Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity

Zexin Li[1 #], Donald Pan[2], Guangshan Wei[1, 3], Weiling Pi[1], Jiang-Hai Wang[1], Yongyi Peng[1], Lu Zhang[4, 5], Yong Wang[6], Casey R.J. Hubert[7], Xiyang Dong[1 #] *

[1] School of Marine Sciences, Sun Yat-Sen University, Zhuhai, 519082, China

[2] Department of Ecology and Environmental Studies, The Water School, Florida Gulf Coast University, Fort Myers, FL 33965, USA

[3] Key Laboratory of Marine Genetic Resources, Third Institute of Oceanography, Ministry of Natural Resources, Xiamen, 361005, China

[4] Institute of Advanced Technology, Westlake Institute for Advanced Study, Hangzhou, 310024, China

[5] School of Engineering, Westlake University, Hangzhou, 310024, China

[6] Department of Life Science, Institute of Deep-sea Science and Engineering, Chinese Academy of Sciences, Sanya, 572000, China

[7] Department of Biological Sciences, University of Calgary, Calgary, AB T2N1N4, Canada

[#] These authors contributed equally to this work.

* Correspondence can be addressed to:
Assoc Prof Xiyang Dong (dongxy23@mail.sysu.edu.cn)

Abstract

In marine ecosystems, viruses exert control on the composition and metabolism of microbial communities, thus influencing overall biogeochemical cycling. Deep sea sediments associated with cold seeps are known to host taxonomically diverse microbial communities, but little is known about viruses infecting these microorganisms. Here, we probed metagenomes from seven geographically diverse cold seeps across global oceans, to assess viral diversity, virus-host interaction, and virus-encoded auxiliary metabolic genes (AMGs). Gene-sharing network comparisons with viruses inhabiting other ecosystems reveal that cold seep sediments harbour considerable unexplored viral diversity. Most cold seep viruses display high degrees of endemism with seep fluid flux being one of the main drivers of viral community composition. *In silico* predictions linked 14.2% of the viruses to microbial host populations, with many belonging to poorly understood candidate bacterial and archaeal phyla. Lysis was predicted to be a predominant viral lifestyle based on lineage-specific virus/host abundance ratios. Metabolic predictions of prokaryotic host genomes and viral AMGs suggest that viruses influence microbial hydrocarbon biodegradation at cold seeps, as well as other carbon, sulfur and nitrogen cycling via virus-induced mortality and/or metabolic augmentation. Overall, these findings reveal the global diversity and biogeography of cold seep viruses and indicate how viruses may manipulate seep microbial ecology and biogeochemistry.

Introduction

Marine cold seeps are typically found at the edges of continental shelves and feature mainly of gaseous and liquid hydrocarbons from deep geologic sources[1, 2]. Seep fluids may come from thermogenic oil and gas systems that have been present for long periods of time in lower strata, indicating underlying oil and gas reservoirs[3, 4]. In the context of global climate change, methane and other short-chain alkanes escaping from deep sea cold seep sediments can reach the atmosphere, exacerbating the greenhouse effect[5]. Understanding the biogeochemical cycling in marine sediments associated with cold seeps is thus important for meeting critical energy and climate challenges.

Cold seeps are a chemosynthetic ecosystem and contain an extensive diversity of archaea and bacteria which play important roles in hydrocarbon metabolism[6, 7]. These microbial populations are not only highly active in influencing seep biogeochemistry at the sediment-water interface[8], but also contribute to a variety of biological processes such as sulfate reduction, sulfur oxidation, denitrification, metal reduction and methanogenesis within the seabed[2, 8]. Viruses have also been observed in cold seep sediments. Epifluorescence microscopy of sediments from the Gulf of Mexico revealed that viral-like particle counts and virus-to-prokaryote ratios at cold seeps were significantly higher than in surrounding sediments, suggesting these habitats may be hot spots for viruses[9]. This agrees with elevated microbial activity at cold seeps driven by the availability of energy-rich substrates supplied from below. In addition, novel viruses have also been discovered in methane seep sediments[10]. These findings suggest that cold seeps harbour abundant and undiscovered viruses potentially influencing their microbial hosts and consequently, biogeochemical cycling at cold seeps.

Knowledge of the ecological roles of viruses in deep sea sediments has been limited by difficulties in sampling and extracting viral particles (virions) from sediments[11]. In recent years, developments in sequencing and bioinformatics have enabled the analysis of viruses recovered from metagenomes sequenced without prior virion separation. These methods have greatly advanced viral ecology from the identification of novel viruses to the global distribution of viruses. Studies from a variety of environments such as thawing permafrost[12], mangroves[13], arctic lakes[14], freshwater lakes[15], and particularly seawater[16-18] have suggested that prokaryotic viruses act as key agents in natural ecosystems via a range of interactions with their microbial hosts. Viruses can influence organic carbon and nutrient turnover by top-down control of microbial abundance via lysis of cells and the subsequent release of cellular contents during lytic infection[19]. They can also reprogram host metabolism through horizontal gene transfer, or via auxiliary metabolic genes (AMGs) in their genomes that are expressed during infection. In peatland soils along a permafrost thaw gradient in Sweden, virus-encoded glycoside hydrolases were found to play a role in complex carbon degradation[12]. In freshwater lakes fed with sediment-derived methane, some viruses were found to encode subunits of particulate methane monooxygenase, suggesting that they may augment bacterial aerobic methane oxidation during infection[20]. In recent years, studies are starting to reveal the presence and abundance of viruses in deep sea sediments[11, 21, 22], thus deep sea sediments associated with cold seeps present a unique opportunity to study viruses and their interactions with hosts in a chemosynthetic ecosystem often dominated by anaerobic methane oxidation. Several metagenomic sequencing efforts have been undertaken on cold seep sediments[2] such that the extracted DNA includes genomes of viruses in these sediments, yet most studies have focused exclusively on genomes bacteria and archaea, neglecting the viruses.

In this study, we sought to expand the understanding of viral diversity and the ecological role of viruses in deep sea sediments associated with cold seeps. To this end, 28 publicly available marine sediment metagenomes from seven cold seeps around the world were analyzed to recover genomes of viruses in cold seep communities. Characterizing the diversity of these viral communities enabled predictions about the organisms they may be infecting and identification of AMGs potentially mediating ecological roles of viruses in these habitats. Our findings reveal the global diversity and biogeography of seep viruses and their role in benthic microbial ecology and biogeochemistry.

Methods

Collection of metagenomic datasets for deep sea cold seeps

Metagenomic data sets were compiled from 28 sediment samples collected from seven cold seep sites across the global oceans (**Figure 1)**. These sites were: Haakon Mosby mud volcano (HM); Eastern North Pacific ODP site 1244 (ENP); Mediterranean Sea, Amon mud volcano (MS); Santa Monica Mounds (SMM); Eastern Gulf of Mexico (EGM); Scotian Basin (SB); and Western Gulf of Mexico (WGM) (**Supplementary Table 1 and references therein**). Except for EGM and SB, metagenomic datasets along with metadata were downloaded from NCBI Sequence Read Archive and NCBI BioSample databases (https://www.ncbi.nlm.nih.gov). Sample collection and DNA sequencing of samples from EGM and SB are described in detail elsewhere[23, 24].

Taxonomic profiling of microbial communities

To explore the prokaryotic composition of each sample, 16S rRNA gene fragments (i.e. miTags) were extracted from metagenomic raw reads using the phyloFlash pipeline[25]. Extracted 16S miTags were mapped to the SILVA SSU rRNA reference database (v132)[26] and assigned an approximate taxonomic affiliation (nearest taxonomic unit, NTU).

Metagenomic assembly

Raw reads were quality-controlled by trimming primers and adapters and filtering out artifacts and low-quality reads using the Read_QC module within the metaWRAP pipeline[27]. Quality-controlled reads from each metagenome were individually assembled using MEGAHIT v1.1.3[28] (default parameters). Short contigs (<1000 bp) were removed.

Generation of prokaryotic metagenome-assembled genomes

For each assembly, contigs were binned using the binning module (parameters: --maxbin2 --metabat1 --metabat2) and consolidated into a final bin set using the Bin_refinement module (parameters: -c 50 -x 10) within metaWRAP[27]. All the produced bin sets were aggregated and dereplicated at 95% average nucleotide identity (ANI) using dRep v2.3.2 (parameters: -comp 50 -con 10 -sa 0.95)[29], resulting in a total of 592 species-level metagenome-assembled genomes (MAGs). Taxonomy of each MAG was initially assigned using GTDB-Tk v0.3.3[30] based on the Genome Taxonomy Database (GTDB, http://gtdb.ecogenomic.org) taxonomy R04-RS89[31]. The results were further refined using maximum-likelihood phylogeny inferred from a concatenation of 120 bacterial or 122 archaeal marker genes produced by GTDB-Tk. Bacterial and archaeal trees were built using RAxML v8[32] called as follows: raxmlHPC-HYBRID -f a -n result -s input -c 25 -N 100 -p 12345 -m PROTCATLG -x 12345. Genomes were finally classified using the naming system of the NCBI taxonomy[33].

Identification of viral contigs

Viral contigs were recovered from metagenome assemblies using VirSorter v1.0.5[34] and VirFinder v1.1[35]. Only contigs ≥10 kb were retained, based on the following criteria[17]: (1) VirSorter categories 1, 2, 4 and 5; (2) VirFinder score ≥0.9 and *p*<0.05; (3) both VirSorter categories 1-6 and VirFinder score ≥0.7 and *p*<0.05. The identified contigs from each assembly were then compiled and clustered at 95% nucleotide identity using CD-HIT v4.8.1 (parameters: -c 0.95 -d 400 -T 20 -M 20000 -n 5)[36], producing 2885 viral OTUs (vOTUs). These may represent a mixture of free viruses, proviruses and/or actively infecting viruses[12]. Completeness of viral genomes was estimated using the CheckV pipeline[37]. CheckV and VIBRANT v1.2.1[38] were used to infer temperate lifestyles by identifying viral contigs that contain provirus integration sites or integrase genes.

Comparisons to viral sequences from other environments by protein clustering

To place the 2885 vOTUs in broader context, they were compared to viral contigs in public databases: (i) GOV 2.0 seawater[17] (n=195728); (ii) wetland sediment[39] (n=1212); (iii) Stordalen thawing permafrost[12] (n=1896). For each viral contig, open reading frames (ORFs) were called using Prodigal v2.6.3[40] and the predicted protein sequences were used as input for vConTACT2[41]. We followed the protocol published in protocols.io (https://www.protocols.io/view/applying-vcontact-to-viral-sequences-and-visualizi-x5xfq7n) for the application of vConTACT2 and visualization of the gene-sharing network in Cytoscape v3.7.2[42] (edge-weighted spring-embedded model). Viral RefSeq (v85) was selected as the reference database, and Diamond was used for the protein-protein similarity method. Other parameters were set as default.

Viral taxonomic assignments

To identify the taxonomic affiliations of the vOTUs, ORFs predicated from Prodigal v2.6.3 were aligned against the viral NCBI Viral RefSeq V94 using BLASTp (E-value of <0.0001, bitscore ≥50)[13, 17, 43]. The BLASTp output was then imported into MEGAN v6.17.0 using the Lowest Common Ancestor (LCA) algorithm for taxonomic analysis[44].

Abundance profiles

RPKM (Reads per kilobase per million mapped reads) values were used to represent relative abundances of viruses and microorganisms. To calculate the RPKM values of each viral contig or MAG, quality-controlled reads from each sample were mapped to a viral contig database or to contigs compiled from the 592 MAGs with BamM v1.7.3 'make' (https://github.com/Ecogenomics/BamM). Low quality mappings were removed with BamM v1.7.3 'filter' (parameters: --percentage_id 0.95 --percentage_aln 0.75). Filtered bam files were then passed to CoverM v0.3.1 (https://github.com/wwood/CoverM) to generate coverage profiles across samples (parameters: contig mode for viral contigs, genome mode for MAGs, --trim-min 0.10 --trim-max 0.90 --min-read-percent-identity 0.95 --min-read-aligned-percent 0.75 -m rpkm).

Virus-host prediction

Four different *in silico* methods[12, 39, 45] were used to predict virus-host interactions. *(1) Nucleotide sequence homology*. Sequences of vOTUs and prokaryotic MAGs were compared using BLASTn. Match criteria were ≥75% coverage over the length of the viral contig, ≥70% minimum nucleotide identity, ≥50 bit score, and ≤0.001 e-value. *(2)*

Oligonucleotide frequency (ONF). VirHostMatcher v1.0[46] was run with default parameters, with d_2^* values ≤0.2 being considered as a match. *(3) Transfer RNA (tRNA) match*. Identification of tRNAs from prokaryotic MAGs and vOTUs was performed with ARAGORN v1.265 using the '-t' option[47]. Match requirements were ≥90% length identity in ≥90% of the sequence by BLASTn[18]. *(4) CRISPR spacer match*. CRISPR arrays were assembled from quality-controlled reads using crass v1.0.1 with default parameters[48]. CRISPR spacers were then matched against viral contigs with ≤1 mismatch over the complete length of the spacer using BLASTn. For each matching CRISPR spacer, the repeat from the same assembled CRISPR array was compared against the prokaryotic MAGs using BLASTn with the same parameters, creating a virus-host link. Among potential linkages, *cas* genes of putative microbial hosts were inspected further using MetaErg v1.2.2[49]. Only hits with adjacent *cas* genes were regarded as highly confident signals.

Whenever multiple hosts for a vOTU were predicted, the virus-host linkage supported by multiple approaches was chosen. Otherwise, virus-host linkage determination used a previously-reported[12] priority order of: (1) CRISPR spacer match with adjacent *cas* gene; (2) CRISPR spacer match without adjacent *cas* gene; (3) tRNA match or nucleotide sequence homology; (4) ONF comparison

Functional annotations of MAGs

Each MAG was first annotated using MetaErg v1.2.2[49]. The predicted amino sequences were then used as query for identification of key metabolic markers via METABOLIC v2.0[50]. For phylogenetic analysis of McrA and DsrA, amino acid sequences were aligned using the MUSCLE algorithm[51] included in MEGA X[52]. All positions with less than 95% site coverage were eliminated. The maximum-likelihood phylogenetic tree was constructed in MEGA X using the JTT matrix-based model. The tree was bootstrapped with 50 replicates and midpoint-rooted.

Identification of auxiliary metabolic genes

AMGs were identified based on KEGG, Pfam and VOG annotations using a combination of VIBRANT v1.2.1 and METABOLIC v2.0[50] with default parameters. Manual inspection was used to remove non-AMG annotations. CAZyme (carbohydrate-active enzyme) genes were identified on the dbCAN2[53] web server based on the recognition of the CAZyme signature domain found by at least two out of three tree tools (HMMER + DIAMOND + Hotpep).

Statistical analyses

All statistical analyses were performed in R version 3.6.3. Alpha and beta diversity of viral communities were calculated using vegan package v2.5-6[54]. Shapiro-Wilk and Bartlett's tests were employed to test the data normality and homoscedasticity prior to other statistical analysis. For beta diversity of viral communities, non-metric multidimensional scaling (NMDS) was used to reduce dimensionality using the function capscale with no constraints applied. NMDS was based on Bray-Curtis dissimilarities generated from OTU tables with viral abundances (RPKM) using the vegdist function (method ‘‘bray’’). The grouping of cold seep sites into different types[2] (mineral-prone vs mud-prone) was individually verified using Analysis of similarity (ANOSIM). For comparison between cold seep sites, Shannon index was compared using Analysis of Variance (ANOVA) while Simpson and Chao1 indices were compared using a Kruskal-Wallis nonparametric test. For comparison of cold seep systems, Shannon index was compared using Student's T-test while Simpson and Chao1 indices were compared using Wilcoxon signed-rank test. Pearson correlations were performed using the cor function.

Results and Discussion

To investigate the diversity and ecological function of viruses inhabiting cold seep sediments, a 0.38 Tbp compilation of metagenomic data was recruited from public databases and analysed (**Supplementary Table 1**). Metagenomes were sequenced from 28 sediment samples obtained at seven seabed cold seeps across the global oceans, encompassing gas hydrates, oil and gas seeps, mud volcanoes and asphalt volcanoes (**Figure 1**).

Overview of bacterial and archaeal communities

To assess the overall microbial community structure in these sediments, 16S miTags were extracted from metagenomic reads for taxonomic profiling[25]. Classification of 16S miTags at the phylum level (class level for Proteobacteria) revealed dominant bacterial lineages to be Chloroflexi (on average 23% of bacterial 16S miTags from 28 samples), Atribacteria (23%), *Gammaproteobacteria* (9%), *Deltaproteobacteria* (9%), and Planctomycetes (6%) (**Supplementary Figure 1**). In shallow sediments (<0.2 meters below the sea floor; mbsf), *Gammaproteobacteria* and *Deltaproteobacteria* were present in higher relative abundance, whereas Atribacteria and Chloroflexi predominated in deeper sediments that made up the majority of the sample set. For archaeal lineages, members from *Methanomicrobia* (phylum Euryarchaeota) were on average 30% of archaeal miTags, followed by Bathyarchaeota (TACK group) at 18%, and Lokiarchaeota (Asgard group) at 16% (**Supplementary Figure 2**).

Assembly and binning of metagenomes resulted in 592 high- or medium-quality[55] microbial MAGs clustering at 95% ANI, nominally representing species-level groups[56]. These 460 bacterial and 132 archaeal MAGs spanned 46 known and four unclassified phyla (**Figure 2a** and **Supplementary Table 2**). Within the domain Bacteria, members of Chloroflexi (n=119 MAGs), *Deltaproteobacteria* (n=67) and Planctomycetes (n=44) were highly represented. Within the domain Archaea, MAGs were mainly affiliated with *Methanomicrobia* (n=41), Bathyarchaeota (n=21) and Lokiarchaeota (n=18). Based on the read coverage of MAGs among the samples, no single MAG was found to be present in all seven cold seeps (**Supplementary Table 3**). All seven regions harboured MAGs belonging to *Deltaproteobacteria* (n=67), Planctomycetes (n=44), WOR-3 (n=15), Bacteroidetes (n=14), Heimdallarchaeota (n=12) and Atribacteria (n=11).

Viruses from cold seep sediments are diverse and novel

From the 28 bulk shotgun metagenomes, 39154 putative viral sequences were obtained, manually filtered and then clustered at 95% ANI to represent approximately species-level taxonomy[17, 57]. This gave rise to 2885 non-redundant cold seep vOTUs, each represented by contigs ≥10kb in size, including four that were ≥200 kb (**Supplementary Table 4**) possibly corresponding to huge viruses[58]. Completeness of metagenome-assembled viral genomes or genome fragments was estimated using CheckV[37], giving rise to four different quality tiers: complete genomes (10.3% vTOUs), high-quality (4.3%), medium-quality (11.3%), and low-quality (57.7%), with the remainder 16.4% being undetermined (**Supplementary Figure 3**).

Cold seep vOTUs were abundant across all sediment samples (**Supplementary Table 5**), however a large majority (84%) of vOTUs were only present within a single cold seep site. Further analysis of viral distribution across the seven cold seep sites (ANOSIM, R=0.802, p=0.0001) also shows that cold seep viruses display a high degree of endemism, similar to what was found previously in methane seep prokaryotic communities[59]. Viral Shannon diversity, Simpson diversity and Chao1 richness were all observed to be significantly different (p<0.05) between the seven sites (**Supplementary Table 6**). To arrange samples into environmentally meaningful groups, the seven cold seeps were designated as mineral-prone or mud-prone systems according to their fluid flow regime[2]. Low-flux, mineral-prone systems have longer geologic history with slower emission of fluids, e.g., gas hydrates (i.e. ENP, SMM and SB) and oil and gas seeps (i.e. EGM) whereas younger mud-prone systems are high-flux, such as mud volcanoes (i.e. HM and MS), asphalt volcanoes (i.e. WGM), brine pools and brine basins. Non-metric multidimensional scaling (NMDS) analysis revealed clear dissimilarity between viral communities in mineral-prone and mud-prone systems (ANOSIM, R=0.558, p<0.001; **Figure 3a**). For the most part, viral communities from mineral-prone systems clustered together, however SB_0 (surface sediment from 0.0 mbsf) deviated from other Scotian Basin viral communities as well as those from other mineral-prone seeps. Other factors thus also contribute to the structuring of the viral community, possibly including sediment depth (**Figure 3a**). Shannon diversity, Simpson diversity and Chao1 richness of viral communities were significantly higher in mineral-prone than in mud-prone seep systems (**Figure 3b**). Overall these results suggest that fluid flux is an important driver of viral community compositions in cold seep sediments.

To investigate the relationship between cold seep vOTUs and publicly available virus sequences from a broader diversity of ecosystems, a gene-sharing network was constructed using vConTACT2[41]. Such a weighted network can assign sequences into viral clusters (VCs) at approximately the genus level. Cold seep sediments, seawater, wetland and permafrost vOTUs were grouped into 3082 VCs (**Figure 4a** and **Supplementary Table 7**). Only 17 VCs were shared amongst all ecosystems (**Figure 4b**). The limited extent of clustering between viral genomes sampled from the various ecosystems may reflect a high degree of habitat specificity for viruses. Among cold seep sediment viruses, 1742 out of 2885 vOTUs were clustered into 804 VCs, with the majority (78.7%) not encountered in any other ecosystem. This suggests that most cold seep viruses may be endemic to cold seeps (**Figure 4b** and **Supplementary Table 8**). Among the 2885 cold seep vOTUs, only 162 clustered with wetland-derived vOTUs, 154 with seawater-derived vOTUs, and 95 with permafrost-derived vOTUs (**Supplementary Table 8**). Very few cold seep viral vOTUs (~0.7%) clustered with taxonomically known genomes from Viral RefSeq (**Figure 4a**), which is a much lower proportion compared to viruses recently discovered in soils using a similar approach[60]. Similarly, attempted taxonomic assignment of cold seep vOTUs using whole genome comparisons against 2616 known bacterial and archaeal viruses from NCBI RefSeq (version 94) left >96% unclassified. The remainder were assigned to the *Caudovirales* order, specifically *Podoviradae* (n=35), *Myoviradae* (n=34) and *Siphoviradae* (n=27) (**Figure 4c**). These analyses show that cold seep sediments harbour considerable unexplored viral diversity.

Viral lifestyles, virus-host linkages and host-linked viral abundance

Comparing sequence similarity, oligonucleotide frequencies, tRNA sequences and CRISPR-spacers[61], putative hosts were predicted for 14.2% of the 2885 cold seep vOTUs (**Supplementary Table 9**). Consistent with previous observations[61, 62], most of these vOTUs are predicted to have narrow host ranges, with only 54 vOTUs potentially exhibiting a broader host range across several phyla. 26 vOTUs were linked to both bacterial and archaeal hosts, suggesting existence of viral infection across domains (**Figure 2b**). To minimise the impact of potential false positives, 203 low-confidence host predictions were excluded from the analysis. For virus-host pairs with the greatest confidence, predicted prokaryotic hosts spanned 9 archaeal and 23 bacterial phyla, with the most frequent predictions being Thorarchaeota (19% of virus-host pairs) and Chloroflexi (14%) (**Figure 2a**). A considerable proportion (40%) of cold seep vOTUs were linked to archaea, including members of Bathyarchaeota, the Asgard group, *Methanomicrobia*, Thaumarchaeota and *Thermoplasmata*. Such broad ranges for archaeal viruses have not been reported previously in natural systems[12, 61]. Based on the presence of functional marker genes within MAGs, predicted hosts included two aerobic methanotrophic *Methylococcales* (**Supplementary Table 10**), 13 anaerobic methane-oxidizing archaea (e.g. ANME-1 and ANME-2, **Supplementary Figure 5a**), one non-methane multi-carbon alkane oxidizer within *Methanosarcinales* (**Supplementary Figure 5a**), 16 sulfate reducers mostly belonging to *Deltaproteobacteria* (**Supplementary Figure 5b**), and numerous respiring and fermentative heterotrophs (**Supplementary Table 10**). The genome of the sulfate reducer *Desulfobacterales* 8_GM_sbin_oily_21 also harboured genes possibly encoding akyl-/arylalkylsuccinate synthases related to anaerobic degradation of longer alkanes and aromatic hydrocarbons. These results suggest that viruses may influence the carbon and sulfur cycling via the lysis of populations mediating biogeochemical processes in cold seeps, where sulfate reduction is coupled to the anaerobic oxidation of methane and other seeping hydrocarbons. Predicted hosts were also identified within the candidate phyla radiation (six vOTUs are predicted to infect Patescibacteria) and DPANN archaea (13 vOTUs are predicted to infect Pacearchaeota or Aenigmarchaeota). Due to limited metabolic capabilities and small cell sizes, many CPR and DPANN organisms are likely to be obligate symbionts of other bacteria and archaea[63]. The impact of viral infection on obligate symbionts and any consequences for the larger organisms hosting those symbionts are not yet known, although it has been suggested that they may protect those hosts from viral predation[63].

Based on abundances determined by read mapping, targeted hosts were predicted for >20% of the cold seep viral community (**Figure 5a**). When grouped at the phylum level (class level for Proteobacteria and Euryarchaeota), the composition of predicted microbial hosts agreed well with that of their viruses (**Figure 5b**). This is supported by regression modelling of the abundances of hosts and lineage-specific viruses (**Figure 5c**). By applying metagenomic read recruitment, most viruses have higher genome coverage compared to their hosts, suggesting that most taxa may be undergoing active viral replication and possibly lysis at the time of sample collection[64]. Lineage-specific virus/host abundance ratios (i.e. VHR) for most taxa were greater than one with Thorarchaeota being the highest at $10^{2.5}$ (**Figure 5d**), indicating a high level of active viral genome replication. This is in accordance with the presence of higher abundances of viral particles detected by epifluorescence microscopy in cold seep sediments compared to non-cold seep sediments in the Gulf of Mexico[9]. Thus in cold seep sediments, viral lysis may be a major top-down factor[65], contributing to significant microbial mortality. In addition, based on their contigs containing integrase genes and/or being located within their host genomes, at least 372 cold seep vOTUs were predicted to be lysogenic (i.e. temperate viruses, **Supplementary Figure 4** and **Supplementary Table 4**).

Viral AMGs involved in carbon, sulfur and nitrogen transformations

To further understand how viruses might affect the biogeochemistry of cold seep sediments, viral contigs encoding AMGs that supplement host metabolism during infection were examined. Overall, cold seep viruses tended to encode AMGs for cofactor/vitamin and carbohydrate metabolism. A significant portion also encoded AMGs for amino acid and glycan metabolism (**Figure 6a**). We identified 70 genes encoding carbohydrate-active enzymes (CAZymes), related to the initial breakdown of complex carbohydrates, with 22 of them affiliated to glycoside hydrolyases (**Figure 6b**). These 22 genes, spanning 16 glycoside hydrolase families (**Supplementary Table 11**), were predicted to function in polymer hydrolysis, typical of bacteria and/or archaea (e.g. Planctomycetes and Thorarchaeota)[66]. Two *mmoB* genes encoding soluble methane monooxygenase regulatory protein B were identified in viral contigs, which might be associated with aerobic methane oxidation[67, 68]. No other AMGs directly related to key functional genes involved in initial activation of hydrocarbons were identified. However, many genes potentially involved in downstream hydrocarbon biodegradation pathways were identified, e.g., acetate-CoA ligase (*acd*), acetyl-CoA synthetase (*acs*), acetyl-CoA decarbonylase/synthase (*cdhD* and *cdhE*), 5,6,7,8-tetrahydromethanopterin hydro-lyase (*fae*), anaerobic carbon-monoxide dehydrogenase (*cooS*), 5,10-methylenetetrahydromethanopterin reductase (*mer*), and heterodisulfide reductase subunit C2 (*hdrC2*) (**Supplementary Table 12**). These genes might aid in bacterial fermentative or respiratory consumption of metabolites produced from oxidation of hydrocarbons and other complex substrates.

The most common AMG related to sulfur metabolism within the viral contigs was phosphoadenosine phosphosulfate reductase (*cysH*), predicted to participate in assimilatory sulfate reduction (**Supplementary Table 12**). Viral *cysH* has also been found in viral sequences obtained from the rumen[69], a deep freshwater lake[15] and sulfidic mine tailings[70]. Other related enzymes in the assimilatory sulfate reduction pathway including adenylylsulfate kinase (*cysC*) and cysteine synthase (*cysK*) were also identified but only in relatively small number of viral sequences (**Figure 6c**). These genes likely facilitate host utilization of reduced sulfur compounds during infection, providing viruses with some fitness advantage. AMGs related to sulfate assimilation were less prevalent in mud-prone systems (**Figure 6c**), possibly due to low sulfate availability in mud-prone systems, as a result of low sulfate intrusion into sediments caused by rapid rates of upward fluid flow from the subsurface[2]. Two *dsrC* genes were identified in viral contigs, and may be involved in dissimilatory sulfur metabolism[16]. One viral contig encoded a sulfur dioxygenase (*sdo*) for facilitating sulfur oxidation, with the predicted host being a *Deltaproteobacteria* (**Supplementary Tables 9 and 12**). Numerous contigs contained *nosD* (encoding a nitrous oxidase accessory protein) and two contigs contained *nrfA,* (encoding cytochrome c nitrite reductase; **Supplementary Table 12**), suggesting that viruses might also manipulate nitrogen cycling in cold seep sediments[71].

Conclusions

Due to the challenges of deep sea sediment sampling and laboratory cultivation of microbial communities along with their viruses, the roles that viruses play in influencing microbial mortality, ecology and evolution remains largely unexplored in marine sediments associated with cold seeps[21, 72]. In this study, in-depth exploration of untargeted *de novo* metagenomic data successfully revealed novel, abundant and diverse bacterial and archaeal viruses. Many of the putative microbial hosts for seep viruses belong to taxonomic groups with no cultured representatives. These results therefore expand the diversity of archaeal viruses, especially those infecting important archaeal lineages in hydrocarbon seep microbiomes, e.g., members of the Euryarchaeota, Bathyarchaeota, and the Asgard group. While a significant portion of the viruses appear to be lysogenic, the high read coverages for many viral genomes suggest that viral lysis is a major source of microbial mortality and biomass turnover in cold seep sediments. Virus encoded AMGs, including genes related to carbon, sulfur, and nitrogen metabolism, may augment the metabolism of prokaryotic hosts during infection, potentially altering biogeochemical processes mediated by cold seep microorganisms. As subsurface reservoirs of prokaryotic diversity and hotspots of microbial activity, cold seeps additionally represent oases of viruses and viral activity. Much remains to be revealed about the contribution of viruses to the functioning of cold seeps and other marine environments, especially with respect to their potential role in horizontal gene transfer which was not addressed in this study. With only a fraction of vOTUs identified here able to be classified, and many of them predicted to infect poorly characterized taxa, there remain large gaps in understanding the microbiology of these environments.

References

1. Suess, E. Marine cold seeps and their manifestations: geological control, biogeochemical criteria and environmental conditions. *Int J Earth Sci* **103**, 1889-1916 (2014).

2. Joye, S.B. The Geology and Biogeochemistry of Hydrocarbon Seeps. *Annu Rev Earth Planet Sci* **48**, 205-231 (2020).

3. Etiope, G. et al. A thermogenic hydrocarbon seep in shallow Adriatic Sea (Italy): Gas origin, sediment contamination and benthic foraminifera. *Mar Petrol Geol* **57**, 283-293 (2014).

4. Kennicutt, M.C. in Habitats and Biota of the Gulf of Mexico: Before the Deepwater Horizon Oil Spill. (ed. C.H. Ward) 275-358 (Springer New York, New York, NY; 2017).

5. Ruppel, C.D. & Kessler, J.D. The interaction of climate change and methane hydrates. *Rev Geophys* **55**, 126-168 (2017).

6. Kniemeyer, O. et al. Anaerobic oxidation of short-chain hydrocarbons by marine sulphate-reducing bacteria. *Nature* **449**, 898-901 (2007).

7. Jaekel, U. et al. Anaerobic degradation of propane and butane by sulfate-reducing bacteria enriched from marine hydrocarbon cold seeps. *ISME J* **7**, 885-895 (2013).

8. Teske, A. & Carvalho, V. Marine hydrocarbon seeps: microbiology and biogeochemistry of a global marine habitat. (Springer Nature, 2020).

9. Kellogg, C.A. Enumeration of viruses and prokaryotes in deep-sea sediments and cold seeps of the Gulf of Mexico. *Deep-Sea Res Pt II* **57**, 2002-2007 (2010).

10. Bryson, S.J., Thurber, A.R., Correa, A.M., Orphan, V.J. & Vega Thurber, R. A novel sister clade to the enterobacteria microviruses (family Microviridae) identified in methane seep sediments. *Environ Microbiol* **17**, 3708-3721 (2015).

11. Pan, D., Morono, Y., Inagaki, F. & Takai, K. An Improved Method for Extracting Viruses From Sediment: Detection of Far More Viruses in the Subseafloor Than Previously Reported. *Front Microbiol* **10**, 878 (2019).

12. Emerson, J.B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* **3**, 870-880 (2018).

13. Jin, M. et al. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* **7**, 58 (2019).

14. Labbe, M., Girard, C., Vincent, W.F. & Culley, A.I. Extreme Viral Partitioning in a Marine-Derived High Arctic Lake. *mSphere* **5**, e00334-00320 (2020).

15. Okazaki, Y., Nishimura, Y., Yoshida, T., Ogata, H. & Nakano, S.I. Genome-resolved viral and cellular metagenomes revealed potential key virus-host interactions in a deep freshwater lake. *Environ Microbiol* **21**, 4740-4754 (2019).

16. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689-693 (2016).

17. Gregory, A.C. et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114 (2019).

18. Coutinho, F.H. et al. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun* **8**, 15955 (2017).

19. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N.A. Phage puppet masters of the marine microbial realm. *Nat Microbiol* **3**, 754-766 (2018).

20. Chen, L.X. et al. Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat Microbiol*, doi: 10.1038/s41564-41020-40779-41569 (2020).

21. Backstrom, D. et al. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. *mBio* **10**, e02497-02418 (2019).

22. Cai, L. et al. Active and diverse viruses persist in the deep sub-seafloor sediments over thousands of years. *ISME J* **13**, 1857-1864 (2019).

23. Dong, X. et al. Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nat Commun* **10**, 1816 (2019).

24. Dong, X. et al. Thermogenic hydrocarbons sustain diverse subseafloor microbial communities in deep sea cold seep sediments. *bioRxiv*, doi: 10.1101/2020.1102.1102.928283 (2020).

25. Gruber-Vodicka, H.R., Seah, B.K.B. & Pruesse, E. phyloFlash – rapid SSU rRNA profiling and targeted assembly from metagenomes. *bioRxiv*, doi: 10.1101/521922 (2019).

26. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-D596 (2013).

27. Uritskiy, G.V., DiRuggiero, J. & Taylor, J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

28. Li, D. et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3-11 (2016).

29. Olm, M.R., Brown, C.T., Brooks, B. & Banfield, J.F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864-2868 (2017).

30. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P. & Parks, D.H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, doi: 10.1093/bioinformatics/btz1848 (2019).

31. Parks, D.H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol*, doi: 10.1038/s41587-41020-40501-41588 (2020).

32. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

33. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res* **40**, D136-D143 (2012).

34. Roux, S., Enault, F., Hurwitz, B.L. & Sullivan, M.B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

35. 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).

36. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152 (2012).

37. Nayfach, S., Camargo, A.P., Eloe-Fadrosh, E., Roux, S. & Kyrpides, N. CheckV: assessing the quality of metagenome-assembled viral genomes. *bioRxiv*, doi: 10.1101/2020.1105.1106.081778 (2020).

38. 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).

39. Dalcin Martins, P. et al. Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. *Microbiome* **6**, 138 (2018).

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

41. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, 632-639 (2019).

42. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).

43. Castelan-Sanchez, H.G. et al. Extremophile deep-sea viral communities from hydrothermal vents: Structural and functional analysis. *Mar Genomics* **46**, 16-28 (2019).

44. Huson, D.H., Auch, A.F., Qi, J. & Schuster, S.C. MEGAN analysis of metagenomic data. *Genome Res* **17**, 377-386 (2007).

45. Tominaga, K., Morimoto, D., Nishimura, Y., Ogata, H. & Yoshida, T. In silico Prediction of Virus-Host Interactions for Marine Bacteroidetes With the Use of Metagenome-Assembled Genomes. *Front Microbiol* **11**, 738 (2020).

46. Ahlgren, N.A., Ren, J., Lu, Y.Y., Fuhrman, J.A. & Sun, F. Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res* **45**, 39-53 (2017).

47. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**, 11-16 (2004).

48. Skennerton, C.T., Imelfort, M. & Tyson, G.W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res* **41**, e105 (2013).

49. Dong, X. & Strous, M. An Integrated Pipeline for Annotation and Visualization of Metagenomic Contigs. *Front Genet* **10**, 999 (2019).

50. Zhou, Z., Tran, P., Liu, Y., Kieft, K. & Anantharaman, K. METABOLIC: a scalable high-throughput metabolic and biogeochemical functional trait profiler based on microbial genomes. *bioRxiv*, doi: 10.1101/761643 (2019).

51. Edgar, R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

52. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549 (2018).

53. Zhang, H. et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **46**, W95-W101 (2018).

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

55. Bowers, R.M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725-731 (2017).

56. Jain, C., Rodriguez, R.L., Phillippy, A.M., Konstantinidis, K.T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).

57. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat Biotechnol* **37**, 29-37 (2019).

58. Al-Shayeb, B. et al. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425-431 (2020).

59. Ruff, S.E. et al. Global dispersion and local diversification of the methane seep microbiome. *Proc Natl Acad Sci U S A* **112**, 4015-4020 (2015).

60. Trubl, G. et al. Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems* **3**, e00076-00018 (2018).

61. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425-430 (2016).

62. Roux, S., Hallam, S.J., Woyke, T. & Sullivan, M.B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* **4**, e08490 (2015).

63. Castelle, C.J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* **16**, 629-645 (2018).

64. Jarett, J.K. et al. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J*, doi: 10.1038/s41396-41020-40705-41394 (2020).

65. Orsi, W.D. Ecology and evolution of seafloor and subseafloor microbial communities. *Nat Rev Microbiol* **16**, 671-683 (2018).

66. Garron, M.L. & Henrissat, B. The continuing expansion of CAZymes and their families. *Curr Opin Chem Biol* **53**, 82-87 (2019).

67. Kim, H. et al. MMOD-induced structural changes of hydroxylase in soluble methane monooxygenase. *Sci Adv* **5**, eaax0059 (2019).

68. Walters, K.J., Gassner, G.T., Lippard, S.J. & Wagner, G. Structure of the soluble methane monooxygenase regulatory protein B. *Proc Natl Acad Sci U S A* **96**, 7877-7882 (1999).

69. Anderson, C.L., Sullivan, M.B. & Fernando, S.C. Dietary energy drives the dynamic response of bovine rumen viral communities. *Microbiome* **5**, 155 (2017).

70. Gao, S.M. et al. Depth-related variability in viral communities in highly stratified sulfidic mine tailings. *Microbiome* **8**, 89 (2020).

71. Kuypers, M.M.M., Marchant, H.K. & Kartal, B. The microbial nitrogen-cycling network. *Nat Rev Microbiol* **16**, 263-276 (2018).

72. Zheng, X. et al. Extraordinary diversity of viruses in deep-sea sediments as revealed by metagenomics without prior virion separation. *Environ Microbiol*, doi: 10.1111/1462-2920.15154 (2020).

Data availability

Sequences of 2885 viral contigs and 592 de-replicated metagenome-assembled genomes can be found at figshare (DOI: 10.6084/m9.figshare.12922229). All other data are available from the corresponding author upon request.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (No. 41906076), the Fundamental Research Funds for the Central Universities (No. 19lgpy90), and Guangzhou Marine Geological Survey (No. 2019C-15-229). We thank Benjamin Bolduc for help with vContact2 software, Simon Roux for help with host assignment, and Chuwen Zhang for helpful comments.

Author contributions

X.D. designed this study. X.D., Z.L., and W.P. analyzed metagenomic data. X.D., Z.L., D.P., and G.W. interpreted data. Z.L., Y.P., and L.Z. performed viral diversity analyses. C.R.J. H. contributed part of the data. X.D., Z.L., D.P., and C.R.J. H. wrote the paper, with input from other authors.

Competing interest

The authors declare no conflict of interest.

Figure Legends

Figure 1 Geographic distribution of sampling sites where metagenomic data were collected. Locations of cold seep sites indicating the site name and abbreviation, sampling depth range in meters below seafloor (mbsf) and water depth in meters below sea level (m).

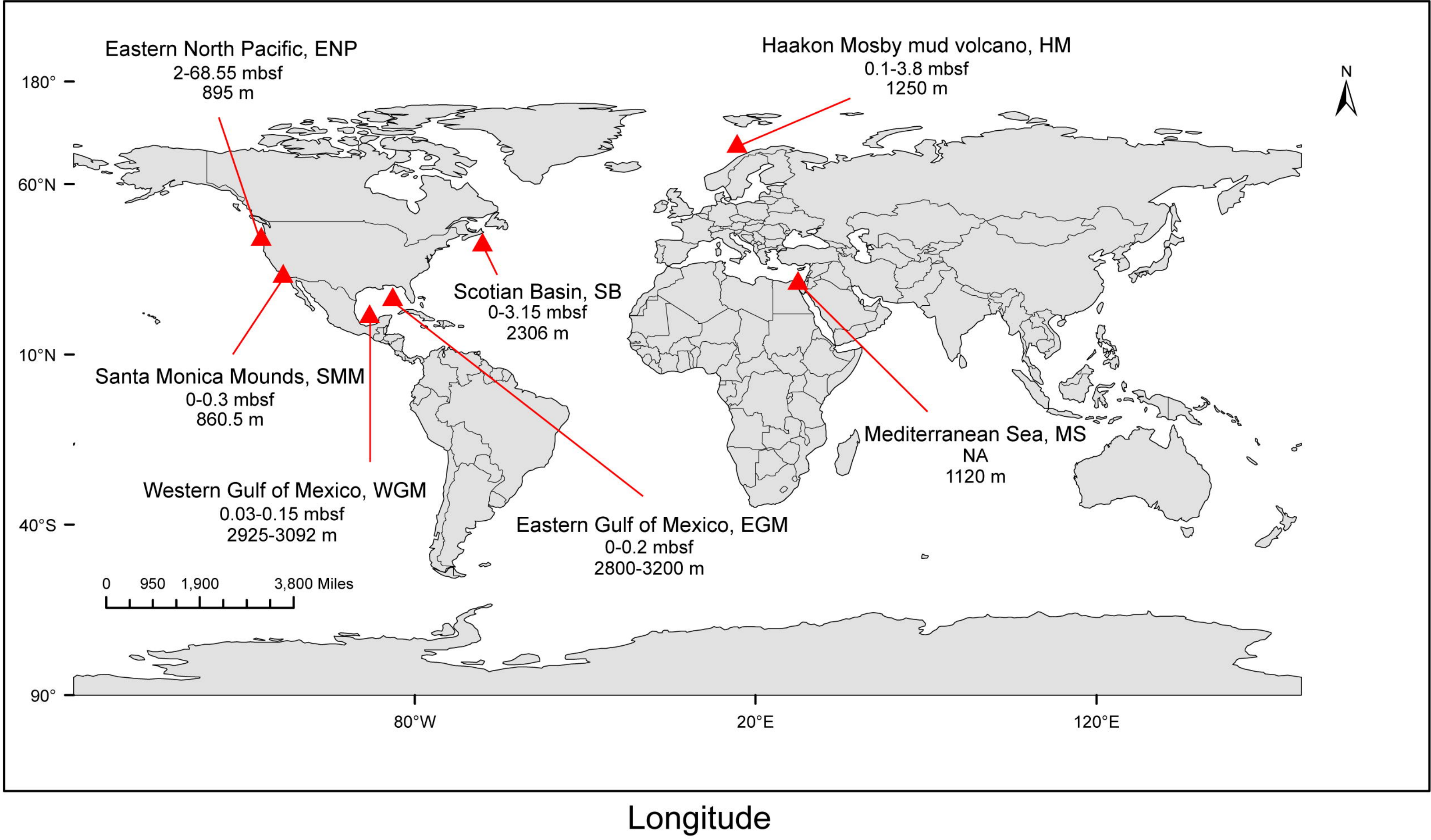
Figure 2 Cold seep virus-host linkages. (a) Maximum-likelihood phylogenetic tree of bacterial and archaeal MAGs at the phylum level (class level for Proteobacteria and Euryarchaeota), inferred from a concatenated alignment of 120 bacterial or 122 archaeal single-copy marker genes. Clades outlined by solid lines represent lineages predicted to include a host for one or more viral OTUs (number of vOTUs predicted to have a host within a clade is shown in grey circles). (b) Network of putative virus-host linkages. Edges indicate putative virus-host pairs. Large nodes represent bacterial (blue) or archaeal (pink) hosts. Small nodes represent vOTUs coloured according to host ranges: grey, host-specific infection at or below phylum level; brown, cross-phylum infection; black, cross-domain infection.

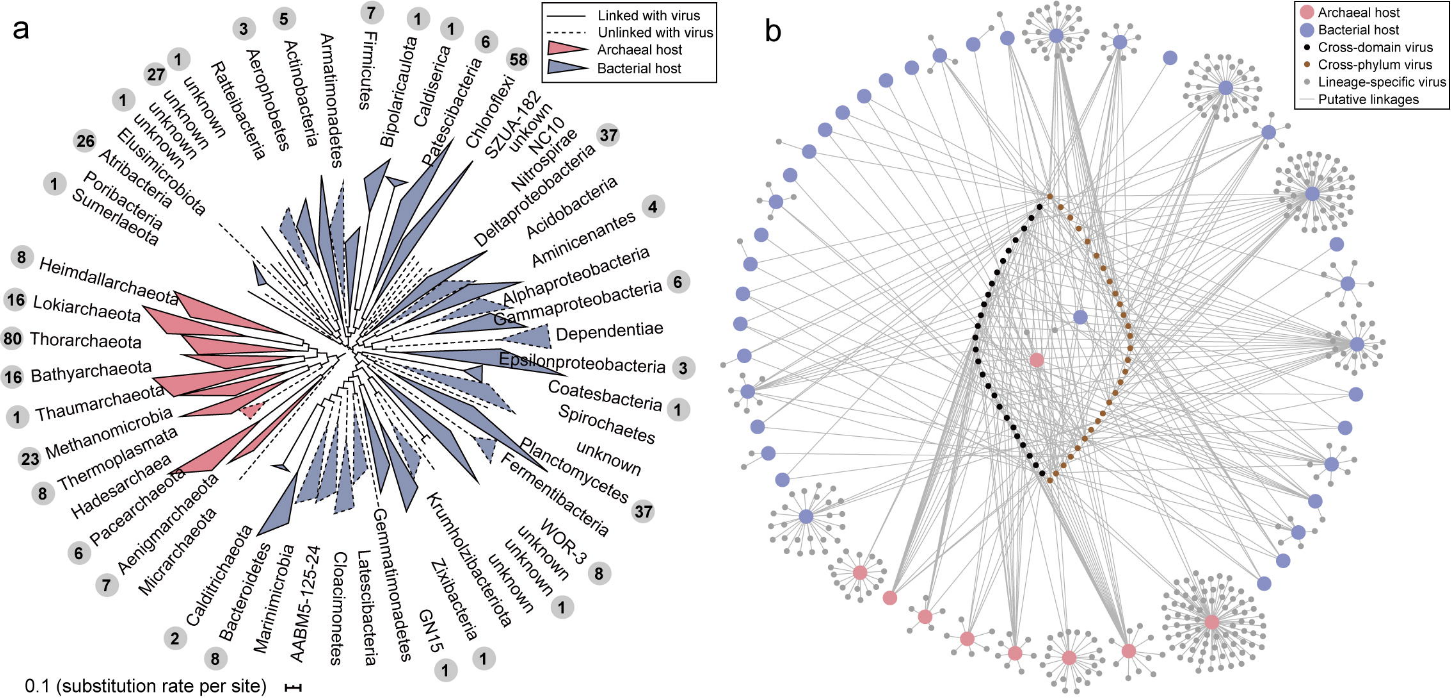
Figure 3 Comparison of viral community diversity between mineral-prone and mud-prone cold seeps. (a) NMDS analysis of a Bray-Curtis dissimilarity matrix calculated from RPKM values of vOTUs. ANOSIM was applied to test for the difference between viral communities in mineral-prone (n=19) or mud-prone (n=9) systems. (b) Shannon, Simpson and Chao1 indices of the viral community diversity from mineral-prone and mud-prone cold seeps. Asterisks denote significance, with * indicating *p*<0.05, and ** indicating *p*<0.01.

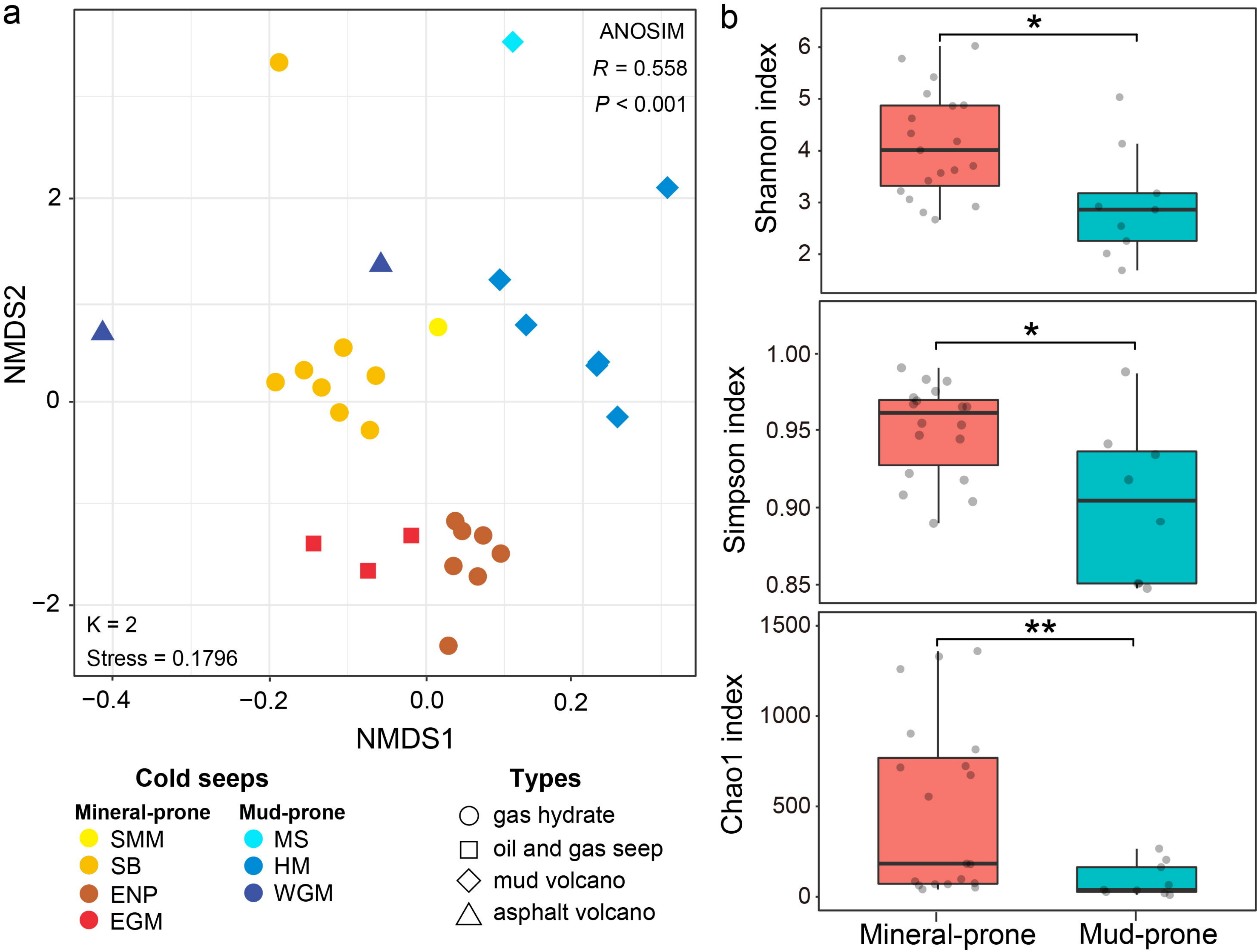
Figure 4 Taxonomic diversity of cold seep viruses. (a) Gene-sharing network of viral sequence space based on assembled viral genomes from cold seep sediment, wetland, permafrost, seawater and RefSeq prokaryotic viral genomes. Nodes represent viral genomes and edges indicate similarity based on shared protein clusters. (b) Venn diagram of shared viral clusters among the four environmental virus data sets and RefSeq. (c) Taxonomic assignments of vOTUs.

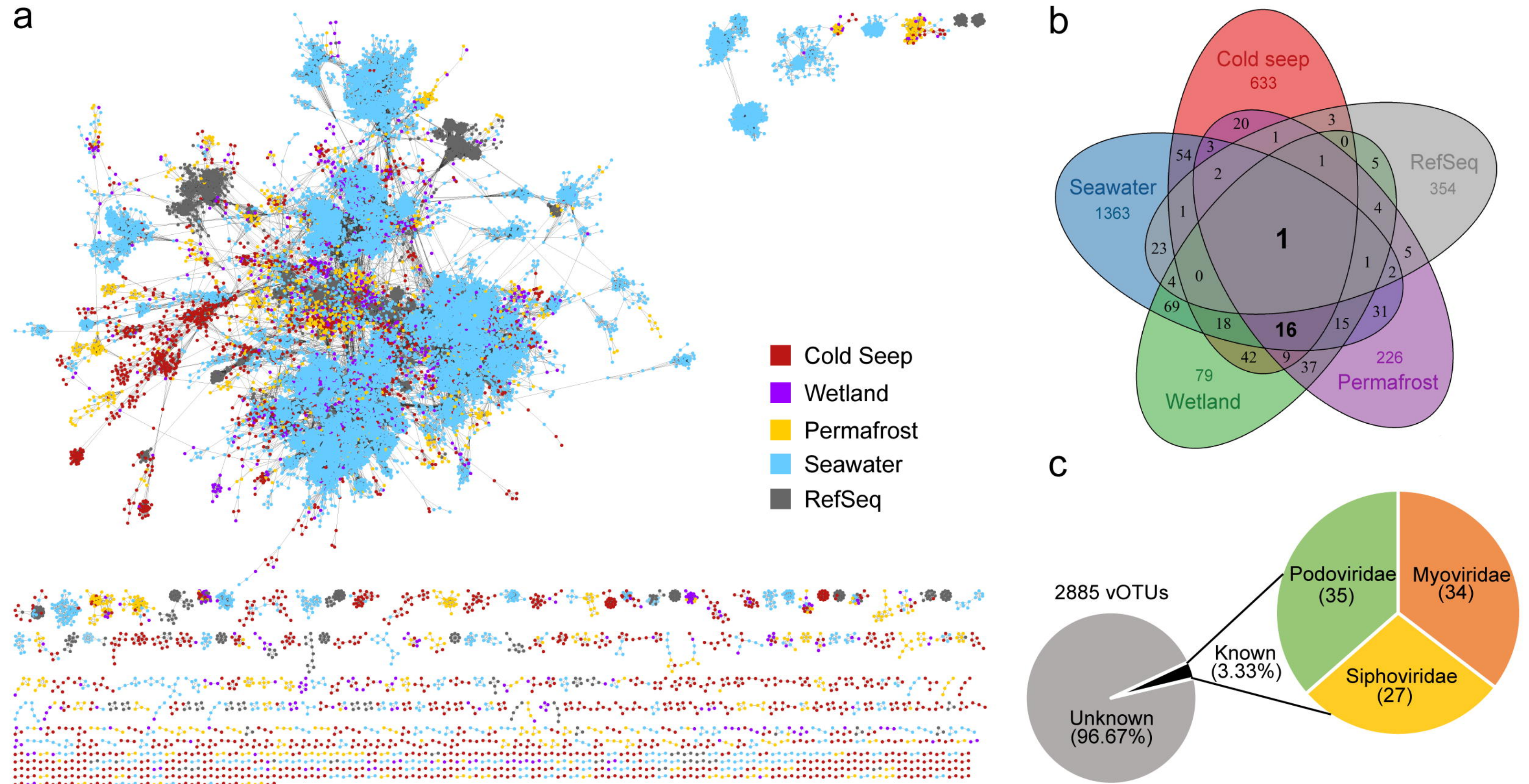
Figure 5 Relative abundance patterns of viruses and their predicted hosts in cold seep sediments. (a) Percentage of vOTUs based on relative abundance in which a host was predicted or not. (b) Relative abundances of vOTUs and their predicted hosts grouped by the host taxonomy (c) Significant Pearson correlation between relative abundances of viruses and their hosts (calculated by normalized mean coverage depth, reads per kilobase mapped reads: RPKM). (d) Lineage-specific virus-host abundance ratios (VHR) of for all predicted microbial hosts. The red line indicates a 1:1 ratio. Predicted hosts in (b) and (c) are in different colours as shown in the colour bar on the right side.

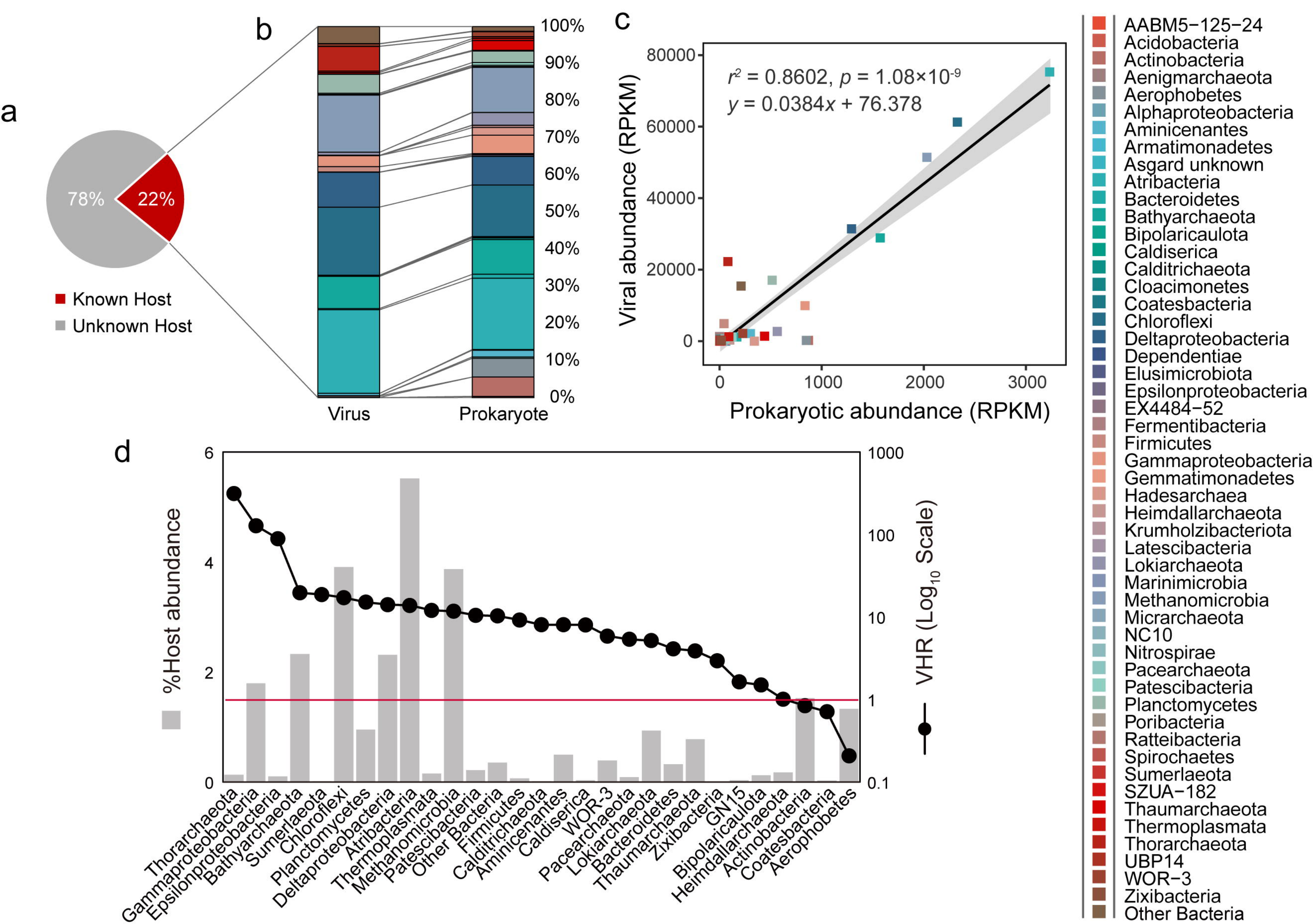
Figure 6 Profiles of virus-encoded auxiliary metabolic genes (AMGs). (a) Classification of AMGs into KEGG metabolic categories. (b) Classification of viral ORFs encoding carbohydrate-active enzymes. (c) Relative abundance of AMGs associated with sulfur metabolism. Gene abbreviations: sulfur dioxygenase (*sdo*), dissimilatory sulfite reductase related protein (*dsrC*), cysteine synthase (*cysK*), phosphoadenosine phosphosulfate reductase (*cysH*), adenylylsulfate kinase (*cysC*).

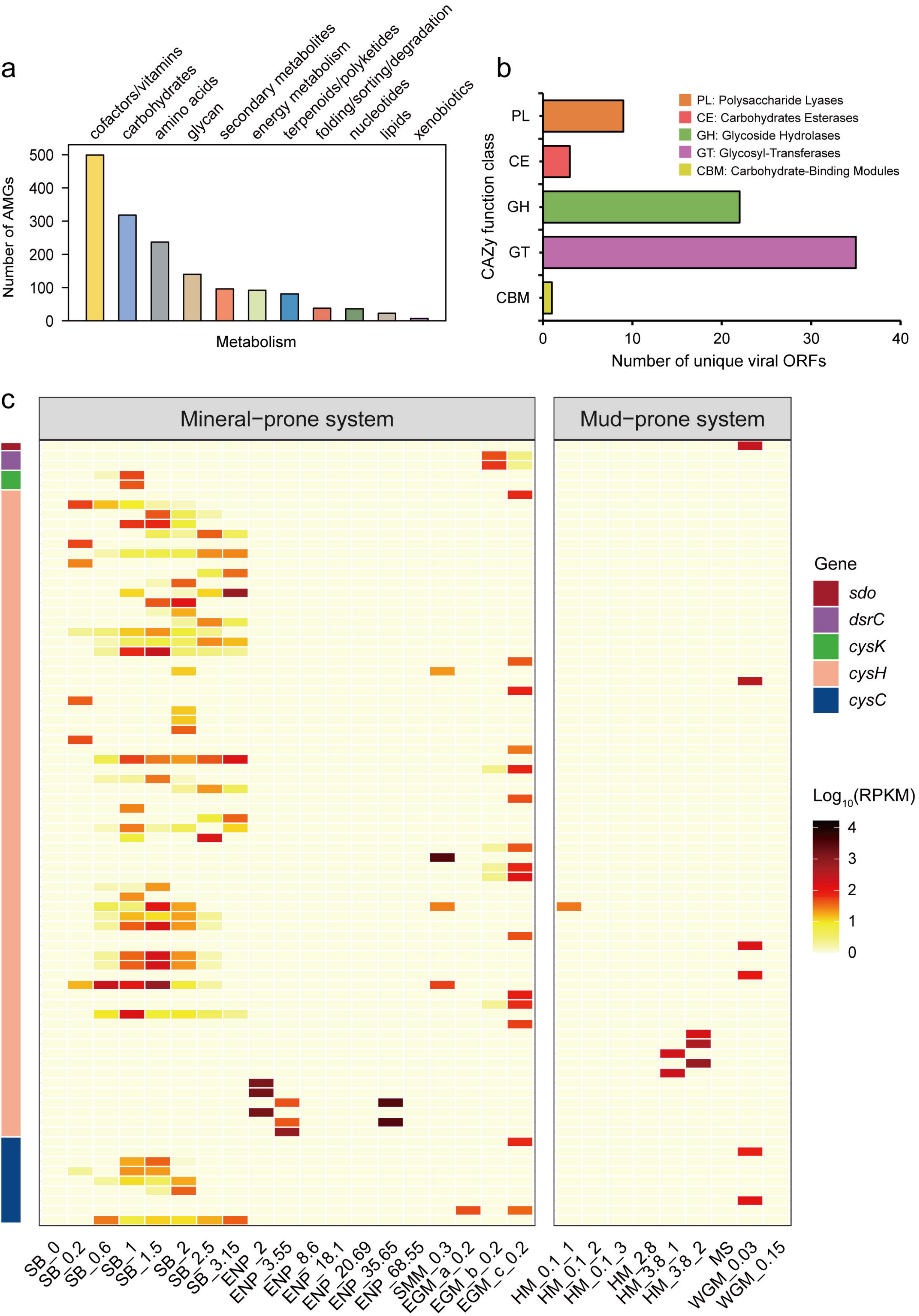

Eastern North Pacific, ENP
2-68.55 mbsf
895 m
Haakon Mosby mud volcano, HM
0.1-3.8 mbsf
1250 m
N
Scotian Basin, SB
0-3.15 mbsf
2306 m
Santa Monica Mounds, SMM
0-0.3 mbsf
860.5 m
Mediterranean Sea, MS
NA
1120 m
Western Gulf of Mexico, WGM
0.03-0.15 mbsf
2925-3092 m
Eastern Gulf of Mexico, EGM
0-0.2 mbsf
2800-3200 m
0
950
1,900
3,800 Miles
180°
60°N
10°N
40°S
90°
80°W
20°E
120°E
Latitude
Longitude


a
Linked with virus
Unlinked with virus
Archaeal host
Bacterial host
Heimdallarchaeota 8
Lokiarchaeota 16
Thorarchaeota 80
Bathyarchaeota 16
Thaumarchaeota 1
Methanomicrobia 23
Thermoplasmata 8
Hadesarchaea
Pacearchaeota 6
Aenigmarchaeota 7
Micrarchaeota
Calditrichaeota 2
Bacteroidetes 8
Marinimicrobia
AABM5-125-24
Cloacimonetes
Latescibacteria
Gemmatimonadetes
GN15 1
Zixibacteria 1
Krumholzibacteriota
unknown 1
unknown
unknown
WOR-3 8
Fermentibacteria
Planctomycetes 37
unknown
Spirochaetes
Coatesbacteria 1
Epsilonproteobacteria 3
Dependentiae
Gammaproteobacteria 6
Alphaproteobacteria
Aminicenantes 4
Acidobacteria
Deltaproteobacteria 37
Nitrospirae
NC10
unkown
SZUA-182
Chloroflexi 58
Patescibacteria 6
Caldiserica 1
Bipolaricaulota 1
Firmicutes 7
Armatimonadetes
Actinobacteria 5
Aerophobetes 3
Ratteibacteria
unknown 1
unknown 27
unknown 1
Elusimicrobiota
Atribacteria 26
Poribacteria 1
Sumerlaeota
0.1 (substitution rate per site)
b
Archaeal host
Bacterial host
Cross-domain virus
Cross-phylum virus
Lineage-specific virus
Putative linkages


a
b
ANOSIM
R = 0.558
P < 0.001
K = 2
Stress = 0.1796
NMDS1
NMDS2
−0.4
−0.2
0.0
0.2
2
0
−2
Cold seeps
Mineral-prone
SMM
SB
ENP
EGM
Mud-prone
MS
HM
WGM
Types
gas hydrate
oil and gas seep
mud volcano
asphalt volcano
Shannon index
Simpson index
Chao1 index
*
*
**
Mineral-prone
Mud-prone
6
5
4
3
2
1.00
0.95
0.90
0.85
1500
1000
500
0


a
Cold Seep
Wetland
Permafrost
Seawater
RefSeq
b
Cold seep
633
RefSeq
354
Seawater
1363
Wetland
79
Permafrost
226
20
3
1
0
3
54
1
5
2
1
4
1
23
1
5
4
0
1
2
69
31
18
16
15
42
9
37
c
2885 vOTUs
Unknown
(96.67%)
Known
(3.33%)
Podoviridae
(35)
Myoviridae
(34)
Siphoviridae
(27)


a
78%
22%
Known Host
Unknown Host
b
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Virus
Prokaryote
c
$r^2 = 0.8602$, $p = 1.08×10^{-9}$
$y = 0.0384x + 76.378$
Viral abundance (RPKM)
80000
60000
40000
20000
0
0
1000
2000
3000
Prokaryotic abundance (RPKM)
d
%Host abundance
6
4
2
0
VHR (Log$_{10}$ Scale)
1000
100
10
1
0.1
Thorarchaeota
Gammaproteobacteria
Epsilonproteobacteria
Bathyarchaeota
Sumerlaeota
Chloroflexi
Planctomycetes
Deltaproteobacteria
Atribacteria
Thermoplasmata
Methanomicrobia
Patescibacteria
Other Bacteria
Firmicutes
Calditrichaeota
Aminicenantes
Caldiserica
WOR-3
Pacearchaeota
Lokiarchaeota
Bacteroidetes
Thaumarchaeota
Zixibacteria
GN15
Bipolaricaulota
Heimdallarchaeota
Actinobacteria
Coatesbacteria
Aerophobetes
AABM5−125−24
Acidobacteria
Actinobacteria
Aenigmarchaeota
Aerophobetes
Alphaproteobacteria
Aminicenantes
Armatimonadetes
Asgard unknown
Atribacteria
Bacteroidetes
Bathyarchaeota
Bipolaricaulota
Caldiserica
Calditrichaeota
Cloacimonetes
Coatesbacteria
Chloroflexi
Deltaproteobacteria
Dependentiae
Elusimicrobiota
Epsilonproteobacteria
EX4484−52
Fermentibacteria
Firmicutes
Gammaproteobacteria
Gemmatimonadetes
Hadesarchaea
Heimdallarchaeota
Krumholzibacteriota
Latescibacteria
Lokiarchaeota
Marinimicrobia
Methanomicrobia
Micrarchaeota
NC10
Nitrospirae
Pacearchaeota
Patescibacteria
Planctomycetes
Poribacteria
Ratteibacteria
Spirochaetes
Sumerlaeota
SZUA−182
Thaumarchaeota
Thermoplasmata
Thorarchaeota
UBP14
WOR−3
Zixibacteria
Other Bacteria


a
cofactors/vitamins
carbohydrates
amino acids
glycan
secondary metabolites
energy metabolism
terpenoids/polyketides
folding/sorting/degradation
nucleotides
lipids
xenobiotics
Number of AMGs
500
400
300
200
100
0
Metabolism
b
CAZy function class
PL
CE
GH
GT
CBM
🟧 PL: Polysaccharide Lyases
🟥 CE: Carbohydrates Esterases
🟩 GH: Glycoside Hydrolases
🟪 GT: Glycosyl-Transferases
🟨 CBM: Carbohydrate-Binding Modules
0
10
20
30
40
Number of unique viral ORFs
c
Mineral−prone system
Mud−prone system
SB_0
SB_0.2
SB_0.6
SB_1
SB_1.5
SB_2
SB_2.5
SB_3.15
ENP_2
ENP_3.55
ENP_8.6
ENP_18.1
ENP_20.69
ENP_35.65
ENP_68.55
SMM_0.3
EGM_a_0.2
EGM_b_0.2
EGM_c_0.2
HM_0.1_1
HM_0.1_2
HM_0.1_3
HM_2.8
HM_3.8_1
HM_3.8_2
MS
WGM_0.03
WGM_0.15
Gene
🟥 *sdo*
🟪 *dsrC*
🟩 *cysK*
🟧 *cysH*
🟦 *cysC*
Log$_{10}$(RPKM)
4
3
2
1
0