

Global diversity and ecological functions of viruses inhabiting oil reservoirs

Received: 27 October 2023

Accepted: 30 July 2024

Published online: 08 August 2024

Liyun An¹, Xinwu Liu², Jianwei Wang², Jinbo Xu³, Xiaoli Chen^{2,4}, Xiaonan Liu², Bingxin Hu², Yong Nie²✉ & Xiao-Lei Wu^{1,2,3,4,5}✉

Oil reservoirs, being one of the significant subsurface repositories of energy and carbon, host diverse microbial communities affecting energy production and carbon emissions. Viruses play crucial roles in the ecology of microbiomes, however, their distribution and ecological significance in oil reservoirs remain undetermined. Here, we assemble a catalogue encompassing viral and prokaryotic genomes sourced from oil reservoirs. The catalogue comprises 7229 prokaryotic genomes and 3,886 viral Operational Taxonomic Units (vOTUs) from 182 oil reservoir metagenomes. The results show that viruses are widely distributed in oil reservoirs, and 85% vOTUs in oil reservoir are detected in less than 10% of the samples, highlighting the heterogeneous nature of viral communities within oil reservoirs. Through combined microcosm enrichment experiments and bioinformatics analysis, we validate the ecological roles of viruses in regulating the community structure of sulfate reducing microorganisms, primarily through a virulent lifestyle. Taken together, this study uncovers a rich diversity of viruses and their ecological functions within oil reservoirs, offering a comprehensive understanding of the role of viral communities in the biogeochemical cycles of the deep biosphere.

The global dependence on fossil-fuel energy and the escalating levels of atmospheric carbon dioxide present a pressing challenge to humanity¹. In the realm of subsurface energy and carbon reservoirs, oil reservoirs have long harbored diverse microorganisms, and have been a subject of growing interest since the 1930s². Recent research has intensified the focus on understanding the presence and role of microorganisms in these deep subsurface environments^{3,4}, driven by their dual impact on energy production and carbon emissions^{5,6}.

Despite extreme environmental conditions, oil reservoirs still offer a variety of niches to support various microorganisms⁷, including hydrocarbon oxidizing microorganisms, fermentative microorganisms, sulfate reducing microorganisms (SRMs), nitrate reducing microorganisms, and methanogens^{8,9}. Microorganisms in oil reservoirs have adapted to these extreme conditions and contribute to important ecological processes, such as the transformation of crude oil via biodegradation and other geochemical cycles^{5,10}. Furthermore, microbial

H₂S production by SRMs is a prevalent metabolic process within subsurface oil reservoirs¹¹, and this can lead to the corrosion of metal equipment and infrastructure, souring of oil, and health risks, resulting in significant economic and ecological costs¹¹.

While the importance of microorganisms within oil reservoirs has become increasingly recognized, the mechanisms governing the assembly and regulation of microbial communities in these deep subsurface environments remains unclear. Microbial communities in oil reservoirs undergo fluctuations during production-related engineering interventions, particularly when surface activities establish connections with subsurface reservoirs¹². Additionally, while variation partitioning of beta-diversity in oil reservoirs revealed that chemical properties and physical conditions explained a substantial fraction of variation in microbial beta-diversity, more than 70% of the community variation could not be explained¹³. Viruses are directly associated with host physiology and mortality, and further influence microbial

¹College of architecture and environment, Sichuan University, Chengdu 610065, China. ²College of Engineering, Peking University, Beijing 100871, China.

³School of Earth and Space Sciences, Peking University, Beijing 100871, China. ⁴Institute of Ocean Research, Peking University, Beijing 100871, China.

⁵Institute of Ecology, Peking University, Beijing 100871, China. ✉e-mail: nieyong@pku.edu.cn; xiaolei_wu@pku.edu.cn

community dynamics in aquatic environments¹⁴. Therefore, viruses in the unique environment of oil reservoirs may be one of the major contributors to the dynamic changes observed in the microbial communities within these reservoirs.

Viruses are the most abundant biological entities on the planet and are found in nearly all environments¹⁵, even in extreme environments such as Antarctic soils¹⁶, thawing permafrost¹⁷, and cryoconite holes of glaciers¹⁸. They play pivotal roles in natural ecosystems by interacting with microbial hosts and thereby exerting significant influence over global biogeochemical cycles, such as the global cycling of nutrients, energy flow, and food web dynamics¹⁹. Viruses can function as predators, regulating microbial abundance while releasing organic matter and inorganic nutrients through cell lysis²⁰. Furthermore, viruses can actively regulate and rebuild host metabolism via auxiliary metabolic genes (AMGs) in their genomes that are expressed during infection. Recent studies have uncovered the presence of a substantial number of viruses within oil reservoir environments^{21,22}. A meta-analysis of eight production wells indicated that the abundance of viruses is approximately $3 \times 10^8 \text{ mL}^{-1}$. The viral abundance in production wells during water flooding is higher than that in production wells during microbial flooding²¹. In addition, a previous study recovered about 800 reference viral genomes (length $\geq 10 \text{ kb}$), revealing viruses are widespread in hydraulically fractured wells²³. Laboratory experiments showed that prophage-induced dominant microorganism lysis releases intracellular metabolites that can sustain key fermentative metabolisms²³, supporting the persistence of microorganisms in this ecosystem. These findings reveal the potential significance of viruses in shaping microbial communities and reprogramming microbial metabolisms within the oil reservoir ecosystem.

In addition, oil reservoirs harbor a broad diversity of uncultured microorganisms and novel metabolic pathways. Zhou et al. suggested that the archaeon '*Candidatus Methanoliparum*' alone has the ability to degrade various large hydrocarbons into methane⁸. Meckenstock et al. discovered complex microbial communities inhabiting small water droplets within the oil phase, and it is widely believed that microbial degradation occurs within these microenvironments²⁴. Diverse functional gene groups identified from *Pseudomonas* in the oil phase were significantly differed from those in the corresponding water phases²². This indicated that oil reservoirs are a unique environment, containing uncultured microorganisms and their co-evolved novel viruses. Therefore, the study of oil reservoir viruses is conducive to revealing global viral diversity.

However, the study of viruses in oil reservoirs is still in its early stages, with only a small number of viral genomes having been identified to date²³. Moreover, existing studies on viromes in oil reservoirs have predominantly focused on a limited spatial scale and specific sample types^{21,23}. Consequently, a comprehensive understanding of viral communities in oil reservoirs remains elusive, and little is known about the distribution patterns and assembly processes of viral communities in these environments.

In this study, to gain insights into the diversity, virus-host interactions, and potential ecological roles of viruses within oil reservoirs, we assemble a catalogue encompassing viral and prokaryotic genomes sourced from 182 oil reservoir metagenomes, and analyze the ecological roles of viruses in regulating the community structure of sulfate reducing microorganisms through combined microcosm enrichment experiments and bioinformatics analysis. Our findings reveal that viruses are widely distributed in oil reservoirs and viruses not only inhibit the growth of SRMs, but also shape the structure of the sulfate reducing microbial subcommunity.

Results

Overview of prokaryotic genome and viral genome catalogues in oil reservoirs

To explore the diversity and ecological function of viruses inhabiting oil reservoirs, we collected 59 oil reservoir produced water samples

from across China and performed whole shotgun metagenomic sequencing. In order to compile catalogs of both prokaryotic and viral genomes from these metagenomic datasets, we created an analysis pipeline (see Supplementary Fig. 1). Employing this pipeline, we analyzed the 59 metagenomes newly generated in this study, plus an additional 123 metagenomes obtained from public databases, originated from oil reservoir samples collected from Europe, Asia, North America, and South America (Fig. 1a). Consequently, we successfully constructed catalogs of both prokaryotic and viral genomes using a total of 182 metagenomic datasets, including 59 newly generated metagenomes from our study and 123 metagenomes obtained from public databases. The prokaryotic genome catalogue consisted of 7229 microbial metagenome-assembled genomes (MAGs) that met or exceeded medium quality ($>50\%$ completeness and $<10\%$ contamination), encompassing 6,686 bacterial and 543 archaeal MAGs, spanning 72 bacterial and 9 archaeal phyla. The bacterial community was dominated by Proteobacteria (2402 MAGs), while the archaeal community was predominantly represented by Halobacteriota (339 MAGs). The viral catalogue of the oil reservoirs consisted of 33,657 putative viral genomes, all exceeding 10 kb in size. This catalog featured 6,218 viral genomes of medium- to high-quality ($>50\%$ completeness). To assess the coverage of viral communities within the oil reservoir, we conducted sequencing of virus-like particles (VLPs) from 8 out of the 59 produced water samples mentioned above²⁵. The result revealed that 77.32% of the VLPs could be mapped to the viral catalogue, suggesting that the viral catalogue extensively covered the viral population within the oil reservoir environment²⁶.

We clustered the 6218 viral genomes into 3886 viral Operational Taxonomic Units (vOTUs) with a sequence similarity threshold of 95%, a level that approximates species-level taxonomy²⁷. The viral richness of all samples ranged from 1 to 1020 with an average of 206 (Fig. 1b). Rarefaction analysis showed that the number of detected vOTUs was saturated (Supplementary Fig. 2), suggesting that our study provided a reasonably comprehensive sampling of the viral communities within oil reservoirs. We assigned the vOTUs to taxonomic ranks based on the latest ICTV classification using PhaGCN2.0 and geNomad v1.9, and found that 44.49% of vOTUs could not be confidently annotated at the family level (Fig. 1c), in addition, only 5.20% of vOTUs could be found in the IMG/VR v3 dataset, underscoring the presence of a substantial number of unidentified viruses within oil reservoirs. Annotated vOTUs were predominantly assigned to the class Caudoviricetes (formerly known as the order Cadovirales; 54.48%), including families such as Peduoviridae (24.94%), Casadabanvirus (7.54%), Hendrixvirinae (2.86%), and Casjensviridae (2.47%). Furthermore, we identified three core vOTUs that were present in more than 50% ($n = 91$) of all 182 oil reservoir metagenomes analyzed in this study. Notably, these core vOTUs represented a mere 0.08% of the total vOTUs. The majority of vOTUs (84.64%) were detected in less than 10% of the samples, highlighting the heterogeneous nature of viral communities within oil reservoirs. Additionally, our analysis predicted 1722 and 1774 vOTUs as virulent and temperate viruses, respectively, while the lifestyles of the remaining vOTUs ($n = 390$) could not be determined (Fig. 1d).

Biogeography of viral communities in oil reservoirs

To investigate how viral communities assembled within oil reservoirs, we compared the viral diversity and composition across various oil reservoir samples. We found significant disparities in viral diversity and composition associated with different geographic locations (Supplementary Fig. 3a, Fig. 2a, ANOSIM, $R = 0.59$, $P = 0.001$), and this disparity was not due to the sequencing depth of the samples included in the study (Supplementary Fig. 3b, Spearman's correlation coefficient $\rho = -0.04$, $P = 0.80$). These results suggested that geographic location serves as the primary determinant of viral variation between samples on an intercontinental geographic scale. This discovery aligns with previous findings in studies of viral communities in cold seep and acid

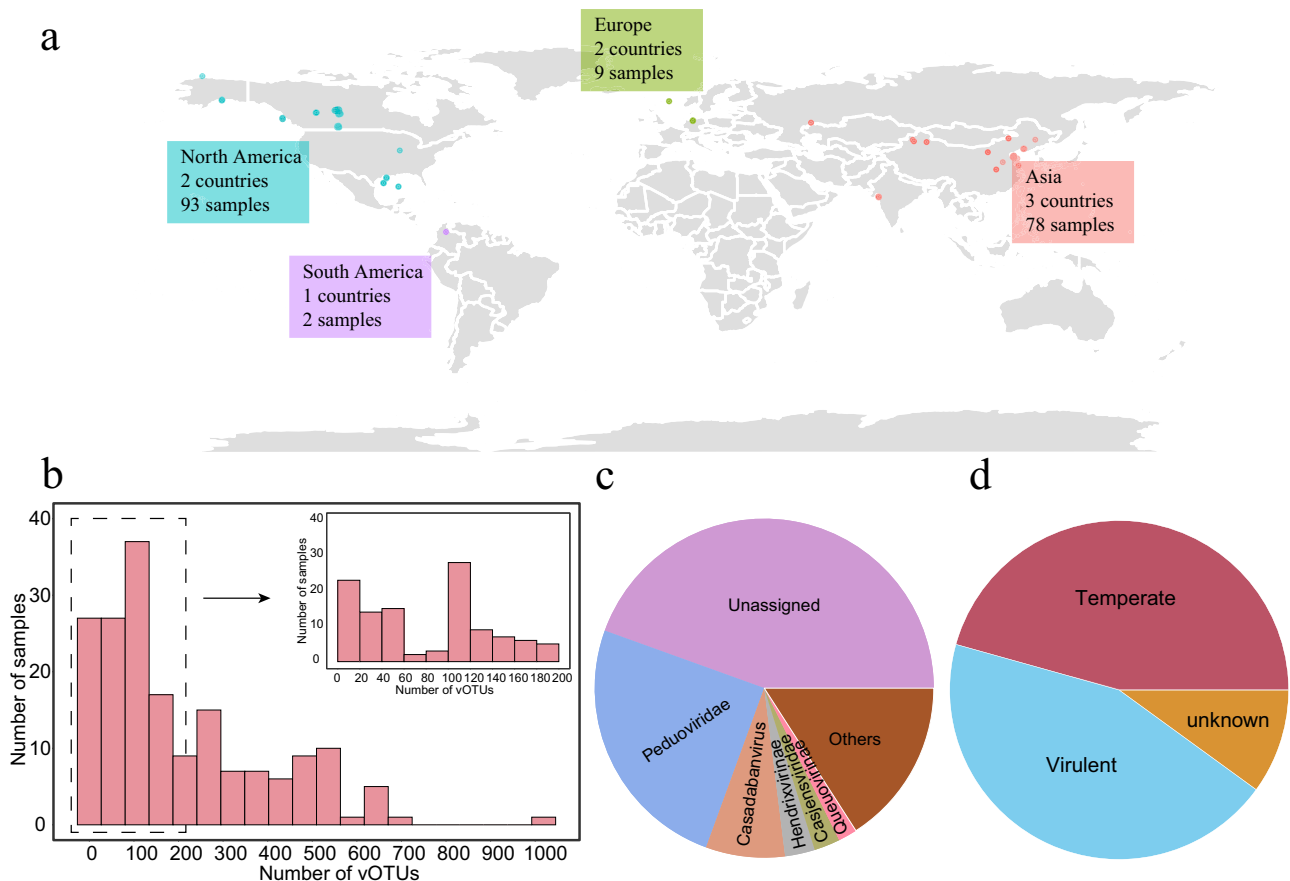


Fig. 1 | Overview of viruses in oil reservoir ecosystems. **a** Geographic distribution of collected oil reservoir samples. **b** Distribution of vOTUs across samples. **c** Relative proportion and taxonomic classification of vOTUs based on the latest

ICTV classification using PhaGCN2.0 and geNomad v1.9. **d** Predicted lifestyles of vOTUs within the oil reservoirs.

mine drainage environments^{28,29}. Within the dataset employed in this study, the majority of oil reservoir samples from Canada and China were dominated by Peduoviridae, while a few Chinese oil reservoir samples exhibited dominance by Casadabanvirus. The majority viruses in samples from oil reservoirs in the United States remained unclassified at the family level (Supplementary Fig. 3c).

Furthermore, we observed significant negative distance-decay relationships across all samples based on the Bray-Curtis similarities (1 – dissimilarity) of viral communities (slope = -0.05 , $P < 0.001$) (Fig. 2b). Mantel tests indicated that the compositions of viral communities were significantly related to geographic distance (Spearman's correlation coefficient $\rho = 0.30$, $P = 0.001$) (Supplementary Table 1). To determine the relative importance of geographic factor in structuring viral communities, we conducted multiple regression analysis using multiple regression on matrices (MRM) approach, MRM results showed that 16.53% of the variability in viral community similarity was explained by geographic distance (partial regression coefficient $b = -0.37$, $P = 0.0001$). To explore the mechanisms for viral community assembly within oil reservoirs, we performed a null model analysis. The result revealed that stochastic processes played a more substantial role than determinism in governing viral community assembly in oil reservoirs (Fig. 2c). Moreover, we utilized environmental metadata associated with this study's 59 newly sequenced oil reservoir metagenomes to investigate the influence of environmental factors on the viral community. Correlations of individual physicochemical parameters with alpha diversity showed that the alpha diversity of viral community was significantly correlated with pH, K^+ , Mg^{2+} ($P < 0.05$; Fig. 2d). The Shannon-Wiener and richness index all increased as pH increased, but all decreased as K^+ and Mg^{2+} concentrations increased

(Fig. 2d). In contrast, no significant correlation between viral alpha diversity and temperature was found (Supplementary Fig. 3d). For beta diversity among viral communities, Mantel tests showed that the similarity in viral community composition among samples was significantly correlated with environmental distance (Spearman's correlation coefficient $\rho = 0.27$, $P = 0.0001$) and geographic distance (Spearman's correlation coefficient $\rho = 0.19$, $P = 0.001$). MRM was used to further identify the relative contributions of environmental factors versus geographic distance to viral community similarity (Supplementary Table 2). In oil reservoirs, the MRM model explained 14.42% of the variability in viral community similarity ($P = 0.0001$), with geographic distance, pH, and temperature being the most important variables explaining viral community similarity (partial regression coefficient $b = -0.14$, -0.11 , and -0.071 , respectively, $P < 0.0001$). NMDS analysis also illustrated the significant dissimilarities in viral communities between different pH (Fig. 2e) or temperature (Fig. 2f) oil reservoirs (pH: ANOSIM, $R = 0.28$, $P = 0.001$; temperature: ANOSIM, $R = 0.24$, $P = 0.001$).

Novel viral clusters in oil reservoirs

To uncover novel viruses inhabiting oil reservoirs, we employed vConTACT2³⁰ to construct a gene-sharing network that included vOTUs found in oil reservoirs and from a wide variety of other ecosystems, such as groundwater, sediment, and thermal springs, etc. In this weighted network, all vOTUs were grouped into 4011 viral clusters (VCs) (Fig. 3a). Among these VCs, 1934 were exclusively associated with a single ecosystem, while only two were shared across all ecosystems. The limited overlap of viruses between different ecosystems demonstrated a high degree of habitat specificity among viruses. Within the

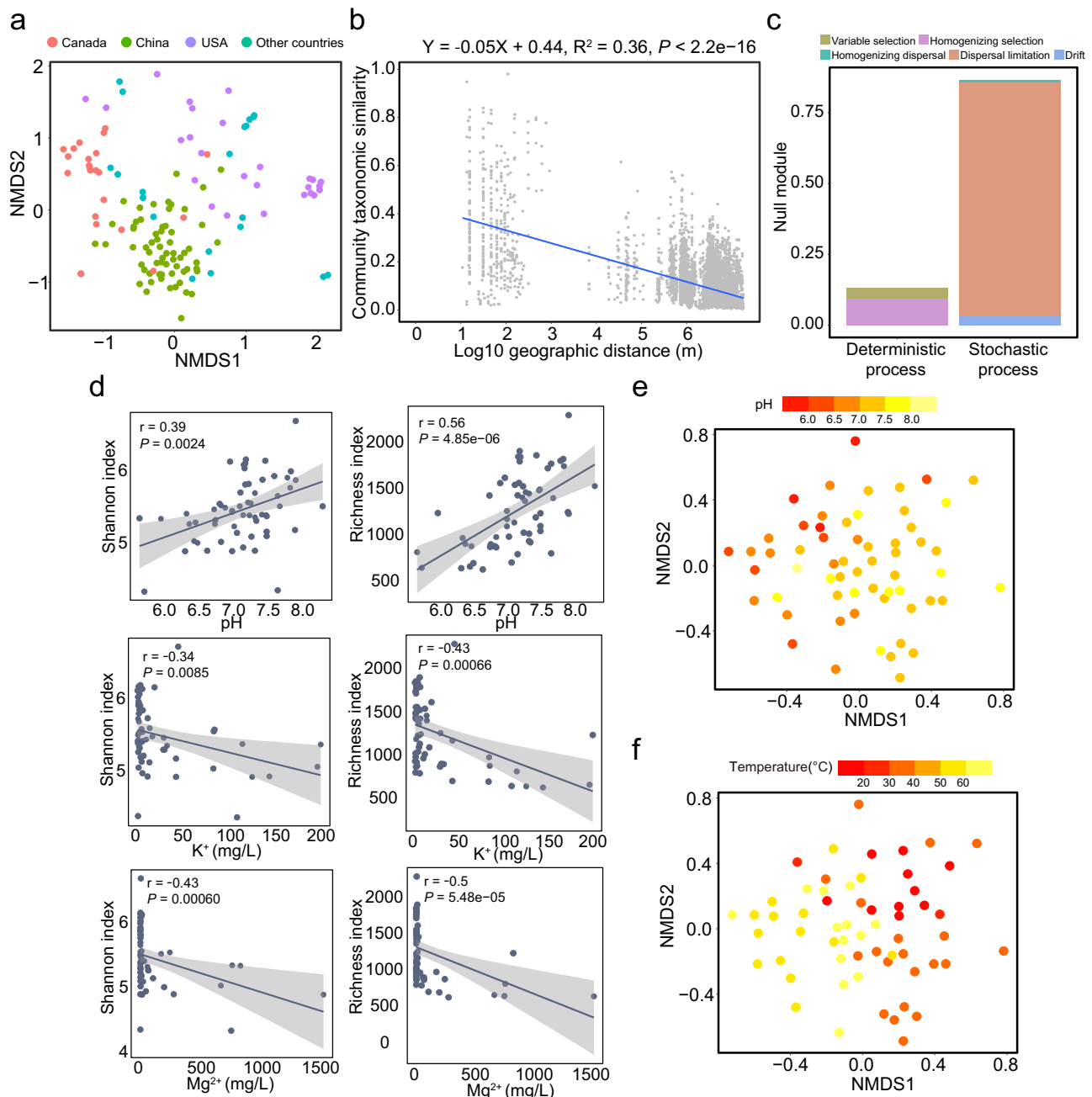


Fig. 2 | The distribution and assembly process of viral communities.

a Nonmetric multidimensional scaling (NMDS) of viral communities colored by sampling sites. **b** Distance-decay relationships based on Bray-Curtis similarity (1 - dissimilarity of viral communities). The blue line denotes the least-squares linear regression across all spatial scales. Fitting equation, adjusted R^2 values, and P values for Distance-decay are based on linear regression and presented above graph. **c** Proportions of viral community assembly processes in oil reservoirs, including variable selection, homogenizing selection, homogenizing dispersal, dispersal

limitations, and drift. **d** Correlations between individual physicochemical parameters (pH, K^+ , and Mg^{2+}) and Shannon-Wiener (left) and richness index (right) were examined using a Spearman's rank test (two-sided P -value), shaded area shows 95% confidence interval around the fitted mean line. **e** NMDS of viral communities from different pH oil reservoirs. **f** NMDS of viral communities from different temperature oil reservoirs. Sample points are colored according to oil reservoirs pH (**e**) and temperature (**f**). Color bars from red to yellow represent values from small to large, respectively.

subset of oil reservoir viruses, 2942 out of 3886 vOTUs were clustered into 805 VCs, with 265 VCs (33%) being unique to oil reservoirs. This finding suggested that about one-third of oil reservoir viruses may be endemic to oil reservoirs. Furthermore, we found that acid mine drainage sediments and groundwater shared a higher number of VCs with oil reservoirs (Fig. 3a). Additionally, we constructed a gene-sharing network based on the vOTUs found in oil reservoirs and those deposited in the Viral RefSeq database. Our analysis showed that only a small percentage ($n = 495$, 14.66%) of vOTUs from oil reservoirs

clustered with taxonomically known genomes from Viral RefSeq (Fig. 3b). These results indicated that oil reservoirs possess a vast array of undescribed viruses.

Functional genes encoded by viromes in oil reservoirs

To unravel the functional roles of viruses within oil reservoirs, we clustered all 346,145 predicted protein-coding genes derived from the oil reservoir viral genomes into 322,060 gene clusters. We found that 61.32% of viral gene clusters (accounting for 63.90% of total viral

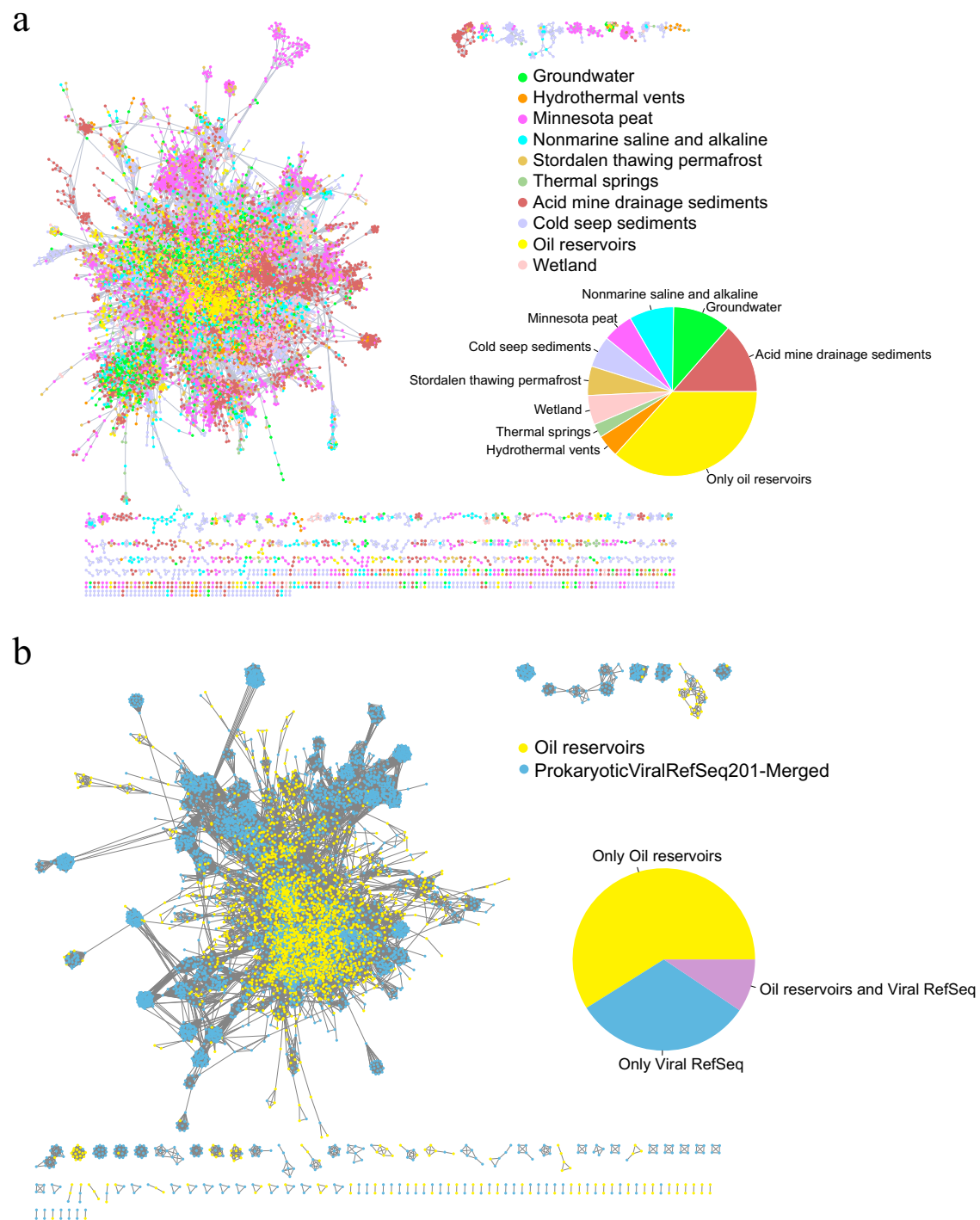


Fig. 3 | Taxonomic diversity of viruses in oil reservoirs. a Gene-sharing network of viruses across various environments, including oil reservoirs, groundwater, cold seep sediments, thermal springs, wetland, acid mine drainage sediments, Minnesota peat, non-marine saline and alkaline, hydrothermal vents, and Stordalen thawing permafrost. The nodes in the network represent viruses, while the edges indicate similarity based on shared protein clusters. Node color represents the

origin environment of the viruses. Pie chart showing the relative proportion of shared viral clusters between oil reservoirs and the other nine environmental virus datasets. **b** Gene-sharing network of viruses from oil reservoirs and RefSeq prokaryotic viruses. Pie graph showing the relative proportion of shared viral clusters between oil reservoirs and RefSeq prokaryotic viruses.

genes) lacked functional annotation against the eggNOG database. The largest gene cluster derived from viromes in oil reservoirs predominantly encoded proteins associated with replication, recombination and repair (L) functions. (Supplementary Fig. 4a).

Furthermore, we identified 105 putative auxiliary metabolic genes (AMGs) that might participate in host metabolism and adaptation to

the environment, mainly involved in carbon utilization ($n = 32$), energy metabolism ($n = 25$), miscellaneous metabolism (MISC, $n = 34$), and transporters metabolism ($n = 14$). Among these AMGs, genes involved in cobalamin biosynthesis (including *cobS* and *cobT*) were the most prevalent (Supplementary Fig. 4b), being found in 25 samples. Cobalamins are a class of structurally diverse cofactors containing cobalt³¹

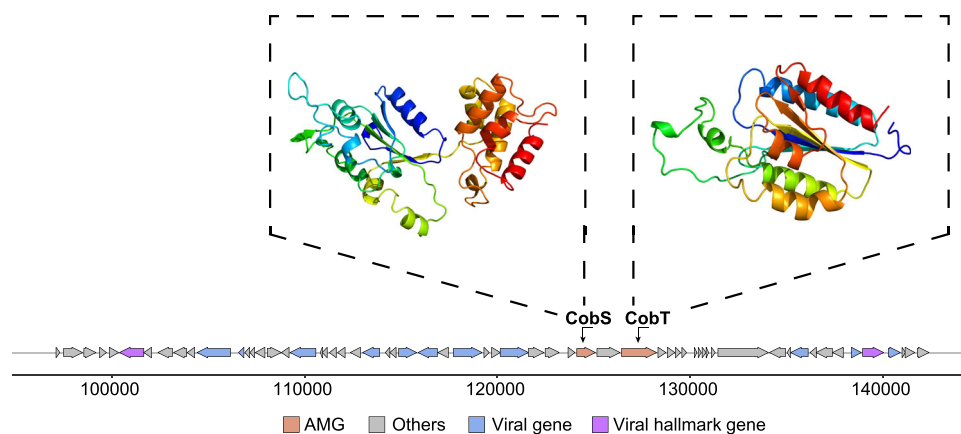


Fig. 4 | Genome map of a representative *cobS* and *cobT* encoding virus. AMGs are in orange, virus-like genes are in blue, viral hallmark genes are in purple, and non-virus-like or uncharacterized genes are in gray. Tertiary structures of selected AMGs based on structural modeling using Phyre2.

that are essential for various biological functions, including amino acid synthesis and carbon metabolism³². VirSorter2 placed the *cobS* and *cobT* within the interior of the viral genome. In addition, the structural model prediction of virus-encoded CobS and CobT at Phyre2 showed 100% confidence (Fig. 4). These findings suggest that viruses in oil reservoirs could potentially influence microbial interactions by regulating the production of cobalamin within their hosts. In addition, we identified 9 AMGs from complete/high-quality viruses that were predicted to participate in energy metabolism (Supplementary Fig. 4c, d, see Supplementary Discussion for more information).

Virus-host interactions in oil reservoirs

To further study the potential impacts of viruses on microbial ecology within oil reservoirs, we investigated in situ virus-host interactions based on CRISPR spacers and tRNA sequences³³. A total of 7197 putative virus-host pairs were predicted (1411 based on CRISPR, 5,786 based on tRNA), in which 1119 viruses (18.00% of the total viruses) were connected to 1217 prokaryotic MAGs (host MAGs, accounting for 16.83% of the total MAGs). The host MAGs were distributed across 3 archaeal phyla and 37 bacterial phyla (Fig. 5a, Supplementary Fig. 5a). The top five most frequently predicted bacterial phyla included Proteobacteria (785 MAGs), Actinobacteriota (69), Desulfobacterota (64), Chloroflexota (50), and Bacteroidota (49). The most frequently predicted archaeal phyla were Halobacteriota (10) and Methanobacteriota (8) (see Supplementary Discussion for more information). In addition, we found 422 host MAGs with the potential for hydrocarbon degradation.

To investigate the relationship between viruses and microbes in oil reservoirs, we conducted a correlation analysis between viral and microbial profiles. Our analysis revealed a significant positive correlation of α -diversity (Shannon diversity) between the viral and prokaryotic communities ($r = 0.68$, $P < 2.2 \times 10^{-16}$; Fig. 5b). This result indicated that the structures of the viral and prokaryotic communities were closely related within oil reservoir environments. To further explore their association, we examined the relationship between the relative abundance of each virus and its predicted host. Our results revealed a positive correlation between viral and microbial relative abundances ($r = 0.68$, $P = 1.05 \times 10^{-149}$; Fig. 5c). Furthermore, a negative correlation was observed between Viral-to-Host Ratios (VHRs) and microbial relative abundance within oil reservoirs, showing that hosts with higher relative abundances tended to exhibit lower VHRs (Fig. 5d). This phenomenon aligns with ‘piggybacking the winner’ model, which posits that viruses exploit their hosts through lysogeny rather than killing them when host density is high. Temperate viruses protect their host cells from infection by closely related viruses via superinfection exclusion³⁴. Consequently, the contribution of temperate viruses to

the host increases with high host abundance, resulting in a ‘more microbes, fewer viruses’ scenario. To further clarify whether viruses affected their prokaryotic hosts in oil reservoirs through a ‘piggybacking the winner’ mechanism, we compared the relative abundance of temperate and virulent viruses with identified host links, and found that although both virulent and temperate viruses exhibited positive correlations in relative abundance with their hosts, the increase in the relative abundance of temperate viruses occurred at a higher ratio than that of virulent viruses (Fig. 5e), and the VHRs of temperate viruses exhibited a milder decline compared to those of virulent viruses (Fig. 5f). Additionally, we found that host relative abundances exhibited a negative correlation with the ratio of virulent and temperate viruses (Fig. 5g), whereas the VHR displayed a positive correlation with the ratio of virulent and temperate viruses (Fig. 5h). Collectively, these findings suggest that oil reservoir viruses affect their hosts mostly in the manner of ‘piggybacking the winner’.

Sulfur metabolism mediated by viruses in oil reservoirs

In oil reservoirs, a prevalent metabolic pathway involves the microbial reduction of sulfate to generate hydrogen sulfide (H_2S), a process primarily conducted by sulfate-reducing microorganisms (SRMs)^{35,36}. H_2S has significant implications for the degradation of infrastructure, reservoir souring, the operational cost of oil production, and the value of crude oil³⁷. Recent studies reported that viruses infecting sulfur-metabolizing microbes could reshape the host’s sulfur metabolism through encoded auxiliary metabolic genes (AMGs)³⁸.

To investigate the potential impact of viruses on microbial reduction of sulfate in oil reservoirs, we identified the MAGs possessing sulfur metabolic capabilities and their associated viruses, including assimilatory sulfate reduction (ASR), thiosulfate oxidation metabolism (TSO), dissimilatory sulfate reduction (DSR), and sulfide oxidation (SO). A total of 484 host MAGs were predicted to possess the pathways for different types of sulfur metabolism (Supplementary Fig. 7c, see Supplementary Discussion for more information). The relative abundance of host MAGs with sulfur metabolic capabilities and their associated viruses were positively correlated (Fig. 6a), and the VHRs exhibited a decline corresponding to the increasing relative abundance of hosts with sulfur metabolic capabilities (Fig. 6b). These results were consistent with the findings observed in the overall MAG population. However, we noticed a parallel increase in the relative abundances of virulent and temperate viruses in tandem with the rise in host relative abundance (Fig. 6c). This difference in relative abundance increase ratio between virulent and temperate viruses was not as pronounced as the trend observed for the entire host population (Fig. 5e, Fig. 6c). In addition, virulent viruses displayed a significantly stronger correlation with hosts engaged in sulfur metabolism when

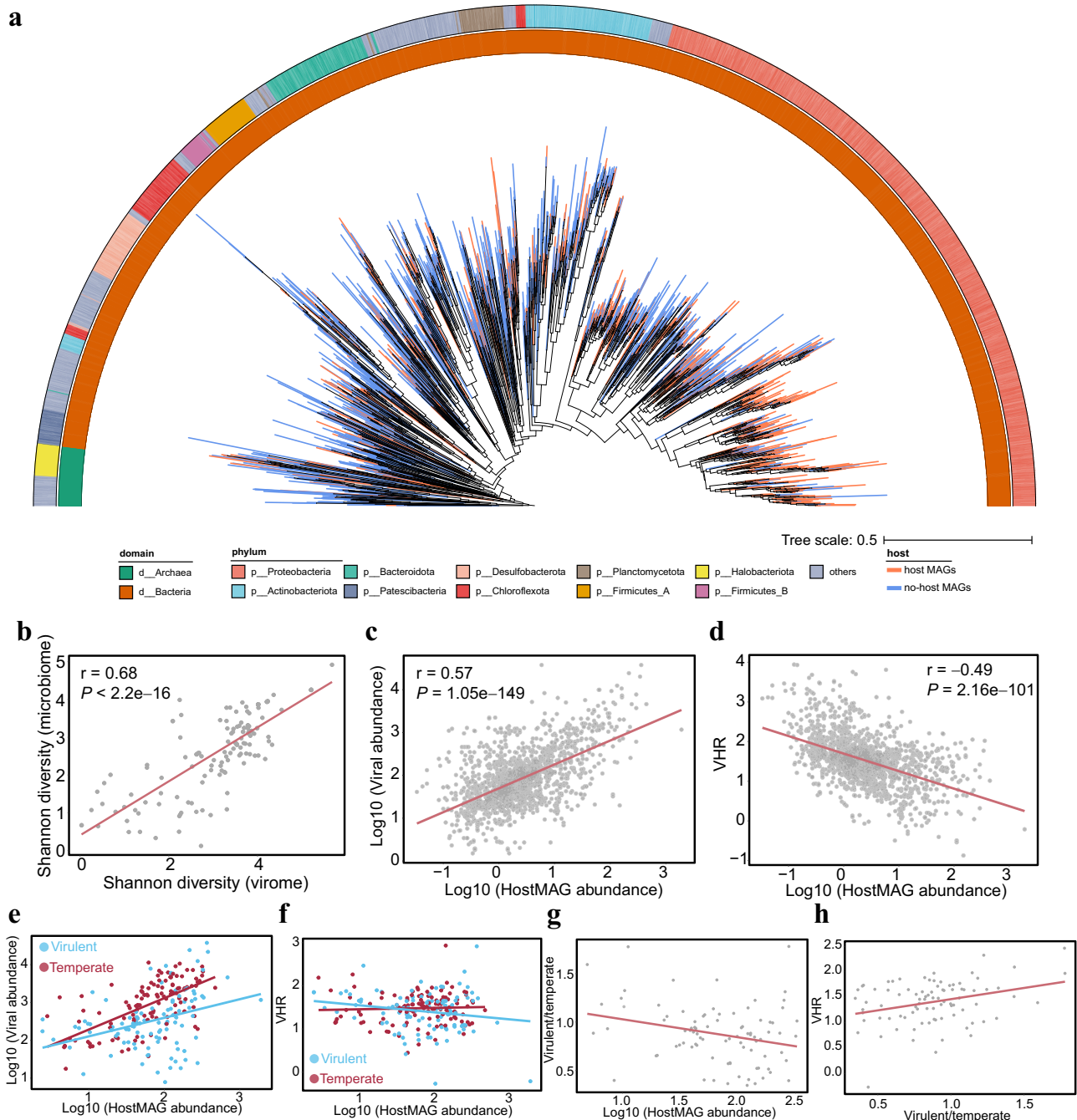


Fig. 5 | Linkages between each virus and its predicted host MAG in oil reservoirs. a Phylum level maximum-likelihood phylogenetic tree of MAGs detected in the oil reservoirs. The color of clades in the tree indicates whether MAGs are host MAGs (orange represents host MAGs, blue represents no-host MAGs). The inner circle represents the domain, and the outer circle represents the phylum as annotated by GTDB. The tree was constructed using PhyloPhlAn and visualized with iTOL. **b** Comparisons of α -diversity (Shannon index) between the virome and the microbiome are presented. **c, d** Correlation between the relative abundance of hosts and viruses (**c**) or VHR (**d**) in the oil reservoirs. In (**b–d**), Correlations are

examined using a Spearman's rank test (two-sided P -value). Regression lines are shown in red. **e, f** Associations between the relative abundance of hosts and viruses (**e**) or VHR (**f**) (temperate and virulent viruses) in different samples. The red points represent temperate viruses, and the blue points represent virulent viruses, each point represents one sample. Regression lines are shown in red and blue for temperate and virulent viruses, respectively. **g, h** Correlation between the relative abundance ratio of virulent-to-temperate viruses and the relative abundance of hosts (**g**) or VHR (**h**). Regression lines are shown in red.

compared to temperate viruses (Fig. 6d). Collectively, these findings point to a relatively more substantial contribution by virulent phages to the regulatory dynamics of sulfur metabolizing communities. Moreover, we found that hosts with DSR capabilities exhibited a relatively stronger correlation with their associated viruses in comparison to other hosts (Fig. 6e). This high correlation could be attributed to the higher prevalence of virulent viruses infecting these hosts compared

to other hosts (Supplementary Fig. 7c). In addition, both virulent and temperate viruses infecting hosts capable of DSR showed a relatively stronger correlation with their hosts compared to other hosts (Fig. 6f, g). These findings collectively suggested that viruses primarily regulated DSR function within oil reservoirs, possibly in a top-down regulatory manner³⁹, where the growth and abundance of SRMs are primarily regulated by the predation of virulent viruses.

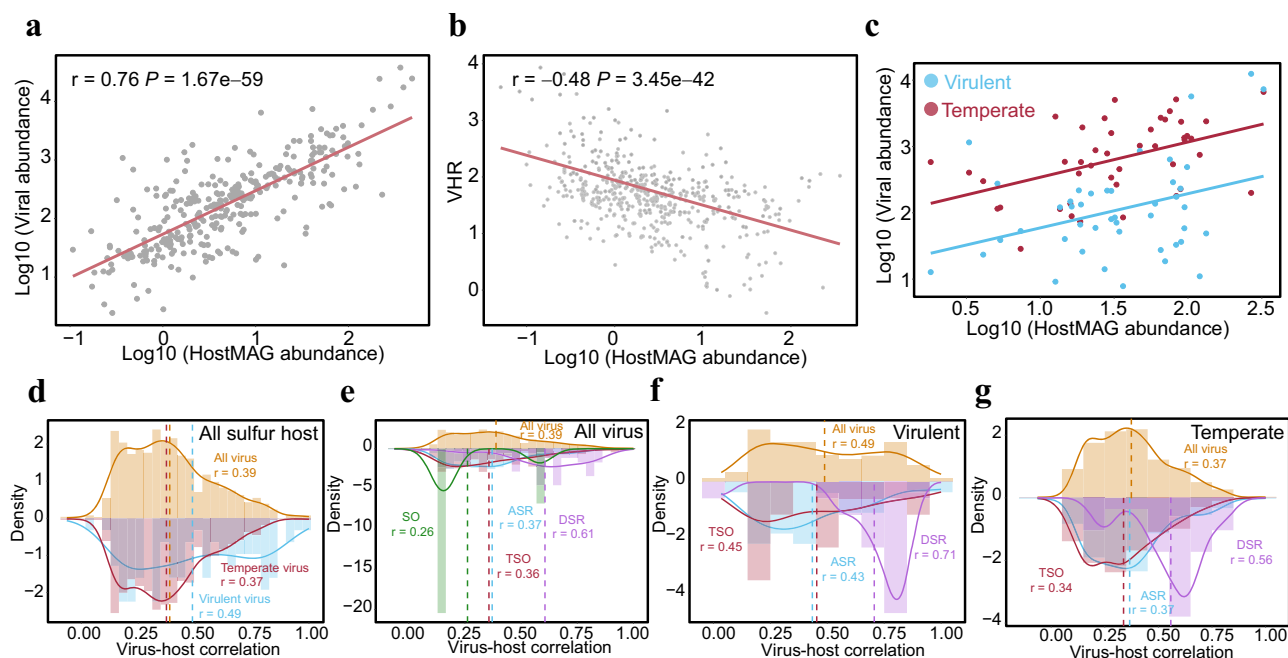


Fig. 6 | The impact of viruses on the sulfur cycle in oil reservoirs. a, b Correlation between the relative abundance of hosts with sulfur metabolic capabilities and viruses (**a**) and VHR (**b**) were examined using a Spearman's rank test (two-sided P value). Regression lines are shown in red. **c** Associations between the relative abundance of hosts with sulfur metabolic capabilities and viruses (temperate and virulent viruses) in different samples. Red points represent temperate viruses, blue points represent virulent viruses, each point represents a single sample. Regression lines are shown in red for temperate and blue for virulent viruses. **d** Distribution of virus-host correlations with sulfur metabolic capabilities. Orange, blue, and red colors represent the distributions of all viruses, virulent, and temperate viruses,

respectively. **e** Comparison of distribution of virus-host correlations with different sulfur metabolic capabilities. Orange color represents the distributions of all viruses, while blue, green, purple, and red colors represent viruses infecting different functional sulfur metabolism groups. **f, g** Distribution of virulent (**f**) or temperate (**g**) viruses and host correlations with sulfur metabolic capabilities. Orange color represents the distributions of all virulent (**f**) or temperate (**g**) viruses, while blue, purple, and red colors represent virulent (**f**) or temperate (**g**) viruses infecting different functional sulfur metabolism groups. Dashed lines indicate the average correlation in the distribution.

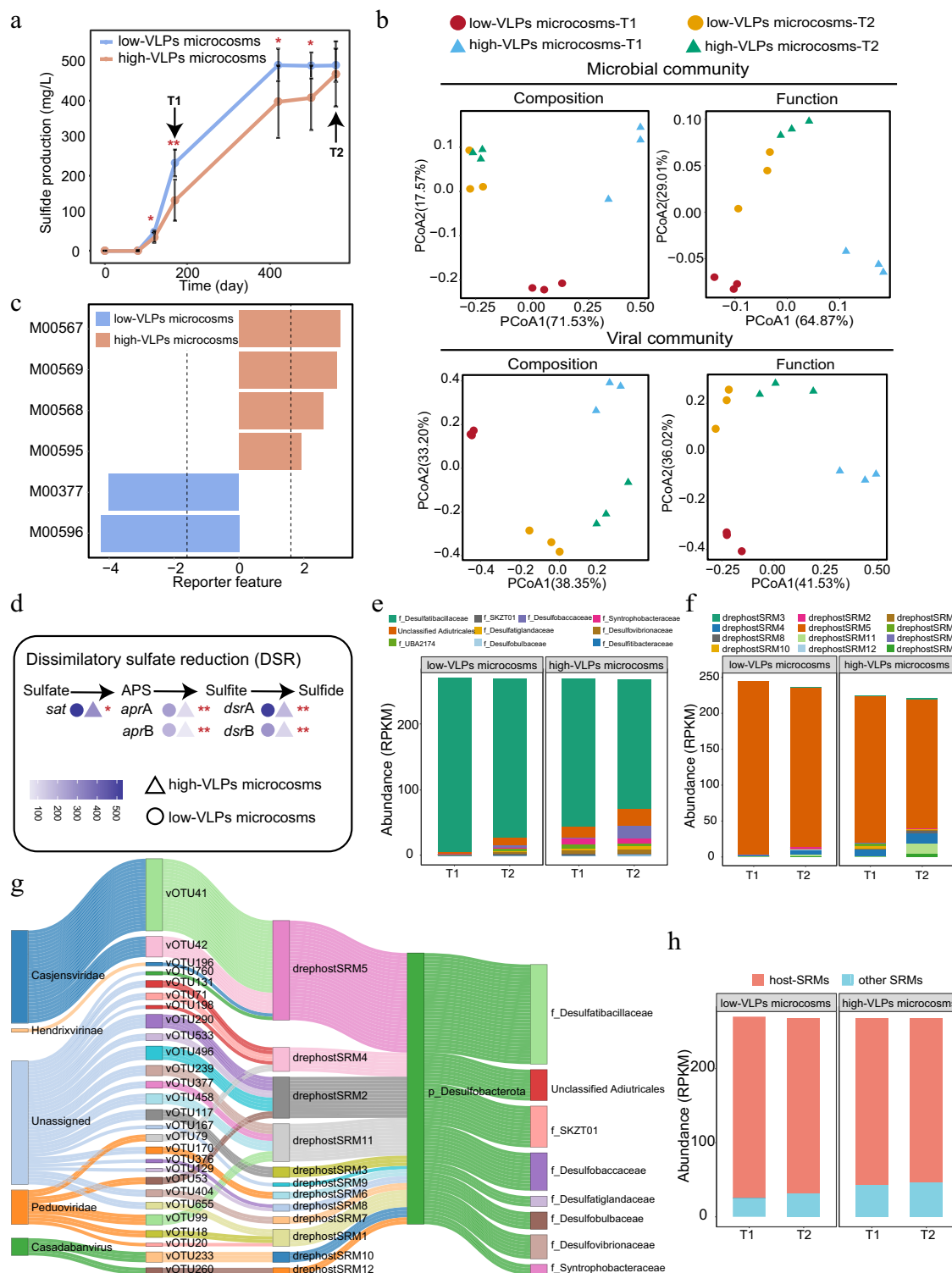
To further investigate the influence of viruses on sulfate reduction, we employed oil reservoir production water from the Huabei Oilfields as the initial inoculum and set up two distinct sets of microcosms, namely low-virus-like particles (VLPs) microcosms and high-VLPs microcosms (Supplementary Fig. 12). In brief, the production water underwent centrifugation, filtration, and ultra-concentration processes to acquire microbial pellets, virus-rich filtrate, and virus-poor filtrate, which were then used to establish two types of microcosms. In the low-VLPs microcosms, microbial pellets were diluted with virus-poor filtrate (Microbes–Viruses), while in the high-VLPs microcosms, microbial pellets were diluted with virus-rich filtrate (Microbes+Viruses). As a result, the high-VLPs microcosms featured a higher initial count of VLPs in comparison to the low-VLPs microcosms (see detailed in method). Over the course of 570 days, we monitored sulfide production, and the viral communities were analyzed at two time points (160 days and 570 days) after the start of incubation (i.e., T1 and T2). Detailed information on putative viral contigs and host MAGs can be found in the Supplementary Discussion.

Throughout the incubation period, we observed significantly lower levels of sulfide in the high-VLPs microcosms (Fig. 7a, Supplementary Data 4). Additionally, at the T1 time point, the diversity of both the overall microbial community and host-MAG subcommunity was significantly higher in the high-VLPs microcosms compared to low-VLPs microcosms (Supplementary Fig. 8c).

To examine the potential influence of initial VLP counts on the composition and function of both the microbial and viral communities, we conducted principal coordinates analysis (PCoA), which revealed distinct dissimilarity patterns in both the composition and function of the microbial community between the two microcosms (ANOSIM,

$R_{\text{composition}} = 0.85$, $P_{\text{composition}} = 0.001$, $R_{\text{function}} = 0.91$, $P_{\text{function}} = 0.001$) (Fig. 7b). Similar to the microbial community, the composition and function of the viral communities showed different phases (ANOSIM, $R_{\text{composition}} = 0.95$, $P_{\text{composition}} = 0.001$, $R_{\text{function}} = 0.96$, $P_{\text{function}} = 0.001$) (Fig. 7b). As the incubation progressed, the differences in diversity and composition of the microbial and viral communities between the two sets of microcosms diminished (Fig. 7b, Supplementary Fig. 8d). Moreover, we compared the functional characteristics related to energy metabolism and hydrocarbon degradation within the low-VLPs and high-VLPs microcosms. Notably, the M00596 module, associated with DSR function, exhibited enrichment in the low-VLPs microcosms at T1 (Fig. 7c, Supplementary Data 5). Additionally, the relative abundance of genes involved in DSR was significantly lower in the high-VLPs microcosms compared to the low-VLPs microcosms at T1 (Fig. 7d). These results indicated a significant correlation between a high number of viruses and a decrease in H_2S production. This decline was attributed to a decrease in the relative abundance of genes involved in DSR within the microcosms.

To gain insight into the potential influence of the initial VLP counts on the SRM community, we compared the relative abundance and composition of SRMs between the different sets of microcosms. In total, we identified 89 SRMs (referred to as all-SRMs) including 44 SRMs associated with viruses (referred to as host-SRMs) (Supplementary Fig. 9, Supplementary Data 6). In the high-VLPs microcosms, the SRM subcommunity displayed a higher diversity than the low-VLPs microcosms, which was similar with the findings observed within the entire microbial community (Fig. 7e, Supplementary Fig. 10a, b). This result could be attributed to the decreased abundance of the dominant SRM (drephostSRM5, *Desulfatibacillaceae*) in the community, which in turn led to an increase in the abundance of rare SRMs, such as



drephostSRM3 (*Desulfatiglandaceae*), *drephostSRM4* (*Unclassified Aditricales*), and *drephostSRM10* (*Syntrophobacteraceae*) (Fig. 7f, g).

Collectively, the results suggested that the high number of viruses in the microcosm restructured the sulfate reducing microbial subcommunity. In addition, we compared the total relative abundance of SRMs within the overall microbial community in different sets of microcosms. We found that the total relative abundances of SRMs in low-VLPs microcosms were significantly higher than in high-VLPs microcosms at T1 ($P=0.024$). Furthermore, we also found that the majority of the SRMs could be associated with identified viruses (Fig. 7h), suggesting that viral infection targeting SRMs was a

significant factor in the reduction in H_2S production. The high initial VLP count primarily attenuated H_2S production by regulating the composition of SRMs.

Discussion

In this study, we conducted large-scale data mining of metagenomes to construct a catalog of viromes from oil reservoirs. This resource contains a wide range of viral genomic diversity that is unique to oil reservoirs. It encompasses diverse and previously uncharacterized viral groups. In addition, we preliminarily verify the potential impact of viruses on the sulfate reducing microbial subcommunity. These

Fig. 7 | Viral impact on dissimilatory sulfate reduction. **a** Sulfide production in microcosms. Differences of sulfide production between low-VLPs microcosms and high-VLPs microcosms were assessed using two-tailed Student's *t*-test. Every point show mean \pm SD with three biological replicates per treatment ($n = 3$). The red asterisks indicate a significantly higher level of sulfide production in the low-VLPs microcosms compared with the high-VLPs microcosms at the corresponding time point ($^*P < 0.05$, $^{**}P < 0.01$). Detailed information of the sulfide production is provided in Supplementary Data 4. **b** Changes in microbial and viral community composition (left) and function (right) at the T1 and T2 sampling time points. **c** Reporter score evaluated the enrichment of modules between the two microcosm types. **d** Comparison of the relative abundance of dissimilatory sulfate reduction related genes in the low-VLPs microcosms ($n = 3$) and high-VLPs microcosms ($n = 3$) at the T1 sampling time. The gradient dark blue circle represents the average RPKM value of the genes in the low-VLPs microcosms. The gradient dark blue triangle

represents the average RPKM value of the genes in the high-VLPs microcosms. The differences of the relative abundance of dissimilatory sulfate reduction related genes between low-VLPs microcosms and high-VLPs microcosms were assessed using two-tailed Student's *t*-test. The significant difference was highlighted by asterisk ($^*P < 0.05$, $^{**}P < 0.01$). Detailed information of the relative abundance of dissimilatory sulfate reduction related genes is provided in Supplementary Data 5. **e** Relative abundance variation of all-SRMs in microcosms at the family level. **f** Relative abundance variation of host-SRMs in microcosms. **g** Predicted virus-host links between virus and host SRMs in microcosms. The right two panels represent host taxonomy colored by phylum and family, and the left two panels show viral taxonomy colored by family and viral clusters (vOTUs), connecting lines show associations between host SRMs and viruses. Detailed information of the link between virus and host SRMs is provided in Supplementary Data 6. **h** Relative abundance proportion between host-SRMs and other SRMs in microcosms.

findings will contribute to a better understanding of the ecological role of viruses in the deep biosphere, such as regulating microbial mortality, structuring the microbial community, and impacting biogeochemical cycling (Supplementary Fig. 11).

Recent metagenomic and virome surveys have uncovered a range of viruses in both aquatic and terrestrial environments, significantly enhancing our understanding of virome diversity^{28,40}. However, our knowledge of viral communities in oil reservoirs is still limited. Many of the viruses recovered from the oil reservoirs were novel with previously uncharacterized genomes. Moreover, the taxonomic annotation ratio of oil reservoir vOTUs was lower than that of vOTUs from deep-sea cold seep sediment⁴¹, hydrothermal vent⁴², soil⁴⁰, acid mine drainage sediment²⁹, and human gut⁴³ environments. This low annotation proportion can primarily be attributed to the absence of complete genomes from viral isolates in oil reservoirs and associated environments. Collectively, our findings demonstrate the uniqueness of viruses in oil reservoirs, emphasizing the limited extent of our knowledge regarding viral diversity in this habitat.

Gene-sharing network analysis showed that viruses from oil reservoirs and other ecosystems formed separate cohesive clusters (Fig. 3a), suggesting that these viruses may possess unique metabolism genes to adapt to oil reservoir environments. Oil reservoirs are relatively independent and stable ecosystems that can be isolated for millions of years prior to industrial production practices, leading to the presence of unique microbiomes⁴. The co-evolution of viruses and microorganisms may lead to the distinct nature of oil reservoir viromes compared to those reported in other environments. This underscores the scarcity of research and datasets on viromes in oil reservoirs. In addition, we found that groundwater and acid mine drainage sediments shared more viral clusters with oil reservoirs, the specific cause of this phenomenon remains elusive.

The associations between viral and host abundance have been described by the Kill-the-Winner and Piggyback-the-Winner theories^{34,44,45}. Density- and frequency-dependent lytic Kill-the-Winner models predict that a high bacterial abundance is associated with a high rate of lytic infections, leading to increased VHRs⁴⁵. On the other hand, Piggyback-the-Winner theory suggests that temperate viruses can protect their host cells from closely related viruses via superinfection exclusion, thus at high host densities, rather than killing their hosts, viruses might switch their lifestyle from virulent to temperate life cycle and replicate integrated into their host genomes, resulting in decreased VHRs^{44,45}. This trend has been observed in various ecosystems, from soil to freshwater to human lungs^{46,47}. In oil reservoirs, the associations between viral and host abundance also support the Piggyback-the-Winner theory. The ratio between virulent and temperate viruses decreased with increasing host density (Fig. 5g), suggesting that temperate viruses are a more successful strategy for viral replication at high host densities in oil reservoirs. While this study has shed light on the relationship between viruses and their hosts within oil reservoirs, there may exist potential biases in our virome analysis workflow. First,

while the applied CRISPR arrays and tRNA enabled us to uncover numerous relationships between viruses and microorganisms, it is crucial to acknowledge that not all microorganisms possess CRISPR arrays, and CRISPR arrays may not be assembled and binned into MAGs. Thus, the full virus-host interactions in the oil reservoirs could not be examined. The viral binning method can offer a more effective approach for fragmented multi-contig viral assemblies, allowing for a more precise clustering of viral and microbial populations and facilitating the direct investigation of virus-host interactions⁴⁸. Second, the VHRs are underestimated when relying solely on metagenomic data, which is a common problem in current research. To address this limitation in future studies, the absolute number of viruses in the filtrate can be quantified by introducing an internal reference (a certain biomass of known viruses) before filtration to quantify the loss after filtration.

SRMs are a significant microbial group within oil reservoirs³⁶. Our association analysis showed a significant correlation between hosts capable of dissimilatory sulfate reduction and viruses. To further validate this conclusion, we conducted a microcosm experiment and discovered that viruses decreased sulfide production by reducing the abundance of genes involved in dissimilatory sulfate reduction. Additionally, we observed that the sulfate reducing microbial subcommunity followed the 'kill the winner' model. In microcosms with high initial count of virus-like particles, the relative abundance of dominant SRMs decreased due to viral infection. This release of niche space fostered the growth of other lower abundance SRMs. Therefore, in this way, viruses regulate the diversity and structure of the sulfate reducing microbial subcommunity. In summary, viruses not only inhibit the growth of SRMs, but also shape the structure of the sulfate reducing microbial subcommunity. The current study has significantly advanced our understanding of the role of viruses in the sulfur cycle within oil reservoirs. In addition to causing the souring of oil reservoirs, the hydrogen sulfide (H_2S) produced by SRMs also influences the cost of oil production and the value of crude oil³⁷. In future oil exploration and production, it may be possible to introduce viruses for the prevention and treatment of souring in oil reservoirs.

In conclusion, this study presents a viral catalogue of oil reservoirs, revealing the rich diversity of viruses in these environments. Furthermore, our study elucidates a wide array of host-virus interactions and provides evidence for the substantial impact of viruses on microbial sulfate reduction within oil reservoirs. These results shed light on the ecological roles of viruses and their hosts in oil reservoirs.

Methods

Collection of metagenomic data sets for oil reservoirs and geochemical analysis

A total of 59 oil reservoir production water samples were collected from five provinces across China. All samples were collected from the wellheads of each production well, where the oil and water mixture fluid were pumped out. Mixture fluid from wellheads was collected

directly into clean and sterilized 5 L sampling bottles until the bottles were filled to exclude oxygen. All samples were placed in an icebox and transported to the laboratory immediately and stored at 4 °C until DNA extraction. For better separation of oil and water from the mixture fluid, all bottles filled with production mixtures were allowed to stand with gravitational precipitation for 12 h at 4 °C. Subsequently, five hundred milliliters of the water phase were collected for each sample for total microbial genomic DNA extraction. Microbial cells were obtained after filtering through 0.22 µm-pore-size polycarbonate membranes (45 mm diameter; Millipore, Bedford, MA, United States). The polycarbonate membranes with the collected microbial cells were cut into small pieces using sterile scissors, and placed into sterile centrifuge tubes for DNA extraction (FastDNA® SPIN Kit for Soil, MP Biomedicals, USA), following the manufacturer's instructions. Extracted DNA was used for library preparation with NEB Next® Ultra II™ DNA Library Prep Kit for Illumina® (New England Biolabs, USA) and sequenced on Illumina NovaSeq 6000 platform (150 bp, paired-end reads). In total, 380 billion raw sequence reads were retained for analysis (Supplementary Data 1). For the geochemical analysis, the concentrations of cations and anions in the produced waters were analyzed using an ion chromatograph (SHINE CIC-DI60) with an SH-AC-3 column (for cation analysis) and an SH-CC-9 column (for anion analysis). In addition, 123 publicly available oil reservoir metagenomes datasets were downloaded from NCBI Sequence Read Archive (SRA) database in Nov 2021 (Supplementary Data 1).

Collection of virus-like particles (VLPs) and extraction of viral DNA

To estimate whether the metagenome derived viral genome catalogue could cover the viral communities in oil reservoirs, we collected virus-like particles (VLPs) and sequenced viral DNA of additional 8 samples from oil reservoirs. Briefly, after obtaining filtrates from the water phase, the 300 ml filtrates were further used for the extraction of VLPs. To obtain VLPs, the filtrates were firstly filtered through tangential flow filtration equipment (TFF) with a 100 kDa tangential flow membrane package (Sartorius, VIVAFLOW 50 100,000 MWCO, Germany), and then continuously concentrated until reaching an ~1 mL volume using 100 kDa centrifugal filter units (Amicon® Ultra-15, Ultracel-100K, Millipore, Germany). 720 µL of concentrated solution was treated with 1000 U/mL DNase I (37 °C, 2 h) (Roche, China) before viral DNA extraction. Total viral DNA was extracted using a phenol-chloroform-isopentanol method. Then, extracted total viral DNA was used to construct sequencing libraries and sequenced. Firstly, extracted total viral DNA was amplified using REPLI-g Cell WGA & WTA Kit (Qiagen, Germany), the products were then used to construct sequencing libraries using the NEB Next® Ultra II™ DNA Library Prep Kit for Illumina® (New England Biolabs, USA) following the manufacturer's recommendations. In total, 74 million raw sequence reads were retained for analysis (Supplementary Data 1).

Processing of metagenomic sequence data and generation of prokaryotic metagenome-assembled genomes (MAGs)

The metagenomic raw reads were examined using FastQC v0.11.9 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), low-quality sequences, primers, and adaptors were trimmed using the Trimmomatic v0.39⁴⁹ (parameters: LEADING:2 TRAILING:2 SLIDINGWINDOW:4:20 MINLEN:50). The trimmed reads were independently assembled using MEGAHIT v1.2.9⁵⁰ (parameters: -presets metasensitive) and/or SPAdes v3.15.4⁵¹ (parameters: -meta, -k 21,33,55,77,99,127). For each assembly, contigs were binned using the binning module (parameters: -metabat2 -maxbin2 -concoct) and consolidated into a final bin set using the Bin_refinement module (parameters: -c 50 -x 10) within metaWRAP v1.2.1⁵². All produced bin sets were aggregated and de-replicated at 95% average nucleotide identity (ANI) using dRep v3.2.2⁵³ (parameters: -comp 50 -con 10 -nc

0.30 -pa 0.9 -sa 0.95). The completeness and contamination of MAGs were assessed using the lineage-specific module within CheckM v1.1.3⁵⁴ with default parameters, which generated 4521 MAGs that met or exceeded the medium-quality thresholds (>50% completeness and <10% contamination), including 2708 estimated to be near-complete (>90% completeness and <5% contamination).

The taxonomy of each MAG was assigned using GTDB-Tk v1.5.0⁵⁵ based on the Genome Taxonomy Database (GTDB, <http://gtdb.ecogenomic.org>) taxonomy (release202). The maximum-likelihood phylogenetic trees of MAGs were constructed based on a concatenated dataset of 400 universally conserved marker proteins using PhyloPhlAn v3.0.64⁵⁶ and visualized using iTOL v5⁵⁷. RPKM (Reads per kilobase per million mapped reads) values were used to represent the relative abundances of MAGs. To mitigate the impact of viral abundance on MAGs abundance, we initially removed virus contigs from MAGs before determining RPKM values, which were calculated using CoverM v0.6.1 (<https://github.com/wwood/CoverM>) (parameters: coverm genome -min-read-percent-identity 0.95, -min-read-aligned-percent 0.75, -contig-end-exclusion 0 and -m rpkm).

Functional annotation of MAGs and phylogenetic analysis of the DsrAB proteins

Open reading frames (ORFs) of these MAGs were predicted with Prodigal v2.6.3⁵⁸ (parameters: -m meta). The predicted ORFs were annotated using eggNOG-mapper v2.0.1⁵⁹ and the eggNOG Orthologous Groups database (version 5.0)⁶⁰ in diamond mode. Annotated KO numbers were used for inferring the pathway encoded in each MAG. MAGs that encoded a complete pathway for sulfur metabolism was utilized for further analysis. In addition, we utilized the CANT-HYD⁶¹ database by employing the -cut_nc noise cut-off to annotate genes involved in the hydrocarbon degradation (Supplementary Data 7).

For phylogenetic analysis of DsrAB sequences, DsrAB sequences from both the MAGs and previously reported reference DsrAB sequences^{11,62} were utilized for the phylogenetic analysis (Supplementary Fig. 7a, b, Supplementary Fig. 9), which could help to distinguish reductive and oxidative type DsrAB. The DsrAB sequences were aligned using MUSCLE v3.8⁶³ with default parameters. The alignments were then filtered using TrimAL v1.4⁶⁴ (parameters: -cons 50). The concatenated DsrAB tree was constructed using RAxML⁶⁵ (parameters: -f a -m PROTGAMMAJTT -p 12345 -x 12345 -N 100). The Newick files with the best tree topology were visualized using iTOL v5⁵⁷.

Viral contig identification, dereplication, virus operational taxonomic unit (vOTU) clustering, and calculation of abundances

Viral contigs were recovered from assembled contigs using VirSorter v2.1⁶⁶ and DeepVirFinder v1.0⁶⁷. Only viral contigs ≥10 kb were retained, based on the following criteria: (1) Viral contigs identified only by VirSorter v2.1⁶⁶ (parameters: -exclude-lt2gene); (2) Viral contigs identified only by DeepVirFinder v1.0⁶⁷ (parameters: score ≥ 0.9 and p < 0.05); (3) Viral contigs identified by both VirSorter v2.1⁶⁶ and DeepVirFinder v1.0⁶⁷. The identified viral contigs from each sample were clustered into virus operational taxonomic units (vOTUs) using the parameters 95% average nucleotide identity (ANI) and 85% alignment fraction of the smallest scaffolds based on the scripts (<https://bitbucket.org/berkeleylab/checkv/src/master/>) provided in CheckV v0.8.1⁶⁸. Next provirus boundaries were identified in the vOTUs and host contamination removed using CheckV v0.8.1⁶⁸. RPKM values were used to represent the relative abundances of vOTUs. The RPKM values of vOTUs were counted using CoverM v0.6.1 (parameters: coverm contig -min-read-percent-identity 0.95, -min-read-aligned-percent 0.75, -contig-end-exclusion 0 and -m rpkm). Viral lifestyle was predicted by both VIBRANT v1.2.1⁶⁹ and CheckV v0.8.1⁶⁸, while the remaining vOTUs with at least 90% completeness that display no prophage signals or lysogeny-specific genes were considered as

potential virulent viruses^{28,33,70,71}. In addition, to determine overlaps between oil reservoir vOTUs and viral contigs in the IMG/VR v3 dataset, we used rapid genome clustering to identify vOTUs that share 95% average nucleotide identity (ANI) and 85% alignment fraction of the smallest scaffolds with viral contigs in the IMG/VR v3 dataset based on the scripts provided in CheckV v0.8.1.

Viral taxonomic assignments, viral function annotation, and identification of auxiliary metabolic genes (vAMGs)

Open reading frames (ORFs) of vOTUs were predicted with Prodigal v2.6.3 (-p meta -g 11 -f gff). To understand the taxonomy of vOTUs, as suggested previously⁷², we used PhaGCN2.0⁷³ and geNomad v1.9⁷⁴ with the ICTV classification to explore the taxonomic affiliation of vOTUs at the family level. The results from these two tools were considered; for a given genome, (1) it was assigned as 'unclassified' if both tools failed to assign it, or it was assigned to different taxa, and (2) it was assigned to the taxonomic level determined by one of the tools if the other failed to assign. Detailed information of the vOTUs quality and taxonomy is provided in Supplementary Data 2. Additionally, to understand the function of vOTUs, the predicted viral proteins were first merged and dereplicated using CD-HIT v4.7⁷⁵ (parameters: -c 0.90 -s 0.8 -n 5 -g 1 -d 0). The dereplicated viral proteins were assigned to the eggNOG Orthologous Groups database (version 5.0) using eggNOG-mapper v2.0.1 (-m diamond) to identify COG functional classifications.

Moreover, in order to identify putative viral auxiliary metabolic genes (AMGs) that may have a role in host metabolism during the infection cycle, we used DRAM-v v1.3.5⁷⁶ to recover putative AMGs from all viruses (n = 6,218). Because DRAM-v requires VirSorter output, we re-ran 6,218 viruses through VirSorter v2.1. For DRAM-v output, as suggested previously^{28,77,78}, to be conservative, we manually scanned the annotation output to improve the confidence in AMG identification, in particular, only putative AMGs with an auxiliary score <4 were retained, and no viral flag (F), transposon flag (T), viral-like peptidase (P), or attachment flag (A) could be present, and putative AMGs that did not have a gene ID or a gene description were also discarded⁷⁷. In addition, putative AMGs predicted to be involved in organic nitrogen, nucleotide metabolism, and predicted to be glycosyl transferases or ribosomal proteins were removed from downstream consideration, because some viruses can encode their own glycosyl transferases⁷⁹. Moreover, to avoid false-positive results for selected AMGs caused by possible contamination from host sequences, we searched for the presence of viral hallmark genes or virus-like genes upstream and downstream of the putative AMGs, only putative AMGs located between or alongside two viral hallmark genes or virus-like genes were considered high-confidence viral AMGs and used for further analysis. Finally, genome maps for six viruses encoding AMGs were visualized based on DRAM-v and VirSorter2 annotations, protein structural homology searches were performed using the Phyre2 web portal⁸⁰ with a confidence threshold of = 98%, coverage threshold of >80%, and identity threshold of >30%.

Statistical analysis of viral community

All statistical analyses were performed in R version 4.0.0⁸¹. In all statistical analyses, no statistical method was used to predetermine sample size, no data were excluded from the analyses. The world map was generated by the function `map_data` ("world") in the "ggplot2" package⁸². All histogram plots were created using the `geom_histogram` function in the "ggplot2" package, all bar plots were created using the `geom_bar` function in the "ggplot2" package, and all pie plots were created using the `pie` function. Alpha and beta diversity of viral communities were calculated using "vegan" v2.5-7 package^{81,83}. Nonmetric multidimensional scaling (NMDS) was conducted based on Bray-Curtis dissimilarities generated from vOTUs tables with viral abundances

(RPKM) using the "vegdist" function. To further determine the significant difference of viral community composition between different continents, a similarity analysis (ANOSIM) was performed using the "anosim" function. In addition, to understand distance-decay relationships of viral community, pairwise geographic distances between samples were calculated from the latitude and longitude coordinates using the "geosphere" library, and the relationships between geographic distances and viral community similarities (1 - dissimilarity of the Bray-Curtis metric) were calculated by ordinary least-squares regressions.

To determine the contribution of different ecological processes to community assembly, null model analysis was carried out using the framework described by Stegen et al.⁸⁴. The null model expectation was generated using 999 randomizations in R. Two metrics, including β -nearest taxon index (β NTI) and Bray-Curtis-based Raup-Crick (RC_{Bray}), were calculated to divide community assembly into five processes, namely, homogeneous selection, variable selection, dispersal limitation, homogeneous dispersal, and drift. β NTI > 2 indicates heterogeneous selection, β NTI < -2 indicates homogeneous selection. $|\beta$ NTI| < 2 and RC_{Bray} < -0.95 indicate homogenizing dispersal, $|\beta$ NTI| < 2 and RC_{Bray} > 0.95 indicate dispersal limitation, $|\beta$ NTI| < 2 and RC_{Bray} < 0.95 indicate drift assembly⁸⁴. Heatmaps were created using the `geom_raster` function in the "ggplot2" package.

Moreover, we utilized the 59 newly sequenced metagenomes that contained sufficient environmental information data in this work to perform Mantel tests and multiple regression analysis using multiple regression on matrices (MRM) approach. Mantel tests with 9999 permutations within the "vegan" package to examine the correlations between viral community similarity and geographic distance or environmental distance (Spearman correlation). Because there was strong collinearity among particular environmental factors, before applying MRM, we used variable clustering to assess the redundancy of the environmental variables by the "varclus" function in the "Hmisc" package. The variables with higher correlation (Spearman's ρ > 0.5) were removed from the MRM analysis (including Cl⁻, Na⁺, Ca²⁺, depth), but kept all other variables in the models.

Network analysis

Protein-sharing network analysis of vOTUs was performed by vConTACT v2.0³⁰. Briefly, all vOTUs of the oil reservoirs were compared to vOTUs (≥ 10 kb) from other ecosystems in previously published data: (1) Wetland sediment (n = 1,075); (2) Stordalen thawing permafrost (n = 1,682)¹⁷; (3) Acid mine drainage sediments (n = 5,184)²⁹; (4) Cold seeps (n = 2,490)²⁸, (5) Minnesota peat (n = 3,566)⁷⁷; and from other ecosystems in IMG/VR v3⁸⁵: (6) Hydrothermal vents (n = 531); (7) Groundwater (n = 1,372); (8) Non-marine Saline and Alkaline (n = 1,756); (9) Thermal springs (n = 268). For each vOTU, ORFs were called using Prodigal v2.6.3³⁸, and the predicted protein sequences were used as input for vConTACT v2.0. The protein sequences of the vOTUs were grouped into protein clusters (PCs) using vConTACT v2.0 (parameters: -rel-mode Diamond -pcs-mode MCL -vcs-mode ClusterONE). The degree of similarity between the vOTUs was calculated based on the number of shared PCs. The networks were visualized by Cytoscape v3.5.1⁸⁶ (<http://cytoscape.org>) using an applied preferred layout model. Additionally, all vOTUs of the oil reservoirs were also compared to the NCBI Prokaryotic Viral RefSeq v201 database using vConTACT v2.0.

Virus-host prediction

Two different in silico methods were used to predict virus-host interactions. (1) tRNA match. ARAGORN v 1.2.38⁸⁷ was used to identify tRNAs from sequences of vOTUs (parameters: -t). Identified tRNAs were compared to metagenomic contigs using `fuzznuc`⁸⁸ from the EMBOSS:6.6.0.0 package with no mismatches allowed. (2) CRISPR spacer match. CRISPR spacers were recovered from metagenomic

contigs using metaCRT (modified from CRT1.2)^{29,89} with default parameters. Extracted spacers were compared to vOTUs using fuzznuc from the EMBOSS:6.6.0.0 package with no mismatches allowed⁹⁰. Detailed information of the virus-host interactions is provided in Supplementary Data 3.

Microcosm experiments setup

To further study the viral impact on sulfate reduction, microcosm experiments were performed. We used the production water samples collected from the Huabei Oilfields (Block B51-11, 38 °C, primary recovery) as seed banks, and constructed two types microcosms with different initial counts of virus-like particles (VLPs), namely low-VLPs microcosms, and high-VLPs microcosms. As shown in supplementary Fig. 12, approximately 500 mL of production water was filtered through sterile gauze to remove crude oil and obtain the fraction containing microbes and viruses. The filtrate was further divided into two parts, part of the filtrate was centrifuged (3500 g at 4 °C for 40 min) to obtain microbial pellets. Another part of the filtrate was filtered through a 0.22-μm-pore-size polycarbonate membrane (45 mm diameter; Millipore, Bedford, MA, United States) to obtain virus-rich filtrate. Afterward, part of the 0.22 μm filtrate was filtered through tangential flow filtration equipment (TFF) with a 100 kDa tangential flow membrane package (Sartorius, VIVAFLW 50 100,000 MWCO, Germany) to obtain virus-poor filtrate. At this point, the basic components required for the microcosms were acquired.

The two types microcosms were set up using the above-mentioned basic components as described below. In brief, all microcosms were set up with clean and sterilized serum bottles (internal volume 600 mL), containing 2 g sterilized crude oil, 30 mL autoclaved Brackish medium, and 300 mL virus-rich filtrate or virus-poor filtrate, amendment with this amount of crude oil (0.6% w/v) follows previous incubations that used a similar concentration⁹¹. Specifically, in high-VLPs microcosms, microbial pellets were diluted three times with virus-rich filtrate (Microbes+Viruses). Similarly, in low-VLPs microcosms, microbial pellets were diluted three times with virus-poor filtrate (Microbes–Viruses). As a result, the high-VLPs microcosms featured a higher initial count of VLPs in comparison to the low-VLPs microcosms. The Brackish medium contained: (g/L): NaCl, 7.0; MgCl₂·6H₂O, 1.20; CaCl₂·2H₂O, 0.10; NH₄Cl, 0.25; KH₂PO₄, 0.20; and KCl, 0.5. The pH of Brackish medium was adjusted to a range of 7.0 ~ 7.3 by addition of HCl (2 M) or NaOH solution (1 M). Then Brackish medium was autoclaved at 121 °C for 30 min. After cooling to room temperature, the following sterile solutions were added: 330 μL of trace elements stock solution, 990 μL of vitamin stock solution, 330 μL of selenite-tungstate solution, 9.9 mL of NaHCO₃ solution (84 g/L), and 330 μL of resazurin solution (1 g/L). The trace elements stock solution contained (g/L): FeSO₄·7H₂O, 2.10; H₃BO₃, 0.03; MnCl₂·4H₂O, 0.10; CoCl₂·6H₂O, 0.19; NiCl₂·6H₂O, 0.024; CuCl₂·2H₂O, 0.002; ZnSO₄·7H₂O, 0.14; Na₂MoO₄·2H₂O, 0.036; and 12.5 mL HCl (25%, v/v). The vitamins stock solution contained (g/L): p-aminobenzoic acid, 0.04; D (+)-biotin, 0.01; nicotinic acid, 0.1; Ca-D (+)-pantothenate, 0.05; pyridoxine-HCl, 0.15; vitamin B₁₂, 0.005; and thiamine-HCl, 0.10. The selenite-tungstate solution contained (g/L): NaOH, 0.4; Na₂SeO₃·5H₂O, 0.006; and Na₂WO₄·2H₂O, 0.008.

All microcosms were sealed with a butyl rubber stopper and an aluminum seal. To create anaerobic conditions, high-purity nitrogen gas (99.99%) was introduced into each microcosm using a sterile needle and continuously passed through for a duration of 1.5 h. Additionally, a small amount of Na₂S solution (48 g/L) were added to each microcosm until the color of the indicator in the bottle changed from pink to colorless. In addition, high-VLPs microcosms and low-VLPs microcosms were supplemented with sulfate to a final concentration of 4 g/L at the beginning of the culture. All treatments were set up in triplicate and were incubated at 30 °C in the dark. During the cultivation process, the concentration of sulfides was measured using

the total Sulfide Quantification Kit (LIANHUA, China). Detailed information of the sulfide production is provided in Supplementary Data 4.

DNA extraction, library construction, and sequencing of microcosms

After 160 days (T1) and 570 days (T2) of incubation, 330 mL of each culture were filtered through 0.22-μm-pore-size polycarbonate membranes (45 mm diameter; Millipore, Bedford, MA, United States) to obtain microbial cells. Total microbial DNA was extracted and sequencing as described previously. Metagenomes of all microcosms were analyzed as described before.

Statistical analyses of microcosm experiments

All statistical analyses were performed in R version 4.0.0^{81,83}. The difference analysis of H₂S between different microcosms was calculated using two-tailed *t*-test. The alpha diversity metrics, including richness, and Shannon (Shannon-Wiener diversity), were calculated using the ‘vegan’ and ‘picante’ packages. For beta diversity, the principal components analysis (PCoA) based on Bray-Curtis distance and a similarity analysis (ANOSIM) were performed in the ‘vegan’ package. The differential enrichment of KEGG modules related to energy metabolism and hydrocarbon degradation were identified according to their reporter scores and reporter features. The final set of differential enrichment KEGG modules was determined as the union of the KEGG modules based on reporter scores and reporter features. Reporter scores and reporter features were calculated as described previous^{92,93}. To understand the significant change of gene abundance between different microcosms, KEGG homologs were combined and summed according to gene clusters table of non-redundant gene sets. We then conducted difference analysis for KO using two-tailed Student’s *t*-test.

Data availability

Raw metagenomic sequencing data and sequences of metagenome-assembled genomes associated with the study have been deposited in the Genome Sequence Archive (<https://ngdc.cnca.ac.cn/gsa>) under BioProject PRJCA020971. The specific accession numbers for individual samples are listed in Supplementary Data 1. Sequences of 3,886 vOTUs can be found at figshare (<https://doi.org/10.6084/m9.figshare.24298258.v1>). Source data are provided with this paper.

Code availability

All original code has been deposited at GitHub and is publicly available as of the date of publication. Workflows and supporting code can be accessed at the following repository: <https://github.com/ananzhongheng/virus/tree/main/pipeline>.

References

- Sharma, S. et al. Geochemical controls on CO₂ interactions with deep subsurface shales: implications for geologic carbon sequestration. *Environ. Sci.* **23**, 1278–1300 (2021).
- Bastin, E. S., Greer, F. E., Merritt, C. A. & Moulton, G. The presence of sulphate reducing bacteria in oil field waters. *Science* **63**, 21–24 (1926).
- Zhao, J. Y. et al. Thermodynamically favorable reactions shape the archaeal community affecting bacterial community assembly in oil reservoirs. *Sci. total Environ.* **781**, 146506 (2021).
- Vigneron, A. et al. Succession in the petroleum reservoir microbiome through an oil field production lifecycle. *ISME J.* **11**, 2141–2154 (2017).
- Tyne, R. L. et al. Rapid microbial methanogenesis during CO₂ storage in hydrocarbon reservoirs. *Nature* **600**, 670–674 (2021).
- Jones, D. M. et al. Crude-oil biodegradation via methanogenesis in subsurface petroleum reservoirs. *Nature* **451**, 176–180 (2008).
- Liu, Y. F. et al. Metabolic capability and in situ activity of microorganisms in an oil reservoir. *Microbiome* **6**, 5 (2018).

8. Zhou, Z. et al. Non-syntrophic methanogenic hydrocarbon degradation by an archaeal species. *Nature* **601**, 257–262 (2022).
9. Lahme, S., Mand, J., Longwell, J., Smith, R. & Enning, D. Severe Corrosion of Carbon Steel in Oil Field Produced Water Can Be Linked to Methanogenic Archaea Containing a Special Type of [NiFe] Hydrogenase. *Appl. Environ. Microbiol.* **87**, e01819–e01820 (2021).
10. Head, I. M., Jones, D. M. & Larter, S. R. Biological activity in the deep subsurface and the origin of heavy oil. *Nature* **426**, 344–352 (2003).
11. Zhao, F. et al. Simultaneous inhibition of sulfate-reducing bacteria, removal of H₂S and production of rhamnolipid by recombinant *Pseudomonas stutzeri* Rhl: Applications for microbial enhanced oil recovery. *Bioresour. Technol.* **207**, 24–30 (2016).
12. Li, J. et al. Effect of exogenous inoculants on enhancing oil recovery and indigenous bacterial community dynamics in long-term field pilot of low permeability reservoir. *World J. Microbiol. Biotechnol.* **34**, 53 (2018).
13. Yun, Y. et al. Disentangling the distinct mechanisms shaping the subsurface oil reservoir bacterial and archaeal communities across northern China. *Sci. total Environ.* **789**, 148074 (2021).
14. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
15. Suttle, C. A. Viruses in the sea. *Nature* **437**, 356–361 (2005).
16. Adriaenssens, E. M. et al. Environmental drivers of viral community composition in Antarctic soils identified by viromics. *Microbiome* **5**, 83 (2017).
17. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
18. Bellas, C. M., Schroeder, D. C., Edwards, A., Barker, G. & Anesio, A. M. Flexible genes establish widespread bacteriophage pan-genomes in cryoconite hole ecosystems. *Nat. Commun.* **11**, 4403 (2020).
19. Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–159 (2015).
20. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).
21. Zheng, L. et al. Viral abundance and diversity of production fluids in oil reservoirs. *Microorganisms* **8**, 1429 (2020).
22. Cai, M. et al. Crude oil as a microbial seed bank with unexpected functional potentials. *Sci. Rep.* **5**, 16057 (2015).
23. Daly, R. A. et al. Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **4**, 352–361 (2019).
24. Meckenstock, R. U. et al. Oil biodegradation. Water droplets in oil are microhabitats for microbial life. *Science* **345**, 673–676 (2014).
25. Nishijima, S. et al. Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. *Nat. Commun.* **13**, 5252 (2022).
26. Zhao, J. et al. Significant differences in planktonic virus communities between “cellular fraction” (0.22–3.0 µm) and “viral fraction” (<0.22 µm) in the ocean. *Microb. Ecol.* **86**, 825–842 (2023).
27. Roux, S. et al. Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
28. Li, Z. et al. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME J.* **15**, 2366–2378 (2021).
29. Gao, S. et al. Patterns and ecological drivers of viral communities in acid mine drainage sediments across Southern China. *Nat. Commun.* **13**, 2389 (2022).
30. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
31. Shelton, A. N. et al. Uneven distribution of cobamide biosynthesis and dependence in bacteria predicted by comparative genomics. *ISME J.* **13**, 789–804 (2019).
32. Roth, J. R., Lawrence, J. G. & Bobik, T. A. Cobalamin (coenzyme B₁₂): synthesis and biological significance. *Annu. Rev. Microbiol.* **50**, 137–181 (1996).
33. Ma, B. et al. Biogeographic patterns and drivers of soil viromes. *Nat. Ecol. Evol.* **8**, 717–728 (2024).
34. Knowles, B. et al. Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).
35. Wacey, D., Kilburn, M. R., Saunders, M., Cliff, J. & Brasier, M. D. Microfossils of sulphur-metabolizing cells in 3.4-billion-year-old rocks of Western Australia. *Nat. Geosci.* **4**, 698–702 (2011).
36. Fike, D. A., Bradley, A. S. & Rose, C. V. Rethinking the ancient sulfur cycle. *Annu. Rev. Earth Planet. Sci.* **43**, 593–622 (2015).
37. Ma, H. et al. The influence of hydrogen sulfide on corrosion of iron under different conditions. *Corros. Sci.* **42**, 1669–1683 (2000).
38. Kieft, K. et al. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nat. Commun.* **12**, 3503 (2021).
39. Thurber, R. V., Payet, J. P., Thurber, A. R. & Correa, A. M. Virus-host interactions and their roles in coral reef health and disease. *Nat. Rev. Microbiol.* **15**, 205–216 (2017).
40. Liao, H. et al. Response of soil viral communities to land use changes. *Nat. Commun.* **13**, 6027 (2022).
41. Peng, Y. et al. Viruses in deep-sea cold seep sediments harbor diverse survival mechanisms and remain genetically conserved within species. *ISME J.* **17**, 1774–1784 (2023).
42. Cheng, R. et al. Virus diversity and interactions with hosts in deep-sea hydrothermal vents. *Microbiome* **10**, 235 (2022).
43. Zeng, S. et al. A metagenomic catalog of the early-life human gut virome. *Nat. Commun.* **15**, 1864 (2024).
44. Silveira, C. B. & Rohwer, F. L. Piggyback-the-Winner in host-associated microbial communities. *NPJ Biofilms Microbiomes* **2**, 16010 (2016).
45. Chen, X., Weinbauer, M. G., Jiao, N. & Zhang, R. Revisiting marine lytic and lysogenic virus-host interactions: Kill-the-Winner and Piggyback-the-Winner. *Sci. Bull.* **66**, 871–874 (2021).
46. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
47. Kim, M. S. & Bae, J. W. Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J.* **12**, 1127–1141 (2018).
48. Johansen, J. et al. Genome binning of viral entities from bulk metagenomics data. *Nat. Commun.* **13**, 965 (2022).
49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
50. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
51. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
52. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP: a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
53. 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).
54. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes

- recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
55. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
 56. Asnicar, F. et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
 57. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
 58. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
 59. Huerta-Cepas, J. et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
 60. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
 61. Khot, V. et al. CANT-HYD: A Curated Database of Phylogeny-Derived Hidden Markov Models for Annotation of Marker Genes Involved in Hydrocarbon Degradation. *Front. Microbiol.* **12**, 764058 (2021).
 62. Anantharaman, K. et al. Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME J.* **12**, 1715–1728 (2018).
 63. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
 64. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
 65. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
 66. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
 67. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **8**, 64–77 (2020).
 68. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
 69. 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).
 70. Gao, S.-M. et al. Eco-evolutionary dynamics of gut phageome in wild gibbons (*Hoolock tianxing*) with seasonal diet variations. *Nat. Commun.* **15**, 1254 (2024).
 71. Luo, X. Q. et al. Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome* **10**, 190 (2022).
 72. Chen, L. & Banfield, J. F. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. *Nat. Microbiol.* **9**, 737–750 (2024).
 73. Jiang, J. Z. et al. Virus classification for viral genomic fragments using PhaGCN2. *Brief. Bioinforma.* **24**, bbac505 (2023).
 74. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* (2023).
 75. 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).
 76. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).
 77. Ter Horst, A. M. et al. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* **9**, 233 (2021).
 78. Zhao, J. et al. Novel viral communities potentially assisting in carbon, nitrogen, and sulfur metabolism in the upper slope sediments of mariana trench. *mSystems* **7**, e0135821 (2022).
 79. Xiang, Y. et al. Crystal structure of a virus-encoded putative glycosyltransferase. *J. Virol.* **84**, 12265–12273 (2010).
 80. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
 81. Ihaka, R. & Gentleman, R. R. A Language for Data Analysis and Graphics. *J. Computational Graph. Stat.* **5**, 299–314 (1996).
 82. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
 83. Oksanen, J. B. F., Kindt R., Legendre P., Minchin P. R., O'Hara R.,. Vegan: community ecology package. R package version 22-1. (2015).
 84. Stegen, J. C. et al. Quantifying community assembly processes and identifying features that impose them. *ISME J.* **7**, 2069–2079 (2013).
 85. Roux, S. et al. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764–D775 (2021).
 86. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
 87. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
 88. Rice, P., Longden, I. & Bleasby, A. EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
 89. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinforma.* **8**, 209 (2007).
 90. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
 91. Mayumi, D. et al. Carbon dioxide concentration dictates alternative methanogenic pathways in oil reservoirs. *Nat. Commun.* **4**, 1998 (2013).
 92. Feng, Q. et al. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
 93. Oliveira, A. P., Patil, K. R. & Nielsen, J. Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. *BMC Syst. Biol.* **2**, 17 (2008).

Acknowledgements

We thank the members of the Wu lab at Peking University for the critical discussions of this work. In addition, we thank the High-performance Computing Platform of Peking University for providing the computing platform. This study has received funding from the National Key R&D Program of China (nos. 2018YFA0902100 to X.L.W., and nos. 2021YFA0910300 to Y.N.) and the National Natural Science Foundation of China (nos. 32130004 to X.L.W., nos. 32161133023 to X.L.W., nos. 91951204 to Y.N., and nos. 32170113 to Y.N.).

Author contributions

All authors contributed intellectual input and assistance to this study. L.Y.A. conceived the study and performed the original analysis. X.W.L. added microcosm enrichment experiments validation. J.W.W., and J.B.X. collected raw data. L.Y.A. and Y.N. co-wrote the paper. X.L.C., X.N.L., and B.X.H. revised it. Y.N. and X.L.W. raised the funding for the project. All authors discussed the results and commented on the article.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-51101-6>.

Correspondence and requests for materials should be addressed to Yong Nie or Xiao-Lei Wu.

Peer review information *Nature Communications* thanks Michael Tisza, and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions 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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024