-CoV-2 coronavirus, in 25 uncultivated virophage populations in North American freshwater lakes, and in 44 dsDNA viral populations dominating the oceans, based on various approaches for the extraction of viral genomes [61–63]. F_{ST} values between viral populations in relation to different sediment samples ranged from zero to 0.89 and were on-average 0.048, with 80% of pairwise fixation indices being zero (Fig. 4c). These data reflect that cold seep viral populations are genetically conserved and homogeneous contrary to observations of their microbial hosts [1], suggesting viruses and microbes might undergo different types of environmental selection.

Nucleotide diversity of viral populations is significantly different among viruses infecting different microbial hosts ($p = 0.0003$; Fig. 4d and Supplementary Table 11). Archaeal viruses associated with *Halobacteriota* have the highest nucleotide diversity. Like evolutionary trajectories of microbial populations in cold seeps [1] (e.g., *Asgardarchaeota*, *Halobacteriota*, and *Bacteroidota*), the nucleotide diversity of associated viruses is also depth-dependent in active cold seeps (linear regression; $R^2 = 0.21$, $p = 1.65 \times 10^{-5}$; Fig. 4e). On the other hand, no obvious depth-dependent trends were observed for viruses in the extinct cold seep (linear regression; $R^2 = -0.0048$, $p = 0.40$). This is in agreement with the significant difference for nucleotide diversity between the two cold seep stages (Fig. 4a; $p = 0.051$).

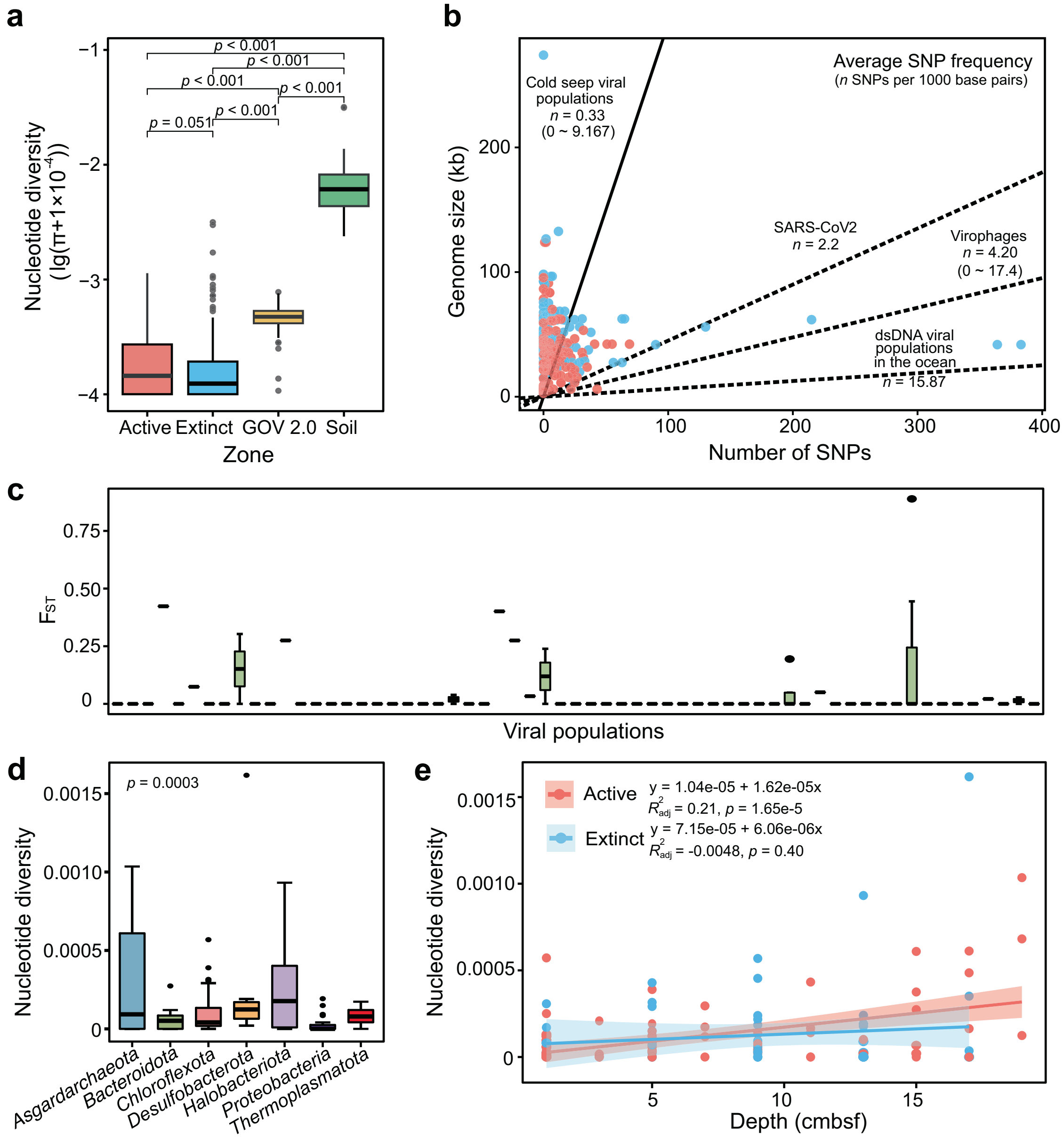

Fig. 4 Genome-wide evolutionary metrics of cold seep viral populations. **a** Nucleotide diversity of viruses from active and extinct cold seeps (this study), seawater sampled at a global scale (GOV 2.0) and soil samples from various land-use types. **b** Comparisons of average SNP frequency among cold seeps viruses, SARS-CoV-2, virophages, and dominant dsDNA populations in the oceans. Linear regressions are indicated for different viral groups. **c** Boxplot showing the F_{ST} values measured as the differences between the same viral populations found in two distinct cold seep samples. **d** Genome-wide nucleotide diversity of viruses infecting different microbial populations. **e** Comparison of nucleotide diversity of viruses infecting dominant populations against sediment depths at the genome level. Linear regressions and adjusted R^2 values are indicated for viruses in active and extinct cold seeps. Shadows of the lines represent 95% confidence intervals. Blue: extinct cold seep; red: active cold seep. Detailed statistics for evolutionary metrics of cold seep viruses from active and extinct cold seeps are provided in Supplementary Tables 11–12.

At the gene level, 90.6% of pN/pS ratios were less than 0.4, much lower ($p < 0.0001$) than those observed for viral assemblages present in underground saline waters from hypersaline springs [64] (Fig. 5a, Supplementary Fig. 10 and Supplementary Table 13), indicating that most cold seep viral genes were under strong purifying selection. However, genes under positive selection were also detected in relation to viral DNA replication, recombination, repair, and maturation (Fig. 5b), including genes encoding TerL, transposase, and leucyl-tRNA synthetase with abnormally high pN/pS values (Supplementary Table 13). Significant differences were exhibited for pN/pS ratios between the two cold seep stages (Fig. 5a; $p < 0.0001$). When grouped according to the functional categories of VOGDB (http://vogdb.org/), nucleotide diversity values were found to be significantly different while no significant differences were observed for pN/pS ratios (Supplementary Fig. 11). Tajima's D values ranged from −9.7 to zero and significantly varied ($p = 1.66 \times 10^{-8}$) between the two cold seep stages (Fig. 5c). A total of 90.5% of viral gene Tajima's D values were found to be zero with no detected SNP. For others, genes under natural selection (Tajima's $D < -2.5$; 6.1%) outnumbered those under neutral processes (Tajima's $D = 0$; 3.4%). The observation of large number of negative values supports the presence of excess rare alleles and recent expansion of cold seep viral populations [65].

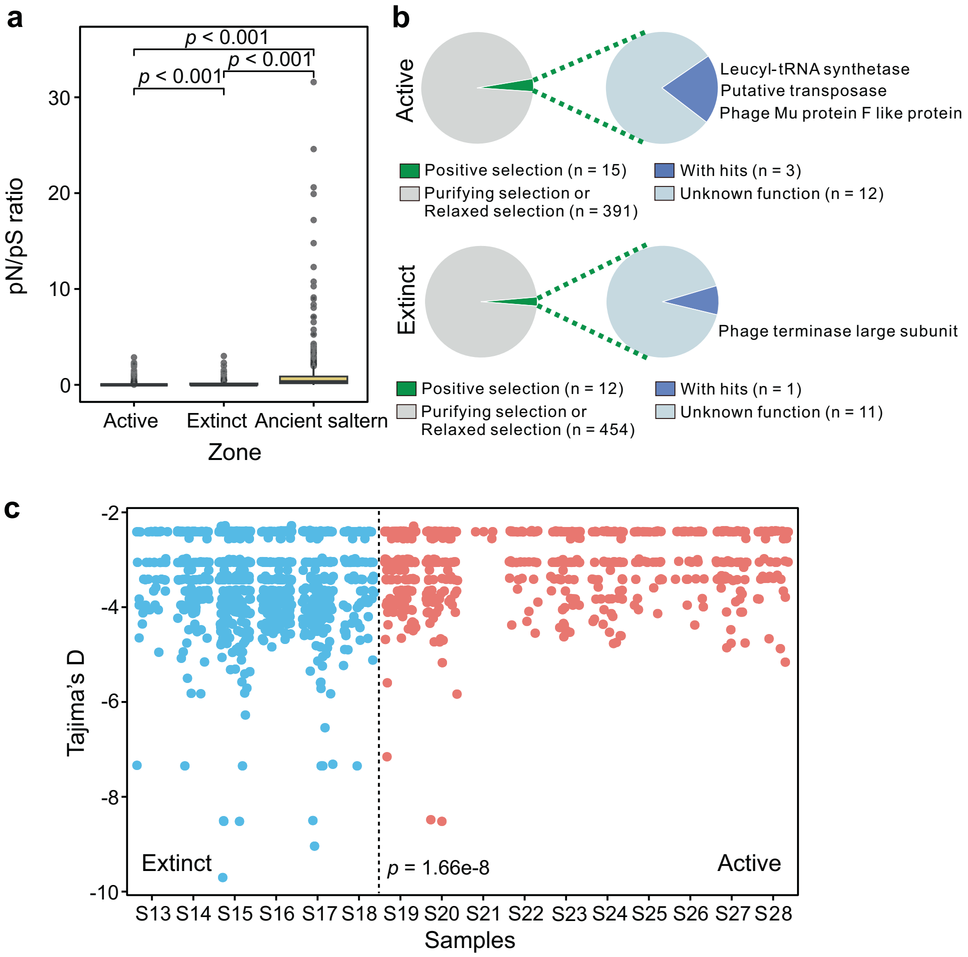

Fig. 5 Gene-wide evolutionary metrics of cold seep viral populations. **a** pN/pS ratio of viral genes from cold seeps (this study) and viral genes from an ancient saltern [64]. **b** Viral genes under positive selection in active and extinct cold seeps. Viral genes are divided into two groups based on pN/pS values, consisting of genes under positive selection (pN/pS≥1) and those under purifying selection or relaxed selection (pN/pS < 1). **c** Tajima's D of viral genes across 16 sediment samples from extinct (blue) and active (red) cold seeps. Detailed statistics for evolutionary metrics of cold seep viral genes are provided in Supplementary Table 13.

CONCLUSIONS

Previous studies of viral ecology and evolution have paid little attention to how subsurface viruses evolve to adjust to their surrounding environment and interact with their hosts [4, 25, 26]. Besides investigating structural and functional characteristics of viral communities, this study highlights evolutionary adaptation patterns of viruses at different sediment depths in cold seeps that are active and extinct. Novel and abundant deep-sea cold seep viruses were identified and observed to vary between active and extinct cold seeps and different sediment depths. These viruses are associated with major lineages of cold seep archaea and bacteria, including many taxonomic groups with no cultured representatives. Cold seep archaea and bacteria have various antiviral defense systems to prevent infections of diverse and abundant viruses, such as RM, AbiEii, SoFIC, and CRISPR-Cas systems. Likewise, their viruses have evolved to harbor a rich repertoire of adaptive strategies to defend against these host systems, including anti-CRISPR (Acr) proteins and methyltransferases. In addition to counter-defense systems to combat microbial hosts, cold seep viruses also contain RTs and AMGs that contribute to viral fitness, as well as alternative genetic code assignments to increase phenotypic diversity. Beyond genetically diverse features of cold seep viruses, their evolutionary trajectories are also surprisingly unique, featuring genetic conservation and homogenous genomes with unexpectedly low microdiversity. Most viral genes generally undergo strong purifying selection, in both the active and extinct cold seep sediments. These findings indicate that multiple factors are likely to determine the evolutionary patterns of cold seep viruses, including microbial hosts, sediment depth, and cold seep geology.

Together, these analyses of evolutionary dynamics of viruses will help guide future studies targeting the viral evolution and virus-host systems in extreme environments. However, it should be noted our results are representative only of double-stranded DNA viruses, such that other viral particles are not incorporated in the extraction process and analysis [9]. Nevertheless, studies with more samples from more locations and covering larger spatial gradients via the combination of metagenomes and viromes as well as single-virus genomics [23, 61] will be necessary to determine if the trends presented here are universal for deep sea subseafloor viral communities.

METHODS

Sample description, metagenomic sequencing and analysis

Metagenomic sequencing was performed on 16 sediment samples collected from the Haima cold seeps in the northern part of the South China Sea (Supplementary Fig. 1). Samples were taken from two active seep sites and one extinct seep site via the R/V Tansuo Yihao using the piloted submersible ShenHai YongShi [28]. Sediment cores penetrated

18 to 20 cm into the seabed. Details for DNA sequencing can be found elsewhere [28] and involved genomic DNA extraction with the MO BIO PowerSoil DNA Isolation Kit followed by sequencing on the MGI sequencing platforms DNBSEQ-T1 or BGISEQ500 (MGI Tech Co., Ltd., China) at BGI-Shenzhen (China).

For assessing microbial community composition, metagenomic reads were screened using singleM v0.13.2 (https://github.com/wwood/singlem) to extract *rplB* operational taxonomic units (OTUs). Quality-control of raw reads, assembly of clean reads into contigs, and generation of metagenome-assembled genomes (MAGs) used the metaWRAP [66] pipeline (v1.3) with details as reported previously [28]. Following depreciation using dRep v3.0.0 (parameters: -comp 50 -con 10 -sa 0.95) [67] a non-redundant set of 487 species-level MAGs was obtained. Taxonomic classifications of bacterial and archaeal genomes were assigned using GTDB-Tk v2.1.1 with the Genome Taxonomy Database using the R207_v2 reference package [68]. The set of 120 bacterial or 53 archaeal marker genes were identified, aligned, concatenated, and trimmed using GTDB-Tk v2.1.1. Genomes are then placed into the domain-specific trees using IQ-TREE v2.0.5 with best-fit models and 1000 ultrafast bootstraps [69, 70]. Bacterial and archaeal trees were visualized and beautified in the Interactive Tree Of Life (iTOL; v6) [71]. DefenseFinder v1.0.2 (parameters: -dbtype gembase) [8] was used to systematically detect antiviral defense systems in MAGs based on MacSyFinder models v1.2.0 in line with MacSyfinder rules [72].

Enumeration of viruses via fluorescence microscopy

Viral particles in sediments were counted by fluorescence microscopy according to a previous protocol [73]. In brief, around ~0.8 g sediment from each sample was transferred into a sterile 50 mL centrifuge tube and promptly fixed in 0.5% glutaraldehyde. Viruses were separated from sediments by vortexing in the dark, incubated in sodium pyrophosphate, and sonicated on ice. Samples were then filtered onto 0.02 µm pore-size membrane filters (Anodisc 25, Whatman), stained with SYBR Green I and observed using a HORIBA Aqualog fluorescence microscope (Tokyo, Japan) with a Leica imaging system. The Find maxima tool of Image J (https://imagej.net) was used to automatically select the fluorescent points [74] with manual curation.

Virus identification, vOTU clustering and taxonomic assignment

Potential single-contig viral genomes were identified from 18 metagenomic assemblies (contigs larger than 10 kb) using DeepVirFinder v1.0 [75], Virsorter2 v2.2.3 [76], and VIBRANT v1.2.1 [77]. Additionally, the MetaviralSPAdes module of SPAdes v3.15.2 was used to assemble viral contigs from metagenomic reads with default parameters [78]. CheckV v1.0.1 (database v1.1) [40] was applied to estimate the completeness and contamination of contigs identified ($n = 6,520$) using the above four methods. Genomes with ≥50% estimated completeness ($n = 488$) were clustered into species-level vOTUs according to MIUViG guidelines (95% average nucleotide identity; 85% aligned fraction) [41]. Clustering used the method for single-contig viral genomes [44] based on the supporting code of the CheckV v1.0.1 repository [15, 40]. Representative viral genomes for each species-level vOTU ($n = 338$) were clustered into viral clusters (VCs) that were roughly equivalent to ICTV (International Committee on Taxonomy of Viruses) prokaryotic viral genera using vConTACT2 v0.11.3 (parameters: --pcs-mode MCL --vcs-mode ClusterONE --rel-mode 'Diamond' --db 'ProkaryoticViralRefSeq94-Merged') enabled by gene-sharing networks [42]. The geNomad v1.3.3 pipeline (genomad end-to-end) [44, 79] was employed for the taxonomic assignment of viral genomes in accordance with the taxonomy contained in ICTV's VMR number 19 (https://ictv.global/). BACPHLIP v0.9.6 (with a minimum score of ≥0.8) [80] and VIBRANT v1.2.1 [77] were used to test if complete viral genomes were likely to be either temperate (lysogenic) or virulent (lytic). Remaining viral genomes were predicted to be lysogenic or unclassified depending on if they contained provirus integration sites or integrase genes based on the annotation provided with each genome.

Host assignments for bacteriophages and archaeoviruses

A total of 2678 bacterial and archaeal MAGs recovered from 68 previously sequenced cold seep sediments were used to serve as the host reference database [1]. Multiple host prediction strategies were used to link viral genomes to their microbial hosts following our previous method [7] complemented with iPHoP, an automated command-line pipeline for host predictions [81] (Supplementary Fig. 4). (i) For CRISPR spacer matches, the CRISPR arrays of cold seep microbial genomes were predicted using the CRISPRidentify v1.1.0 with default parameters [82]. Spacers shorter than 25 bp and CRISPR array with fewer than three spacers were dropped out. CRISPR spacers were aligned with viral genomes with ≤1 mismatch using BLASTn, and the thresholds of 95% identity were selected. Additionally, 1,398,130 spacers from 40,036 distinct genomes in the iPHoP_db_Sept21 database were also used for CRISPR-based predictions by version 1.1.0 of iPHoP [81]. (ii) For the detection of shared tRNA between viruses and hosts, tRNA genes were annotated using tRNAscan-SE v2.0.9 (parameters: -B -A) [83]. Putative host-virus linkages satisfied a threshold of ≥90% length identity over the 95% of the sequences by BLASTn. (iii) For alignment-based matches, viral genomes were aligned with microbial genomes using BLASTn based on their nucleotide sequence homology (e-value ≥ 0.001, nucleotide identity ≥70%, match coverage over the length of viral genomes ≥75% and bitscore ≥50). (iv) For host predictions based on independent signals (k-mer usage profiles and protein content), VirHost-Matcher (VHM) [84], WIsH [85], Prokaryotic virus Host Predictor (PHP) [86], and RaFAH [87] were performed individually using iPHoP v1.1.0. Match criteria were d_2^* values ≤ 0.2 for VHM, *p*-values ≤ 0.05 for WIsH, the predicted 'maxScoreHost' for PHP, and RaFAH_scores>0.14 for RaFAH. The genome was considered to be the host if it belonged to the same family with top hits for each viral genome based on multiple methods.

Identification of counter-defense systems, reverse transcriptases, auxiliary metabolic genes, and alternative genetic codes

For counter-defense systems, Acr-Aca operons were predicted based on the guilt-by-association approach using Acafinder (--Virus; version of Oct 15, 2022) [47]. Methyltransferases and restriction endonucleases of all types of restriction-modification (RM) systems were identified using previous hidden markov model profiles and scripts (https://github.com/oliveira-lab/RMS; version of Mar 16, 2018) [34]. Toxin and Antitoxin genes were identified based on specific domains of TA systems using Metafisher (https://github.com/JeanMainguy/MeTAfisher). Reverse transcriptases (RTs) were predicted and classified through the myRT web-server (https://omics.informatics.indiana.edu/myRT/) [14].

Auxiliary metabolic gene (AMG) identification was performed following previous protocols [7, 88]. Briefly, checkV-trimmed viral genomes were run through VirSorter2 (--prep-for-dramv) to produce the viral-affi-contigs-for-dramv.tab, and then the annotations were done using DRAM v1.2.0 (viral mode; default parameters) [89]. Genes with auxiliary scores of 1-3 and AMG flags of M and F were considered putative AMGs for further validation by manual checking of gene locations. PROSITE [90] was used to analyze the conserved domains of AMGs, and SWISS-MODEL [91] was used for protein structural predictions. Three-dimensional structures of viral AMGs were predicted using ColabFold by combining the fast homology search of MMseqs2 with AlphaFold2 [92, 93]. Genome maps of AMG-containing viral genomes were visualized based on DRAM-v annotations using Easyfig v.2.2.0 (ref. [94]).

Mgcod v1.0.0 was used to identify blocks with specific genetic codes for cold seep viral genomes (parameters: --isoforms) [95]. In this pipeline, MetaGeneMark [96] was applied to find the highest scoring model among four genetic code models: i) the standard genetic code (genetic code 11), ii) a model with the *opal* (TGA) reassignment (genetic code 4), iii) a model with the *amber* (TAG) reassignment (genetic code 15), and iv) a model with the *ochre* (TAA) reassignment (genetic code 101). Identified recoded regions were annotated using eggnog-mapper v2.1.9 (ref. [97]) against the eggNOG database (v5.0) [98].

Macro- and microdiversity analyses of viral populations

Filtered reads from each sample were mapped to 338 single-contig viral genomes that represent each vOTU using Bowtie2 v 2.3.5 [99]. Resulting BAM files, viral genomes, and read counts for each metagenome were used as input for the MetaPop pipeline [100] for pre-processing, macrodiversity and microdiversity analyses. MetaPop was run using the default parameters (--snp_scale both), and genes from viral genomes were predicted using Prodigal v2.6.3 [101]. Macrodiversity estimates include population abundances, alpha-diversity (within community) and beta-diversity (between community) indices. To accurately call SNPs and assess contig-level microdiversity, 207 viral populations with >10× average read depth coverage and >70% length of genome covered were retained for microdiversity analyses [100]. SNP frequencies subsampled down to 10× coverage were used to assess nucleotide diversity (θ and π) at the individual gene and whole-genome levels, as well as fixation indices (F_{ST}; between population microdiversity) and selective pressures on specific genes (pN/pS and Tajima's D).

Statistical analyses

Statistical analyses were performed using R v4.0.0. The normality and variance homogeneity of the data were assessed using Shapiro-Wilk and Bartlett's tests. Wilcoxon tests were used to compare differences in viral microdiversity parameters (π, Tajima's D, pN/pS) across cold seep stages. The Kruskal-Wallis rank-sum test with Chi-square correction was used to compare differences in evolutionary metrics of genomes and genes among different groups and samples. Correlations between microdiversity and sediment depth, defense system numbers, genome sizes, and others parameters were obtained using the linear regression with the fitness and confidence of the regression curves characterized as R^2 and p values, respectively.

DATA AVAILABILITY

MAGs, vOTUs, AMGs, and other related information have been uploaded to figshare (https://doi.org/10.6084/m9.figshare.22303213). Raw metagenomic sequencing data were deposited in NCBI under BioProject ID PRJNA739036.

REFERENCES

1. Dong X, Peng Y, Wang M, Woods L, Wu W, Wang Y, et al. Evolutionary ecology of microbial populations inhabiting deep sea sediments associated with cold seeps. Nat Commun. 2023;14:1127.
2. Dong X, Rattray JE, Campbell DC, Webb J, Chakraborty A, Adebayo O, et al. Thermogenic hydrocarbon biodegradation by diverse depth-stratified microbial populations at a Scotian Basin cold seep. Nat Commun. 2020;11:5825.
3. Suttle CA. Marine viruses-major players in the global ecosystem. Nat Rev Microbiol. 2007;5:801–12.
4. Wirth J, Young M. Viruses in subsurface environments. Annu Rev Virol. 2022;9:99–119.
5. Bryson SJ, Thurber AR, Correa AM, Orphan VJ, Vega Thurber R. A novel sister clade to the enterobacteria microviruses (family *Microviridae*) identified in methane seep sediments. Environ Microbiol. 2015;17:3708–21.
6. Kellogg CA. Enumeration of viruses and prokaryotes in deep-sea sediments and cold seeps of the Gulf of Mexico. Deep Sea Res Part II: Topical Stud Oceanogr. 2010;57:2002–7.
7. Li Z, Pan D, Wei G, Pi W, Zhang C, Wang JH, et al. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. ISME J. 2021;15:2366–78.
8. Tesson F, Herve A, Mordret E, Touchon M, d'Humieres C, Cury J, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun. 2022;13:2561.
9. Coutinho FH, Rosselli R, Rodriguez-Valera F. Trends of microdiversity reveal depth-dependent evolutionary strategies of viruses in the mediterranean. mSystems. 2019;4:e0055419.
10. Muscatt G, Cook R, Millard A, Bending GD, Jameson, E. Ecological and evolutionary patterns of virus-host interactions throughout a grassland soil depth profile. bioRxiv. 2022; https://doi.org/10.1101/2022.12.09.519740.
11. Samuel B, Burstein D. A diverse repertoire of anti-defense systems is encoded in the leading region of plasmids. bioRxiv. 2023; https://doi.org/10.1101/2023.02.15.528439.
12. Pawluk A, Davidson AR, Maxwell KL. Anti-CRISPR: discovery, mechanism and function. Nat Rev Microbiol. 2018;16:12–17.
13. Roux S, Paul BG, Bagby SC, Nayfach S, Allen MA, Attwood G, et al. Ecology and molecular targets of hypermutation in the global microbiome. Nat Commun. 2021;12:3076.
14. Sharifi F, Ye Y. Identification and classification of reverse transcriptases in bacterial genomes and metagenomes. Nucl Acids Res. 2022;50:e29.
15. Nayfach S, Paez-Espino D, Call L, Low SJ, Sberro H, Ivanova NN, et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. Nat Microbiol. 2021;6:960–70.
16. Paul BG, Bagby SC, Czornyj E, Arambula D, Handa S, Sczyrba A, et al. Targeted diversity generation by intraterrestrial archaea and archaeal viruses. Nat Commun. 2015;6:6585.
17. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, et al. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucl Acids Res. 2018;46:11–24.
18. Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T, et al. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature. 2007;449:83–86.
19. Peters SL, Borges AL, Giannone RJ, Morowitz MJ, Banfield JF, Hettich RL. Experimental validation that human microbiome phages use alternative genetic coding. Nat Commun. 2022;13:5710.
20. Borges AL, Lou YC, Sachdeva R, Al-Shayeb B, Penev PI, Jaffe AL, et al. Widespread stop-codon recoding in bacteriophages may regulate translation of lytic genes. Nat Microbiol. 2022;7:918–27.
21. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021;39:727–36.
22. Gregory AC, Zayed AA, Conceicao-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA viral macro- and microdiversity from pole to pole. Cell. 2019;177:1109–1123 e1114.
23. Martinez-Hernandez F, Diop A, Garcia-Heredia I, Bobay LM, Martinez-Garcia M. Unexpected myriad of co-occurring viral strains and species in one of the most abundant and microdiverse viruses on Earth. ISME J. 2022;16:1025–35.
24. Zhong Z-P, Vik D, Rapp J, Zablocki O, Maughan H, Temperton B, et al. Lower viral evolutionary pressure under stable versus fluctuating conditions in subzero Arctic brines. Microbiome. 2023;11:174.
25. Anderson RE. Tracking microbial evolution in the subseafloor biosphere. mSystems. 2021;6:e0073121.
26. Biddle JF, Sylvan JB, Brazelton WJ, Tully BJ, Edwards KJ, Moyer CL, et al. Prospects for the study of evolution in the deep biosphere. Front Microbiol. 2011;2:285.
27. Manrubia S, Lazaro E. Viral evolution. Phys Life Rev. 2006;3:65–92.
28. Li J, Dong X, Tang Y, Zhang C, Yang Y, Zhang W, et al. Deep sea cold seep is an atmospheric Hg sink and MeHg source. Research Square. 2022; https://doi.org/10.21203/rs.3.rs-2323106/v1.
29. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science. 2018;359:eaar4120.
30. Payne LJ, Meaden S, Mestre MR, Palmer C, Toro N, Fineran PC, et al. PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. Nucl Acids Res. 2022;50:W541–550.
31. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. Nat Rev Microbiol. 2020;18:113–9.
32. Millman A, Melamed S, Leavitt A, Doron S, Bernheim A, Hor J, et al. An expanded arsenal of immune systems that protect bacteria from phages. Cell Host Microbe. 2022;30:1556–1569 e1555.
33. Cury J, Bernheim A. CRISPR-Cas and restriction-modification team up to achieve long-term immunity. Trends Microbiol. 2022;30:513–4.
34. Seong HJ, Roux S, Hwang CY, Sul WJ. Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. Microbiome. 2022;10:157.
35. Bernheim A, Bikard D, Touchon M, Rocha EPC. A typical organizations and epistatic interactions of CRISPRs and cas clusters in genomes and their mobile genetic elements. Nucl Acids Res. 2020;48:748–60.
36. Dy RL, Przybilski R, Semeijn K, Salmond GP, Fineran PC. A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism. Nucl Acids Res. 2014;42:4590–605.
37. Horn H, Slaby BM, Jahn MT, Bayer K, Moitinho-Silva L, Forster F, et al. An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. Front Microbiol. 2016;7:1751.
38. Hiraoka S, Sumida T, Hirai M, Toyoda A, Kawagucci S, Yokokawa T, et al. Diverse DNA modification in marine prokaryotic and viral communities. Nucl Acids Res. 2022;50:1531–50.
39. Hwang Y, Roux S, Coclet C, Krause SJE, Girguis PR. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. Nat Microbiol. 2023;8:946–57.
40. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39:578–85.
41. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, et al. Minimum information about an uncultivated virus genome (MIUViG). Nat Biotechnol. 2019;37:29–37.
42. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019;37:632–9.
43. Chen Y, Lyu Y, Zhang J, Li Q, Lyu L, Zhou Y, et al. Riddles of lost city: Chemotrophic prokaryotes drives carbon, sulfur, and nitrogen cycling at an extinct cold seep, South China Sea. Microbiol Spectr. 2023;11:e0333822.
44. Camargo AP, Nayfach S, Chen IA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucl Acids Res. 2023;51:D733–D743.

45. Jian H, Yi Y, Wang J, Hao Y, Zhang M, Wang S, et al. Diversity and distribution of viruses inhabiting the deepest ocean on Earth. ISME J. 2021;15:3094–110.
46. Cheng R, Li X, Jiang L, Gong L, Geslin C, Shao Z. Virus diversity and interactions with hosts in deep-sea hydrothermal vents. Microbiome. 2022;10:235.
47. Yang B, Zheng J, Yin Y. AcaFinder: Genome mining for anti-CRISPR-associated genes. mSystems. 2022;7:e0081722.
48. Landsberger M, Gandon S, Meaden S, Rollie C, Chevallereau A, Chabas H, et al. Anti-CRISPR phages cooperate to overcome CRISPR-Cas immunity. Cell. 2018;174:908–916 e912.
49. Bobonis J, Mitosch K, Mateus A, Karcher N, Kritikos G, Selkrig J, et al. Bacterial retrons encode phage-defending tripartite toxin-antitoxin systems. Nature. 2022;609:144–50.
50. Ma R, Lai J, Chen X, Wang L, Yang Y, Wei S, et al. A novel phage infecting Alteromonas represents a distinct group of Siphophages infecting diverse aquatic copiotrophs. mSphere. 2021;6:e0045421.
51. Hutinet G, Kot W, Cui L, Hillebrand R, Balamkundu S, Gnanakalai S, et al. 7-Deazaguanine modifications protect phage DNA from host restriction systems. Nat Commun. 2019;10:5442.
52. Zhang J, Zheng YG. SAM/SAH analogs as versatile tools for SAM-dependent methyltransferases. ACS Chem Biol. 2016;11:583–97.
53. Zhou YL, Mara P, Vik D, Edgcomb VP, Sullivan MB, Wang Y. Ecogenomics reveals viral communities across the Challenger Deep oceanic trench. Commun Biol. 2022;5:1055.
54. Coutinho FH, Silveira CB, Sebastian M, Sanchez P, Duarte CM, Vaque D, et al. Water mass age structures the auxiliary metabolic gene content of free-living and particle-attached deep ocean viral communities. Microbiome. 2023;11:118.
55. Chen P, Zhou H, Huang Y, Xie Z, Zhang M, Wei Y, et al. Revealing the full biosphere structure and versatile metabolic functions in the deepest ocean sediment of the Challenger Deep. Genome Biol. 2021;22:207.
56. Bobonis J, Mitosch K, Mateus A, Kritikos G, Elfenbein JR, Savitski MM et al. Phage proteins block and trigger retron toxin/antitoxin systems. bioRxiv. 2020; https://doi.org/10.1101/2020.06.22.160242.
57. Neri U, Wolf YI, Roux S, Camargo AP, Lee B, Kazlauskas D, et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell. 2022;185:4023–4037 e4018.
58. Wolf YI, Silas S, Wang Y, Wu S, Bocek M, Kazlauskas D, et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. Nat Microbiol. 2020;5:1262–70.
59. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat Commun. 2021;12:1044.
60. Liao H, Li H, Duan CS, Zhou XY, Luo QP, An XL, et al. Response of soil viral communities to land use changes. Nat Commun. 2022;13:6027.
61. Martinez-Hernandez F, Fornas O, Lluesma Gomez M, Bolduc B, de la Cruz Pena MJ, Martinez JM, et al. Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat Commun. 2017;8:15892.
62. Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. Nat Commun. 2017;8:858.
63. Taiwo IA, Adeleye N, Anwoju FO, Adeyinka A, Uzoma IC, Bankole TT. Sequence analysis for SNP detection and phylogenetic reconstruction of SARS-cov-2 isolated from Nigerian COVID-19 cases. N. Microbes N. Infect. 2022;45:100955.
64. Ramos-Barbero MD, Viver T, Zabaleta A, Senel E, Gomariz M, Antiguedad I, et al. Ancient saltern metagenomics: tracking changes in microbes and their viruses from the underground to the surface. Environ Microbiol. 2021;23:3477–98.
65. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.
66. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome. 2018;6:158.
67. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.
68. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics. 2022;38:5315–6.
69. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
70. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.
71. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucl Acids Res. 2021;49:W293–W296.
72. Abby SS, Neron B, Menager H, Touchon M, Rocha EP. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS One. 2014;9:e110726.
73. Cai L, Jorgensen BB, Suttle CA, He M, Cragg BA, Jiao N, et al. Active and diverse viruses persist in the deep sub-seafloor sediments over thousands of years. ISME J. 2019;13:1857–64.
74. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9:671–5.
75. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020;8:64–77.
76. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9:37.
77. 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. 2020;8:90.
78. Antipov D, Raiko M, Lapidus A, Pevzner PA. Metaviral SPAdes: assembly of viruses from metagenomic data. Bioinformatics. 2020;36:4126–9.
79. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. You can move, but you can't hide: identification of mobile genetic elements with geNomad. bioRxiv. 2023; https://doi.org/10.1101/2023.03.05.531206.
80. Hockenberry AJ, Wilke CO. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ 2021;9:e11396.
81. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: an integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLoS Biol. 2023;21:e3002083.
82. Mitrofanov A, Alkhnbashi OS, Shmakov SA, Makarova KS, Koonin EV, Backofen R. CRISPRidentify: identification of CRISPR arrays using machine learning approach. Nucl Acids Res. 2021;49:e20.
83. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucl Acids Res. 2021;49:9077–96.
84. Ahlgren NA, Ren J, Lu YY, Fuhrman JA, Sun F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucl Acids Res. 2017;45:39–53.
85. Galiez C, Siebert M, Enault F, Vincent J, Soding J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics. 2017;33:3113–4.
86. Lu C, Zhang Z, Cai Z, Zhu Z, Qiu Y, Wu A, et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. BMC Biol. 2021;19:5.
87. Coutinho FH, Zaragoza-Solas A, Lopez-Perez M, Barylski J, Zielezinski A, Dutilh BE, et al. RaFAH: host prediction for viruses of bacteria and archaea based on protein content. Patterns. 2021;2:100274.
88. Pratama AA, Bolduc B, Zayed AA, Zhong ZP, Guo J, Vik DR, et al. Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. PeerJ 2021;9:e11447.
89. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucl Acids Res. 2020;48:8883–8900.
90. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucl Acids Res. 2013;41:D344–347.
91. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucl Acids Res. 2018;46:W296–W303.
92. Mirdita M, Schutze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19:679–82.
93. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
94. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27:1009–10.
95. Pfennig, A, Lomsadze, A & Borodovsky, M Annotation of phage genomes with multiple genetic codes. bioRxiv. (2022); 2022.2006.2029.495998.
96. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucl Acids Res. 2010;38:e132.
97. Cantalapiedra CP, Hernandez-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38:5825–9.
98. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucl Acids Res. 2019;47:D309–D314.
99. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

100. Gregory AC, Gerhardt K, Zhong ZP, Bolduc B, Temperton B, Konstantinidis KT, et al. MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. Microbiome 2022;10:49.

101. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 2010;11:119.

ACKNOWLEDGEMENTS

The work was supported by the Scientific Research Foundation of Third Institute of Oceanography, MNR (No. 2023022 and No. 2022025), the National Key Research and Development Program of China (2022YFC2805500), the Natural Science Foundation of Fujian Province (No. 2023J06042), State Key Laboratory of Marine Geology, Tongji University (No. MGK202303), and China Postdoctoral Science Foundation (2022M723709). We thank Dr. Xiaoyuan Feng and Dr. Min Jin for providing valuable comments.

AUTHOR CONTRIBUTIONS

XD designed this study. XD, YP, and ZL performed the omics analysis. ZZ and ZL counted virus particles in sediments. XD, YP, DP, L-DS, QL, CZ, KJ, and CRJH interpreted the data. JL contributed to data collection. XD, YP, and CRJH wrote the paper, with input from other authors.

COMPETING INTERESTS

The authors declare no competing interest.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41396-023-01491-0.

Correspondence and requests for materials should be addressed to Xiyang Dong.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.