nature communications

Article

https://doi.org/10.1038/s41467-024-51230-y

Unveiling the unknown viral world in groundwater

Received: 4 December 2023

Accepted: 1 August 2024

Published online: 08 August 2024

Check for updates

Zongzhi Wu[1,2,6], Tang Liu[1,3,6], Qian Chen[1,2], Tianyi Chen[1], Jinyun Hu[2], Liyu Sun[1], Bingxue Wang[1], Wenpeng Li[4] & Jinren Ni[1,5]

Viruses as the prevailing biological entities are poorly understood in underground realms. Here, we establish the first metagenomic Groundwater Virome Catalogue (GWVC) comprising 280,420 viral species (≥ 5 kb) detected from 607 monitored wells in seven geo-environmental zones throughout China. In expanding ~10-fold the global portfolio of known groundwater viruses, we uncover over 99% novel viruses and about 95% novel viral clusters. By linking viruses to hosts from 119 prokaryotic phyla, we double the number of microbial phyla known to be virus-infected in groundwater. As keystone ultrasmall symbionts in aquifers, CPR bacteria and DPANN archaea are susceptible to virulent viruses. Certain complete CPR viruses even likely infect non-CPR bacteria, while partial CPR/DPANN viruses harbor cell-surface modification genes that assist symbiont cell adhesion to free-living microbes. This study reveals the unknown viral world and auxiliary metabolism associated with methane, nitrogen, sulfur, and phosphorus cycling in groundwater, and highlights the importance of subsurface virosphere in viral ecology.

Viruses, the most abundant entities on earth, have profound impacts on all organisms and ecosystems[1–3]. There are ~ 2×10^{29} prokaryotic cells living in groundwater, which represent a major component of genetic diversity and manipulate biogeochemical and ecological processes[4,5]. In parallel with in-depth studies of groundwater microorganisms[5–7], some evidence suggests that viruses are also involved in the biogeochemical cycle of the groundwater ecosystem[8,9]. The presence of viruses in aquifers has been previously confirmed, with variable morphology and high abundance (10-fold more than prokaryotes)[10]. Increasing numbers of lytic bacteriophages infecting abundant hosts (e.g., *Pseudomonas*, *Bacillus*, and *Desulfovibrio*) have been identified in phreatic water and deep groundwater[11,12], implying that subsurface environments might be underexplored biotopes in the global virosphere[13].

Although viruses play important roles in host evolution, microbial metabolism, and ecological processes[14,15], only a few viruses could be identified by culture-dependent methods. With the development of next-generation sequencing technology, an increasing number of viral sequences has been identified from meta-omics data[16–19], further deepening our understanding of the virome in different habitats such as oceans[17,20], soil[21,22], human gut[19,23], and wastewater treatment plants[24]. Recent studies of viral diversity and host interaction based on meta-omics data have helped overcome difficulties encountered in capturing groundwater viruses through limited amounts of culture[9,25–27]. Viruses in groundwater displayed massive novelty different from previously known viruses[9,26]. Previously reported virus-host relationship in groundwater, e.g., viruses targeting Altiarchaeota and Firmicutes, has provided an ideal model for viral lifestyle and infection mechanism of some specific taxa[25,27]. However, the great diversity of groundwater prokaryotes (spanning over one hundred phyla) and antiviral systems (such as CRISPR-Cas systems and Restriction-Modification)

[1]Eco-environment and Resource Efficiency Research Laboratory, School of Environment and Energy, Peking University Shenzhen Graduate School, Shenzhen 518055, PR China. [2]Environmental Microbiome and Innovative Genomics Laboratory, College of Environmental Sciences and Engineering, Peking University, Beijing 100871, PR China. [3]College of Chemistry and Environmental Engineering, Shenzhen University, Shenzhen 518060, PR China. [4]Center for Groundwater Monitoring, China Institute of Geo-environmental Monitoring, Beijing 100081, PR China. [5]College of Environmental Sciences and Engineering, Key Laboratory of Water and Sediment Sciences, Ministry of Education, Peking University, Beijing 100871, PR China. [6]These authors contributed equally: Zongzhi Wu, Tang Liu. e-mail: jinrenni@pku.edu.cn

suggest that many unknown virus-host relationships exist but are yet to be identified[5,28]. Besides, the limited information about viral auxiliary metabolic genes (AMGs) in groundwater may have hindered understanding of viral impacts on underground biogeochemical processes[9,26], while frequent horizontal gene transfer and broad accessible host range also imply the urgent necessity for new explorations of viral AMGs involved in carbon, nitrogen, sulfur, and phosphorus metabolisms in groundwater[1,2,9].

More importantly, groundwater ecosystems with typically anoxic or anaerobic environments provide ideal habitats for two keystone taxa, i.e., the candidate phyla radiation (CPR) bacteria and DPANN (an acronym of the names of the five initially found lineages Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanohaloarchaeota, and Nanoarchaeota) archaea[7,29]. CPR bacteria and DPANN archaea, as two remarkable groups in prokaryotic tree of life, usually share conserved traits such as ultrasmall cell size and extremely reduced genome[30,31]. Notably, CPR and DPANN microorganisms lack complete biosynthetic pathways for the synthesis of amino acids and nucleotides[30]. Thus, they usually live as symbionts of other free-living prokaryotes to obtain essential biomolecules[31], and small cells of these symbionts can attach to larger cells of other microbes via diverse cell-surface modifications (e.g., pili, glycosyltransferase, concanavalin, and LamG protein)[30,32]. Moreover, antiviral systems such as Restriction-Modification and CRISPR-Cas were also found in CPR bacteria and DPANN archaea[28,33–35], implying a rich virome infecting these symbionts. Although a recent study reported some aquifer viruses targeted by the CRISPR spacer of Altiarchaeota[27], CPR/DPANN viruses in groundwater ecosystems remain largely unknown, especially in terms of diversity, lifestyle, functional potential, and their roles in the microbial symbiosis.

Here, we aim to explore the enigmatic groundwater virosphere, including viral diversity, virus-host interaction, and AMGs related to biogeochemical cycling. To this end, we leverage ultra-deep metagenomic sequencing (over 30 giga bases per sample) of groundwater microbiome to establish a comprehensive non-redundant Groundwater Virome Catalogue (GWVC) consisting of 280,420 viral operational taxonomic units (vOTUs) at species level. This represents a ~10-fold expansion in the number of existing species of groundwater viruses derived from publicly available viral databases. Importantly, we unveil over 99% novel viruses and about 95% novel viral clusters in groundwater by comparing the GWVC with previously known viruses. The unique viral infection mode of prokaryotes suggests that microbial symbionts represented by keystone taxa like CPR bacteria and DPANN archaea are more susceptible to viral lysis in groundwater. Moreover, diverse AMGs related to methane, nitrogen, sulfur, and phosphorous cycles imply the important role of groundwater viruses in host metabolism and biogeochemical cycling. This study sheds light on the unknown viral world in groundwater, and emphasizes the fundamental importance of subsurface virosphere in future explorations of viral ecology.

Results and Discussion

The GWVC substantially expands groundwater virosphere

By mining metagenomic data (20.8 tera bases) of 607 samples from monitoring wells in seven geo-environment zones across China (Fig. 1a, Supplementary Data 1, and Methods), we constructed the Groundwater Virome Catalogue (GWVC). Four virus identification approaches were applied along with quality control (Methods), and 312,741 viral contigs (≥5 kb) were identified and then clustered at 95% average nucleotide identity (ANI). The generated GWVC consisted of 280,420 non-redundant viral contigs (≥5 kb), representing approximately species-level vOTUs (Fig. 1a). The completeness level of vOTUs in the GWVC varied from short fragments to complete or nearly complete genomes, including 5366 complete, 6092 high-quality, and 15,669 medium-quality genomes (Fig. 1b). Viral genome size of the GWVC ranged from 5 kb to 543.1 kb, and a sum of 107,610 vOTUs possess a length of ≥10 kb. Complete genomes had the largest mean size (49.0 kb), followed by high-quality (41.2 kb), medium-quality (36.1 kb), low-quality (11.0 kb), and not-determined (8.3 kb). Among 14,578 complete or high-quality genomes from the GWVC in the present study and the IMG/VR (groundwater section) (Fig. S1), the GWVC contributed more than 78.6% (11,458 genomes) of uncultivated viruses.

To explore the novelty of GWVC, viral contigs (≥10 kb) and their proteins from the GWVC and the IMG/VR (groundwater, marine, human, surface freshwater, terrestrial, and wastewater) were extracted for comparison analysis (Methods). The GWVC at vOTUs and protein clusters (PCs) level expanded the number of known groundwater viral species 10-fold and PCs 8-fold (Fig. 1c). In the overlapping fraction between the GWVC and the IMG/VR, the number of vOTUs/PCs related to aquatic ecosystems (vOTUs: $n = 156$, PCs: $n = 277{,}529$) were much higher than those related to human systems (vOTUs: $n = 92$, PCs: $n = 22{,}316$) and terrestrial ecosystems (vOTUs: $n = 12$, PCs: $n = 112{,}426$) (Fig. S2). Remarkably, the vast majority of vOTUs/PCs (vOTUs: 99.8%, PCs: 86.3%) were unique to the GWVC (Fig. 1d), indicating the great potential of aquifers to act as large reservoirs of unknown viruses.

To date, the differences in core features of viral genomes in groundwater and surface water remain unclear, though characteristics of prokaryotic genomes have been found to be strongly driven by environmental selection[36]. The comparison of complete viruses in groundwater (GWVC) and surface water (surface freshwater and ocean sections of IMG/VR) indicated that groundwater viruses possess unique genomic, protein, and functional traits (Fig. S3 and Methods). Groundwater viruses are characterized by larger genome size and higher GC content but lower average molecular weight (Fig. S3a, b, c). Amino-acid biosynthetic cost minimization in microorganisms is regarded as a necessarily adaptive strategy under resource limitation in natural environments[37]. Similarly, groundwater viruses in resource-limited groundwater might preferentially use amino acids with lower molecular weights to reduce assimilation costs[37]. Moreover, groundwater viral proteins appear to possess higher nitrogen atoms per residue side chain (N-ARSC) and sulfur atoms per residue side chain (S-ARSC) but lower carbon atoms per residue side chain (C-ARSC) (Fig. S3d, e, f). The elemental composition of microbial proteins is also highly related to resource availability in environments[36,37]. Accordingly, higher N-ARSC and S-ARSC of viral proteins might reflect higher availabilities of nitrogen and sulfur in groundwater, while reducing the carbon usage of viral proteins might contribute to viral survival in carbon-limited conditions. Functionally, the ORFs related to L (replication, recombination, and repair) and S (unknown function) are richer in groundwater viruses (Fig. S3g). Hence, groundwater-specific viral adaptation enables virion reproduction to be maximized under nutrient-limited conditions in groundwater.

Overall, 97.0% of vOTUs (≥10 kb) in the GWVC have the taxonomy assignment. Among these, at class level, the vast majority (95.8%) of vOTUs were assigned to the class Caudoviricetes within the realm Duplodnaviria (Fig. 1e), which contained head-tailed viruses that are common in most natural environments and human hosts[19,38]. In fact, over 94.0% of vOTUs (≥10 kb) in the GWVC could not be taxonomically annotated at order- or family-level (Fig. 1e). We also constructed the Caudoviricetes phylogenetic tree of GWVC vOTUs over high-quality along with NCBI RefSeq viruses based on a concatenated alignment of 77 marker proteins (Fig. S4). Many GWVC viruses appeared as independent clades that were close to certain known families (e.g., Drexlerviridae, and Autographiviridae) and recently proposed families (Casjensviridae, Mesyanzhinovviridae, Zobellviridae, Peduoviridae, and Steigviridae), among which Zobellviridae and Steigviridae (crAssphages) are thought to be widely distributed in global marine[39] and human guts[19], respectively. Many GWVC viruses formed branches with subfamily-level viruses (e.g., Azeredovirinae, Bronfenbrennervirinae, and Tybeckvirinae), greatly expanding the taxonomic diversity of Caudoviricetes. Intriguingly, some vOTUs (3.3%) were classified as

https://doi.org/10.1038/s41467-024-51230-y

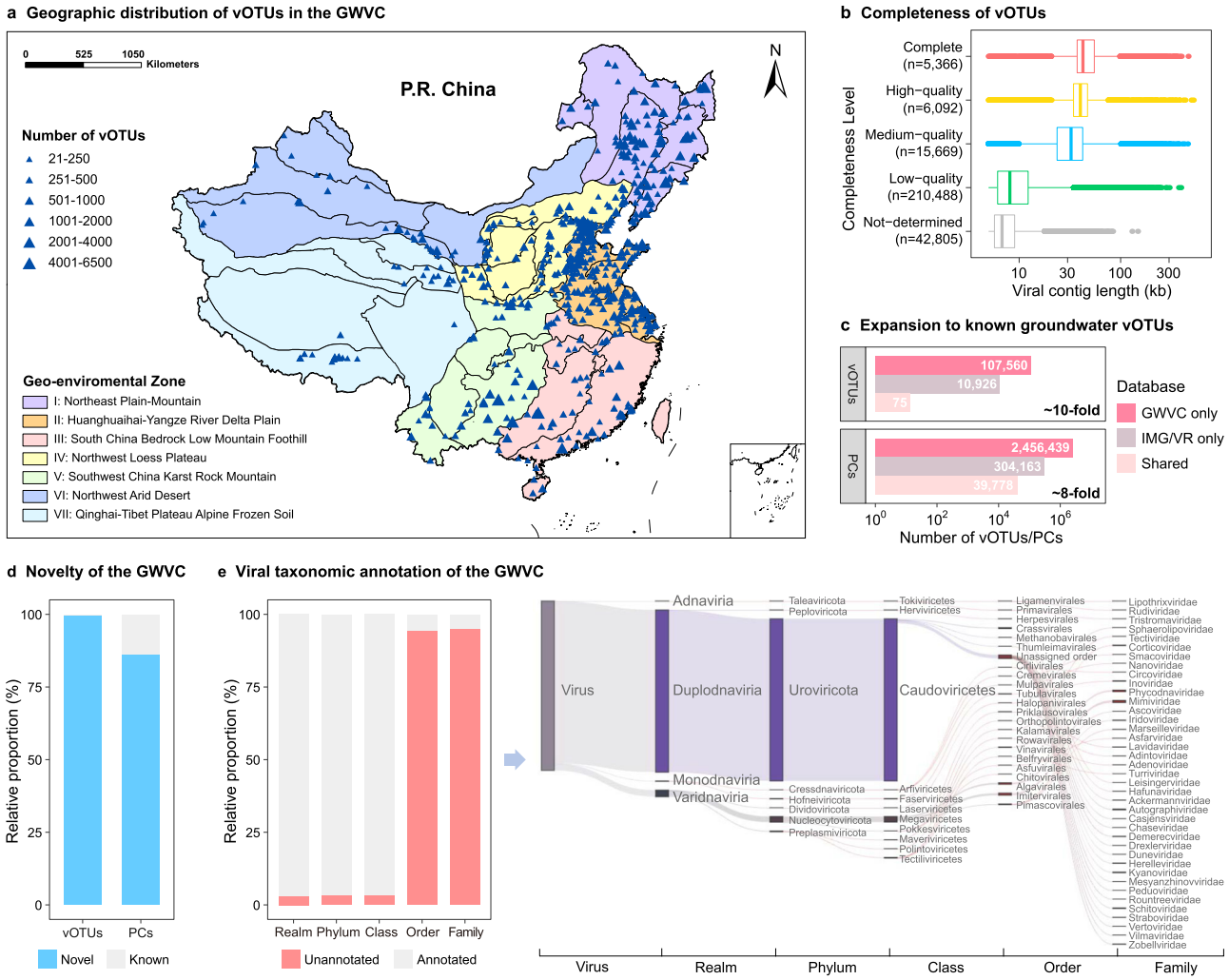

Fig. 1 | Compendium of the Groundwater Virome Catalogue (GWVC). **a** Geographic distribution of vOTUs in the GWVC. Triangle size denotes the number of vOTUs identified at each sampling site. The seven geo-environmental zones were distinguished with different colors. **b** Completeness and length of vOTUs (≥5 kb) in the GWVC. For each boxplot, central line and whiskers indicate the median and 1.5 times the interquartile range. The upper and lower sides of boxes represent the interquartile range between 25th and 75th percentile. Points beyond whiskers are potential outliers. **c** Number of vOTUs (≥10 kb) and their PCs in the GWVC, the IMG/VR (groundwater section), and shared by the two databases. **d** Novelty of vOTUs (≥10 kb) and their PCs in the GWVC based on comparison to the whole IMG/VR. Blue and gray bars represent the proportion of novel and known vOTUs/PCs, respectively. **e** Taxonomic annotation of vOTUs (≥10 kb) in the GWVC. Bar plot shows the proportion of unannotated (red) and annotated (grey) vOTUs/PCs. Sankey plot shows viral taxonomic annotation.

Nucleocytoplasmic Large DNA viruses (NCLDVs) which constitute the phylum Nucleocytoviricota. NCLDVs have complex genomes and large virions similar in size to small cellular organisms, and are usually abundant in marine, terrestrial and wastewaters[18]. Moreover, NCLDVs infect a wide range of eukaryotes, such as protists and algae[18], implying possible impact on groundwater eukaryotes.

Novel viruses occur in varying geo-environmental zones

We used a gene-sharing network approach to cluster viral genomes over medium-quality from the GWVC and the RefSeq database, resulting in 2830 non-singleton viral clusters that contained at least one vOTU from the GWVC (Fig. S5a, Supplementary Data 2, and Methods). Only 5.4% (152 out of 2830) of the GWVC viral clusters overlapped with the RefSeq viruses, suggesting that a larger number of novel viral genera existed in the groundwater samples.

The prevalence and abundance of viral clusters across all monitoring wells were calculated based on read recruitment of viral cluster members. As a result, viruses were found widely distributed at continental scale with a great proportion (79.2%, $n = 2240$) of virus clusters appearing in different geo-environmental zones (Fig. S6a). In the shared virus clusters, 54 viral clusters were prevalent in all geo-environmental zones, with 50 forming new clades on the viral proteomic tree, distinct from known viral families (Fig. S5b). Among these, 11 dominant viral clusters occurred in more than 20% of groundwater monitoring wells and possessed high relative abundance (RPKM ≥ 0.5%) (Fig. S6b, c). A total of 10 novel viral clusters unrelated to known viruses were found in these dominant viral clusters, suggesting many more unknown viral groups could be identified from the GWVC than from existing databases (Fig. S6c). Similar to the results of taxonomic annotations (Fig. 1e), the most dominant viral clusters ($n = 10$) were affiliated with the class Caudoviricetes. Notably, one dominant viral cluster belonged to the family Inoviridae within the realm Monodnaviria. Members of Inoviridae are a large group of viruses evolutionarily and structurally unrelated to Caudoviricetes[40], and possess a single-stranded DNA genome and filamentous virion[41]. Inoviridae are able to establish the chronic infection that release virions without killing the host[41]. Considering the high prevalence of this viral cluster of Inoviridae in aquifers and their unique properties of host interaction[41], they might play an important ecological role in groundwater microbial community.

https://doi.org/10.1038/s41467-024-51230-y

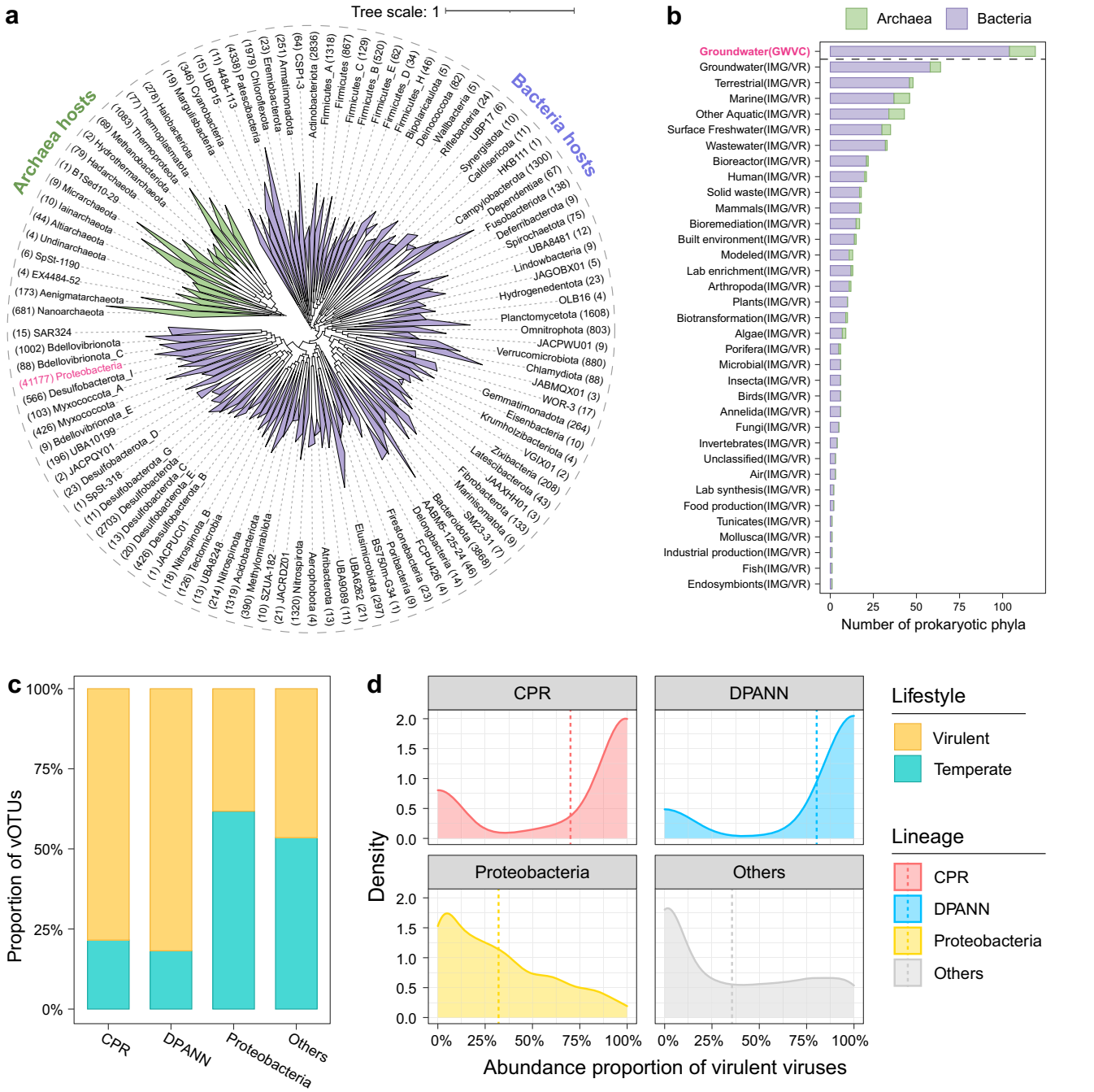

Fig. 2 | Host assignment of viruses in the GWVC. a Phylogeny of bacterial (purple) and archaeal (green) hosts. Number of vOTUs linked to each phylum is indicated by the number next to the host name. Proteobacteria phylum associated with most viruses is highlighted in red. **b** Number of host phyla infected by viruses from the GWVC and IMG/VR. **c** Lifestyle (virulent or temperate) proportions of complete or high-quality viruses infecting CPR bacteria, DPANN archaea, Proteobacteria, and other prokaryotes. Yellow and cyan bars represent the proportion of virulent and temperate vOTUs, respectively. **d** Density plots showing abundance proportion of virulent viruses infecting distinct hosts among samples where virulent or temperate viruses were detected. Dashed lines represent average abundance proportion of virulent viruses.

Viral infection spans an extremely broad spectrum of prokaryotes

To investigate virus-host interaction in groundwater, 34,993 prokaryotic metagenome-assembled genomes (MAGs) with completeness >70% and contamination <10% were reconstructed from 607 metagenomes in this study (Methods). We used four computational approaches to identify 193,952 virus-host connections, resulting in 71,600 vOTUs (25.5%) linked to 21,634 prokaryotic MAGs reconstructed from groundwater metagenomes in this study (Supplementary Data 3 and Supplementary Data 4). At phylum level, viruses were predicted to infect 104 bacteria phyla and 15 archaea phyla (Fig. 2a and Supplementary Data 4). The number of host phyla infected by viruses in groundwater ecosystems was more than twice than that in other ecosystems (e.g., terrestrial, marine, surface freshwater, humans, and wastewaters) (Fig. 2b), implying that groundwater is an underexplored hotspot for virus-host interaction. We found that most vOTUs ($n = 41{,}177$) were linked to Proteobacteria dominant in groundwater microbiome, followed by Patescibacteria (CPR bacteria, $n = 4338$),

Bacteroidota ($n = 3868$), Actinobacteriota ($n = 2836$), and Desulfobacterota ($n = 2703$). Archaea were also predicted to act as hosts for 2510 vOTUs, including viruses of Thermoproteota ($n = 1083$), Nanoarchaeota ($n = 681$), Halobacteriota ($n = 278$), Aenigmatarchaeota ($n = 173$) and Hadarchaeota ($n = 79$). Our results revealed that a total of 4338 and 932 vOTUs were linked to 20 CPR lineages (class-level) and 9 DPANN lineages (phyla-level), respectively (Fig. S7). Among these CPR/DPANN lineages, Paceibacteria (CPR) and Nanoarchaeota (DPANN) were important hosts for groundwater viruses (Fig. S7), unlike Saccharimonadia (CPR) and Altiarchaeota (DPANN) which act as the main hosts for viruses in the digestive tract of mammals[42,43] and the deep terrestrial subsurface[27], respectively. To explore potential viral roles in microbe-mediated biogeochemical cycling, we annotated the functional potentials of host MAGs in methane, nitrogen, and sulfur metabolisms (Methods). We found that numerous viruses ($n = 49,184$, 68.7% of host-linked vOTUs) were linked to prokaryotic hosts involved in methane, nitrogen, and sulfur metabolisms, suggesting potential effects of viral predation on microbial-mediated biogeochemical cycles. The virus-host connections suggest that almost all microbial metabolic processes involved in the canonical methane, nitrogen, and sulfur cycles in groundwater environments[5,7] might be impacted by viral infection (Fig. S8), especially (1) bacterial methane oxidation, and archaeal methanogenesis; (2) bacterial/archaeal dissimilatory nitrate reduction, denitrification, and nitrogen fixation; and (3) bacterial/archaeal dissimilatory sulfate reduction, sulfate disproportionation, and assimilatory sulfate reduction. On the one hand, viruses are able to reprogramme the host cell during infection[1,2], and thus alter the metabolism of key biogeochemical cycling contributors[1]. On the other hand, viral predation can mediate the turnover of abundant biogeochemical cycling microbes[2], and strengthen element cycling in groundwater via viral shunt[1]. In the future, the integration viral impacts into biogeochemical models might help to better predict the element cycling in groundwater.

To investigate potential impacts of viruses on microbial ecology in groundwater, the lineage-specific viral infection dynamics was assessed based on the virus-host abundance pattern. The composition of prokaryotic viruses was found to be highly coupled with their hosts (Fig. S9a), confirmed by the significant Spearman correlation between virus and host abundance ($p < 10^{-5}$, $R = 0.90$) (Fig. S9b). Lineage-specific virus-host abundance ratios revealed a range of virus-host abundance ratios among distinct taxa (Fig. S9c). For almost all lineages, including CPR/DPANN microbes, viral abundances often exceeded host abundances, indicating that microbial symbionts might undergo active viral proliferation like free-living microorganisms. Furthermore, the relationship between viral lifestyle and host in groundwater was revealed through predicting virulent/temperate viruses infecting various hosts (Methods). The proportion of virulent lifestyle was 3.6 and 4.5 times that of temperate lifestyle in CPR and DPANN viruses, respectively, in contrast to 0.62 and 0.87 times in Proteobacteria and other host viruses (Fig. 2c). Among samples where virulent or temperate viruses were detected, viruses linked to CPR/DPANN possessed higher abundance proportion of virulent lifestyle than viruses linked to Proteobacteria and other microorganisms (Fig. 2d), implying that the former prefer a virulent lifestyle but the latter are subject to a temperate lifestyle. Viruses infecting CPR/DPANN symbionts were predominantly virulent viruses, which might kill their host cells by lysis[2] and thereby drive the turnover of CPR/DPANN communities and nutrient cycling in groundwater[1,2]. By contrast, viruses infecting free-living microorganisms (e.g., Proteobacteria) were mainly temperate viruses that can exploit their hosts through lysogeny rather than killing them unless induction events were triggered[2].

CPR/DPANN viruses regulate microbial symbiotic associations in aquifers

We investigated CPR/DPANN virome and their potential impact on the symbiotic relationship in aquifers through constructing a comprehensive dataset of groundwater CPR/DPANN viruses (Methods). A total of 230 CPR viruses and 23 DPANN viruses were identified using CRISPR- and provirus- based methods (Supplementary Data 5), and over 25% of these associations can also be found using other host prediction methods (i.e., nucleotide sequence homology or k-mer frequency match) (Supplementary Data 6). Evidences were found for the association between viruses and CPR CRISPR-Cas systems (Fig. S10a, b, c), though CRISPR-Cas systems in CPR bacteria were not as prevalent as in other prokaryotic lineages[28]. For example, proto-spacers of complete or high-quality viral genomes were matched to spacers in complete CRISPR-Cas systems of CPR lineages (e.g., Paceibacteria, Gracilibacteria, and Dojkabacteria). For CPR-prophage association, clear viral genome integration was found in CPR lineages (e.g., Paceibacteria, ABY1, and Gracilibacteria) based on prophage prediction (Fig. S10d, e, f). The complete CPR viral genomes ($n = 7$) are 36–55 kb in length (average 39 kb), and the two complete DPANN viral genomes were both 41 kb in length. This suggested that CPR/DPANN viruses possessed relatively smaller genomes than viruses infecting other taxa[14,44] (Fig. S11), as if CPR/DPANN themselves contain extremely compact genomes by comparison to other taxa[30]. In the viral gene-sharing networks, CPR viral clusters were closely related (Fig. 3a), and several single modules contained viruses linked to distinct lineages (e.g., Paceibacteria, Dojkabacteria, Microgenomatia, ABY1, and JAEDAM01) (Fig. 3a and Supplementary Data 7), implying that viruses infecting CPR bacteria might be similar in terms of gene compositions.

Considering the symbiosis of CPR/DPANN and other free-living microbes[30], we subsequently verified the potential of the interphylum infection of CPR/DPANN viruses. 7 CPR viruses were predicted to infect non-CPR phyla, but no DPANN viruses were linked to non-DPANN phyla (Supplementary Data 8). Among these co-targeted CPR viruses, three proto-spacers from a complete circular genome of Gracilibacteria phage were matched to spacers from Gracilibacteria and Bacteriota, and the three matched spacers clustered with Cas gene in three CRISPR-Cas systems with different repeats (Fig. 3b). Importantly, the Gracilibacteria phage might be able to replicate in both Gracilibacteria (code 25) and Bacteriota (code 11) leveraging compatible genetic code mechanisms (code 11 and 25). Proto-spacers from another circular Gracilibacteria phage genome compatible with code 11 matched spacers from Bacteriota (Fig. S12a), whereas spacers from Proteobacteria and Bacteriota matched a linear Gracilibacteria phage genome (Fig. S12b). Indeed, no viruses infecting CPR bacteria were isolated, primarily due to the inherent difficulty in propagation of these anaerobic symbionts. However, the predicted results suggested that various laboratory-culturable Bacteriota genomes were linked to Gracilibacteria phage, implying the possibility of obtaining a culture of CPR phage in culturable microorganisms. Given that both isolation- and metagenomics-based studies have reported some phages capable of infecting across distinct bacterial phyla[16,45,46], and thus the broad host range of CPR phages identified in this study warrants further experimental verification. In such circumstances, the possibility of acquisition of such spacer(s) through horizontal gene transfer should be excluded because the CRISPR arrays (both repeat sequences and contiguous spacers) were completely different. These results suggested that CPR microorganisms as small extracellular symbionts might serve as viral bait for free-living microbes, once interphylum infection of CPR phages has occurred in groundwater ecosystems[7,30].

We further investigated how virus-associated functions may augment the metabolic and survival capacities of CPR/DPANN hosts (Fig. 3c and Supplementary Data 9). About 10 CPR phages linked to 4 host lineages (Microgenomatia, Paceibacteria, Dojkabacteria, and ABY1) and 7 DPANN viruses that infected 2 lineages (Altiarchaeota and EX4484-52) encoded Concanavalins/LamG domain proteins. Homology modeling and structure predictions suggest that these viral proteins might be involved in cell adhesion of symbionts to free-living microbes (Fig. S13), indicating viral roles in attachment or biofilm

https://doi.org/10.1038/s41467-024-51230-y

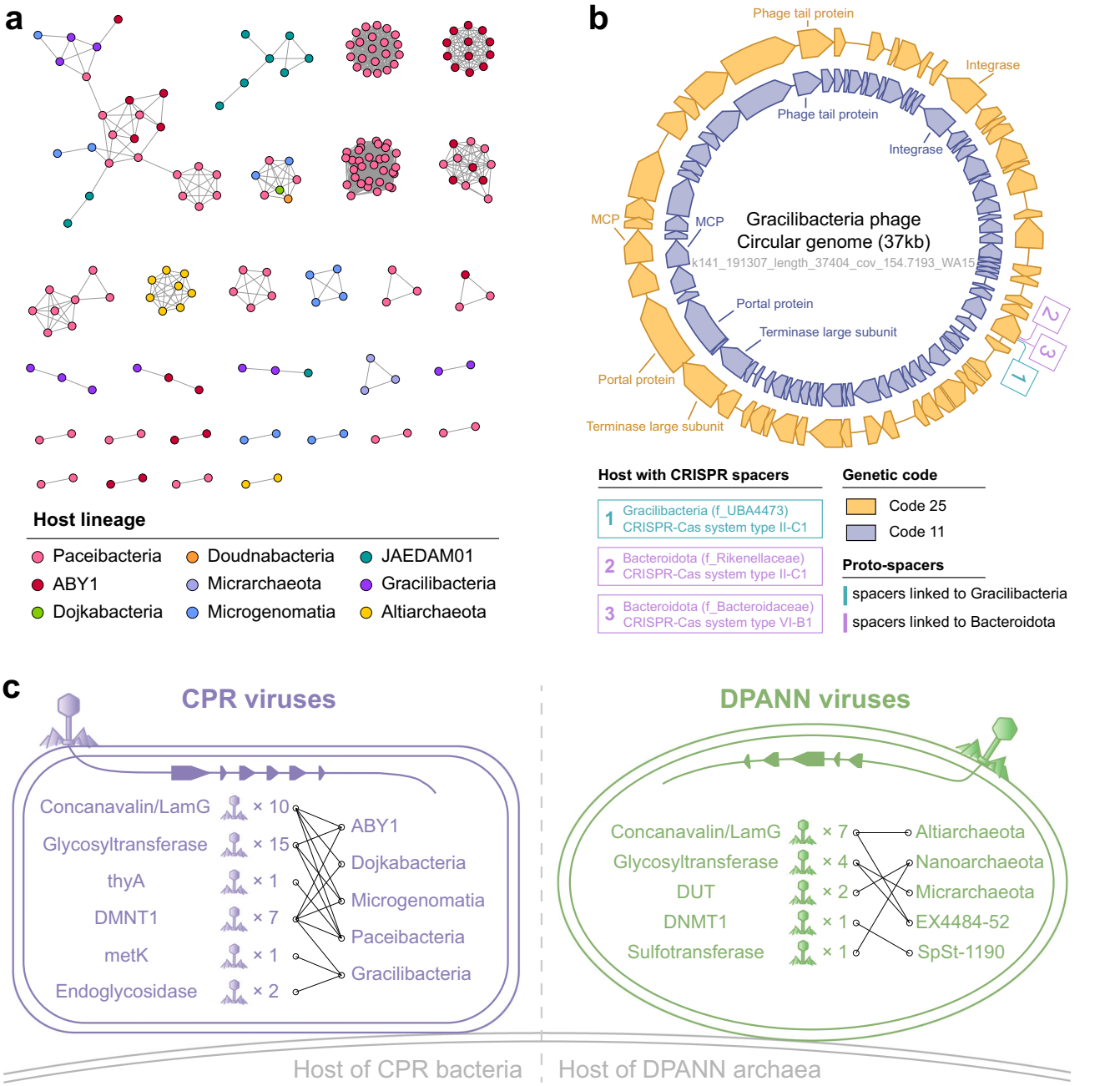

Fig. 3 | CPR bacteria and DPANN archaea host diverse viruses related to symbiosis. a Gene-sharing network of CPR/DPANN viruses identified by CRISPR-based and provirus-based methods. Host taxonomies of viruses are differentiated by colors. **b** Proto-spacers on Gracilibacteria phage were linked to CRISPR spacers from hosts affiliated with non-CPR phylum (Bacteroidota). **c** Schematics of the cell-surface modification genes and other functional genes derived from CPR/DPANN viruses. The number besides the phage icon represents the number of viruses with the corresponding genes. Connecting line between viruses and CPR/DPANN phyla represents the predicted linkage between viruses and their hosts.

formation in the CPR/DPANN organism[30,32]. Many viruses linked to different CPR/DPANN lineages (ABY1, Dojkabacteria, Microgenomatia, Paceibacteria, Micrarchaeota, and EX4484-52) encoded glycosyltransferases involving in glycosylation modification. Similar to CPR/DPANN organisms, glycosyltransferase genes were also found to be common among their viruses, indicating potential viral assistance in cell attachment and cell-surface environment regulation by enhancing the host biosynthesis capabilities of saccharides, polysaccharides and glycoproteins[30,47]. Besides, some AMGs were identified in CPR/DPANN viruses (Supplementary Data 9). For example, DUT genes related to pyrimidine metabolism and DNMT1 genes associated with methionine degradation were detected in CPR/DPANN viruses, suggesting viral contributions to the adaptation of CPR/DPANN microbes with limited metabolic capacity.

Viral auxiliary metabolic genes involved in methane, nitrogen, sulfur, and phosphorus cycles

In the GWVC, the predicted ORFs of viral contigs were widely spanned across 23 COG categories (Fig. S14a). Annotated ORFs mainly occupied the following categories which guaranteed viral transcription and replication: L (replication, recombination and repair), K (transcription), M (cell wall/membrane/envelope biogenesis), and O (post-translational modification, protein turnover, chaperones). A substantial proportion of viral ORFs were assigned to C (energy production and conversion), G

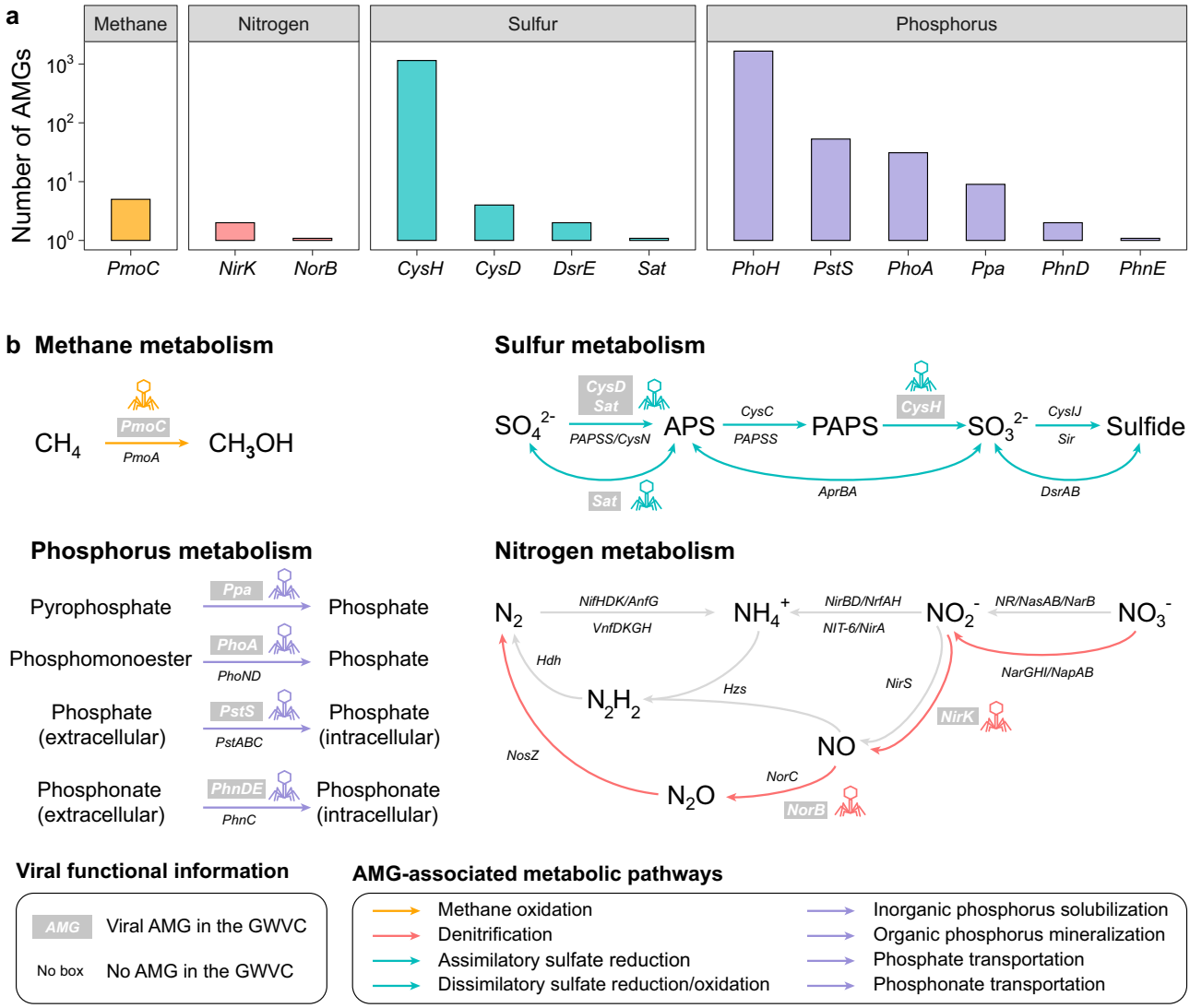

Fig. 4 | Potential viral impacts on microbial metabolisms of methane, nitrogen, sulfur, and phosphorous. a Number of viral AMGs involved in methane, nitrogen, sulfur, and phosphorous metabolisms. **b** Conceptual diagrams of viral auxiliary metabolism involved in prokaryotic metabolic pathways. Viral AMGs identified in the GWVC are highlighted in italics within grey box. The AMG-associated metabolic pathways of methane, nitrogen, sulfur, and phosphorus are distinguished with different colors.

(carbohydrate transport and metabolism), and Q (secondary metabolites biosynthesis, transport and catabolism), suggesting viral functional potential to supplement the host metabolism. Based on the results of CAZy annotation, numerous GH (glycoside hydrolases, $n = 7792$) indicated the potential impact of viruses on the microbial carbohydrate metabolism in groundwater (Fig. S14b).

Viral AMGs might directly affect biogeochemical processes by altering the methane, nitrogen, sulfur, and phosphorous metabolisms (Fig. 4 and Supplementary Data 10). With regard to the methane metabolism, 5 *pmoC* (a methane-oxidizing gene widespread in methanotrophic microorganisms) were found in 5 vOTUs, but no methanogenesis-related AMGs were identified (Fig. 4 and S15; Supplementary Data 10). Methane as a common trace constituent of groundwater could be oxidized by methanotrophic or methylotrophic microorganisms acting as an energy source[7,48]. Viruses with the *pmoC* gene might promote microbial methane oxidation for energy production during the infection cycle (Fig. 4b). Phylogeny suggests that viruses might obtain *pmoC* from Gammaproteobacteria and Alphaproteobateria, probably in two transfer events (Fig. S15). Viral *pmoC* genes acquired from Gammaproteobacteria in our study along with those recently identified in large phages from lake form a virus-specific clade[49], suggesting that these viral *pmoC* genes in surface and subsurface freshwater might share a common origin in the past. Within our knowledge, viral *pmoC* was previously reported in lake[49] and soil[50], but not found in groundwater. For nitrogen-cycling, two kinds of denitrification AMGs (*nirK* and *norB*) were identified in three vOTUs (Figs. 4 and S16; Supplementary Data 10), implying that viruses could be involved in denitrification in aquifers. The phylogenies suggested that these denitrification AMGs might be transferred from Proteobacteria and Bacteroidetes (Fig. S16). Viral *norB* and *nirK* genes have been identified in marine samples[51], yet were not reported in groundwater ecosystems. Four kinds of sulfur-cycling AMGs (*cysH*, *cysD*, *dsrE*, and *sat*) in 1114 vOTUs implied that groundwater viruses might facilitate dissimilatory sulfate reduction/oxidation and assimilatory sulfate reduction (Figs. 4 and S16; Supplementary Data 10). *CysD* and *sat* genes involve in transform SO_4^- to APS (Adenosine 5′-phosphosulfate) in sulfate reduction/oxidation processes. Intriguingly, the most abundant sulfur-cycling AMGs were *cysH* genes ($n = 1149$), which participated in the reduction of PAPS to SO_3^{2-}. Hosts for *cysH* viruses mainly included Proteobacteria, Bacteroidetes, Firmicutes, CPR bacteria, Chlorofexi, and some archaea, and phylogeny further supported the horizontal transfer of *cysH* from these microbial taxa to viruses

(Fig. S17). Virus-associated *cysH* genes have been increasingly found in human and other environmental systems[52,53], and are expected to show great potential as participants in the sulfur cycle in groundwater ecosystems. Additionally, six kinds of phosphorous-cycling AMGs (*ppa*, *phoA*, *phnD*, *phnE*, *pstS*, and *phoH*) identified from 1741 vOTUs suggested the importance of viral auxiliary metabolism in inorganic phosphorus solubilization, organic phosphorus mineralization, and phosphorus transportation (Figs. 4 and S18; Supplementary Data 10). As two major phosphorus-acquisition strategies, viral *ppa* encoding inorganic pyrophosphatase might confer host acquire phosphate via catalyzing the hydrolysis of pyrophosphate into phosphate[54], while viral *phoA* encoding alkaline phosphatase likely release bioavailable phosphate from recalcitrant phosphomonoesters[54]. Phylogenetic analysis suggested that viral *ppa* were transferred from Campylobacterota, Verrucomicrobiota, and Bacteroidota, and viruses might obtain *phoA* mainly from Actinobacteriota and Proteobacteria (Fig. S18). Viral *pstS* and *phnD*/*phnE* might involve in host phosphate transportation and phosphonate transportation[54], respectively. The most abundant phosphorus-cycling AMGs were *phoH* genes ($n = 1661$), which encode a presumed phosphate regulon protein that can be induced under phosphorus starvation[55]. These AMGs such as viral *ppa*, *phoA*, *phnD*, *phnE*, *pstS*, and *phoH* were seldom reported in groundwater, though they were noted in previous studies under surface environments[54,55]. The diverse phosphorous-cycling AMGs might be of significance for groundwater viruses to assist their hosts to cope with phosphorous-limiting stresses[54,55].

In summary, we established the largest Groundwater Virome Catalogue to date containing 280,420 viral species, and unveiled more than 99% novel viruses and about 95% novel viral clusters in the groundwater ecosystem at the continental scale. Our study expanded ~10-fold currently known aquifer viral species and doubled the number of prokaryotic phyla known to be virus-infected in groundwater. Virus-host relation analysis revealed that small-celled microbial symbionts represented by keystone microbes (CPR bacteria and DPANN archaea) in groundwater were more susceptible to viral lysis. Notably, CPR phage appeared capable of infecting free-living bacterial phyla, and CPR/DPANN viruses assisted symbiotic adhesion of cells to free-living cells. Viral AMGs related to methane, nitrogen, sulfur, and phosphorous metabolisms might directly involve in host metabolism and biogeochemical cycling. This study has provided a tremendous opportunity to understand the underexplored viral world in groundwater, and highlighted the significance of subsurface virosphere for future studies of viral ecology.

Methods

Sampling and filtration

In this study, metagenomic sequencing was performed on 607 groundwater samples collected from 525 newly constructed and 82 reconstructed monitoring wells throughout China during 2016–2017 (Fig. 1a and Supplementary Data 1). The monitoring wells were distributed in seven geo-environmental zones (Northeast plain-mountain, Huanghuaihai and river delta Yangtze plain, South China bedrock low mountain foothill, Northwest loess plateau, Southwest China Karst rock mountain, Northwest arid desert, and Qinghai-Tibet plateau Alpine frozen soil) at depths ranging from of 0 to 600 m, taking full consideration of hydrology, geological environment, and groundwater burial conditions. Groundwater was pumped after well flushing and filtered through 0.22 μm polycarbonate membranes (Millipore, USA) to capture microbial organisms. All filtered membranes were frozen at −80 °C for high-throughput sequencing. Groundwater samples for physicochemical analysis were collected in 5 L sterile PET bottles and stored at −20 °C.

DNA extraction and metagenomic sequencing

Total genomic DNA of 607 samples was extracted using the MoBio PowerSoil® kit (MoBio Laboratories, Carlsbad, CA, USA) following the manufacturer's protocol. DNA quantity and quality were determined using a NanoDrop Spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA). Genomic DNA was sequenced by Illumina HiSeq 4000 platform (Majorbio Company, Shanghai, China) with 2 × 150 bp paired-end reads. The sequencing generated 607 metagenomic datasets, encompassing over 116 billion raw reads of length 150 bp (Supplementary Data 1).

Development of the Groundwater Virome Catalogue

All raw reads were trimmed using the Read_qc module (default parameters) of metaWRAP v1.2.3[56]. Clean reads in each sample were then de-novo assembled by the assembly module (--megahit -l 500) of metaWRAP. Assembled contigs were processed for putative viral contig identification using four different viral identification methods (Earth Virome Pipeline[16,57], ViralVerify v1.1[58], VIBRANT v1.2.1[59] and PPR-meta v1.1[60]). Viral Verify, VIBRANT and PPR-meta were used with default parameters. For Earth Virome Pipeline, we expanded the Viral Protein Family (VPF) database by strict detection and manual curation of VPFs generated from groundwater metagenomics data following the recommended protocols[57]. Specifically, we compared the raw VPF database[57] against contigs from 607 groundwater metagenomics using hmmsearch[61]. Contigs with five or more VPFs and a length of ≥ 50 kb were used for further filtrations. Then, we compared these contigs to Kegg Orthology (KO)[62] and protein family (Pfam)[63] databases, and removed those with >10% of genes annotated by KO or >25% of genes annotated by Pfam. Viral proteins derived from retained contigs were de-replicated using USEARCH with 70% identity threshold, and clustered into groups using Markov cluster algorithm[57]. Proteins within clusters were aligned using MAFFT[64], and then VPFs were created using hmmbuild[61]. Generated VPFs in this study were complemented into the raw VPF database for viral identification[16,57].

Putative viral contigs identified by the above four methods were filtered using geNomad v1.7.3[65], and non-viral sequences classified by geNomad were further removed. Viral contigs were merged together for further host contamination removal and completeness estimation using CheckV v1.0.1[66], and viral contigs with a length of ≥5 kb were retained. According to Minimum Information about an Uncultivated Virus Genome (MIUViG)[67], all validated viral contigs ($n = 312,741$) were then clustered at 95% nucleotide identity over 85% coverage using CD-HIT v4.8.1[68] (parameters: -c 0.95 -d 400 -T 20 -M 20000 -n 5). Generated 280,420 non-redundant species-level vOTUs with a length of ≥5 kb constituted the Groundwater Virome Catalogue (GWVC), including 107,610 vOTUs with a length of ≥ 10 kb.

Comparison of viral genomes and proteins to public databases

Viral contigs (≥10 kb) and their encoded proteins in the GWVC were compared against existing viral genome and protein databases in IMG/VR v.3.0[38]. IMG/VR sequences were derived from groundwater, marine, human, surface freshwater, terrestrial, and wastewaters. To identify overlapping vOTUs between IMG/VR and GWVC, all sequences were clustered using CD-HIT at 95% identity over 85% coverage. GWVC amino acid sequences were identified using Prodigal v2.6.3[69] and then clustered with IMG/VR sequences using CD-HIT v.4.8.1 (parameters: -c 0.6 -G 0 -aS 0.8 -n 4).

Calculation of GC content, protein molecular weight, and protein elemental composition

Complete genomes of groundwater viruses from the GWVC and surface-water viruses from surface freshwater and ocean sections of the IMG/VR were selected for calculation of GC content of viral genomes, molecular weight, and carbon/nitrogen/sulfur atoms per residue side chain (C/N/S-ARSC) of viral proteins. Following previous methods[36], GC content, molecular weight, and C/N/S-ARSC were calculated using the python scripts 'get_gc_and_narsc.py' (https://github.com/faylward/pangenomics/).

Viral taxonomy

Taxonomic annotation of vOTUs (≥10 kb) in the GWVC was performed on geNomad v1.7.3[65] with default parameters (https://github.com/apcamargo/genomad). Viral genes of GWVC vOTUs were annotated using taxonomically informative marker profiles of geNomad, and vOTUs were then classified into distinct viral lineages according to Virus Metadata Resource of ICTV (International Committee on Taxonomy of Viruses).

Host assignment and lifestyle prediction

Contigs from each of the 607 groundwater metagenomes were binned using the binning module of metaWRAP[56] (--maxbin2 --concoct --metabat2 options), and generated MAGs were then refined using the bin_refinement module of metaWRAP (-c 70 -x 10 options). Completeness and contamination of MAGs were assessed using CheckM v1.1.2[70], resulting in 34,993 MAGs with >70% completeness and <10% contamination used for host assignment. All genomes were also dereplicated at an estimated species level (ANI ≥ 95%) with dRep v2.5.4[71] (-pa 0.9 -sa 0.95 -cm larger -comp 75 -con 5 -nc 0.30 options). The taxonomy of each genome was assigned using GTDB-Tk v2.1.6 with the GTDB database r207[72,73]. The maximum-likelihood phylogenetic tree inferred from a concatenation of 120 bacterial or 122 archaeal marker genes was also generated using GTDB-Tk.

Four previously reported in silico methods were used to link vOTUs and putative host MAGs[38,52,74,75] in terms of CRISPR spacer match, provirus identified in host genome, nucleotide sequence homology, and k-mer frequency match. First, CRISPR spacers in microbial genomes were detected using minced (https://github.com/ctSkennerton/minced) and then matched against viral contigs with ≤1 mismatch over ≥95% of the spacer length using BLASTn (-word_size 8 -task 'blastn-short'). Second, viral genomes identified as prophages by both geNomad and CheckV were linked to their corresponding host MAGs. Third, nucleotide sequence homology of vOTUs and prokaryotic MAGs were compared using BLASTn[76]. Host predictions were then based on matches of ≥90% nucleotide identity covering ≥2 kb of the virus and putative host sequences. Fourth, Prokaryotic virus Host Predictor (PHP)[77] was run with default parameters to predict viral host based on k-mer frequency match.

Viral lifestyle (virulent/temperate) on complete or high-quality vOTUs was predicted using the geNomad, CheckV and BACPHLIP tools[78]. Integrated proviruses identified by both geNomad and CheckV are considered as temperate viruses. For the remaining complete or high-quality vOTUs, the BACPHLIP based on random forest was used to predict lifestyle, and vOTUs with a greater probability (>90%) in BACPHLIP predictions were classified as virulent or temperate viruses.

Construction of groundwater CPR/DPANN virus dataset

To construct a comprehensive dataset of groundwater CPR/DPANN viruses, we identified putative CPR/DPANN viral genomes from the GWVC and public datasets (IMG/VR and NCBI). Specifically, CPR/DPANN genomes available in NCBI Genbank were collected and filtered for quality using CheckM[70]. As before, the CPR/DPANN genomes (>70% completeness and <10% contamination) were used for predicting CRISPR spacers using minced (https://github.com/ctSkennerton/minced). Extracted spacers of CPR/DPANN genomes from the NCBI and the present study were merged and then matched against viral contigs from the GWVC and the IMG/VR (groundwater section) using BLASTn (-word_size 8 -task 'blastn-short'), allowing a maximum of one mismatch over ≥95% of the spacer length. According to NCBI BioSample annotations, proviruses from CPR/DPANN genomes derived from groundwater environments were also identified using geNomad. As stated above, all CPR/DPANN viruses from the public datasets were filtered and estimated using CheckV. A total of 230 CPR viruses and 23 DPANN viruses were identified using CRISPR- and provirus- based methods, including 90 CPR viruses and 5 DPANN viruses from the GWVC. Host prediction methods based on nucleotide sequence homology and k-mer frequency match were also used to examine these linkages between viruses and CPR/DPANN genomes as stated above.

To examine whether CPR or DPANN viruses can be targeted by spacers of non-CPR or non-DPANN genomes, the spacer database of the iPHoP[79] were compared to CPR/DPANN viruses using BLASTn (-word_size 8 -task 'blastn-short') with a maximum of one mismatch over ≥ 95% of the spacer length, resulting in 7 viruses being found to co-target Gracilibacteria (CPR) and non-CPR phyla. CRISPR-Cas systems of host genomes of these co-targeted viruses were carefully examined using CRISPRCasFinder[80] and classified into different types[33]. Genomic maps of co-targeted Gracilibacteria phages were generated using prodigal[69] (single mode) in terms of genetic codes code 11 and code 25.

Generation of viral clusters using gene-sharing networks

Two viral gene-sharing networks were constructed to generate viral clusters using vConTACT2[81]. One network contained vOTUs over medium-quality from the GWVC and prokaryotic viruses from the NCBI Viral RefSeq (v201). Another network contained CPR/DPANN viruses identified from the GWVC and the public datasets. Visualization of the gene-sharing networks was implemented in Cytoscape v3.7.1[82].

Abundance profiling

RPKM (Reads per kilobase per million mapped reads) values were used to represent relative abundances of vOTUs and their host MAGs. Quality-controlled reads from each sample were mapped to a contig database with Bowtie2[83]. Sam files were sorted using SAMtools[84], and sorted bam files were then passed to CoverM v0.3.1 (https://github.com/wwood/CoverM) to filter low-quality mappings and generate RPKM profiles for all samples (parameters: contig mode for viral contigs, genome mode for prokaryotic MAGs, --trim-min 0.10 --trim-max 0.90 --min-read-percent-identity 0.95 --min-read-aligned-percent 0.75 -m rpkm).

Functional annotation and AMGs identification

Viral protein function was annotated by eggnog-mapper v2.0[85]. Briefly, predicted viral ORFs were annotated based on Diamond blastp search against protein family databases: KEGG[62], COG[86], NCBI-NR[87], Uniref[88], CAZy[89] and VOGDB. Metabolic capacity (methane, nitrogen, and sulfur metabolisms) of host MAGs was analyzed by searching predicted ORFs against a curated set of KEGG, TIGRfam[90], Pfam[63] and custom HMM profiles[91] using METABOLIC v.4.0 (https://github.com/AnantharamanLab/METABOLIC)[91].

For reliable AMGs identification[17,52,92], we performed VirSorter2[93] (--prep-for-dramv) on identified viruses to generate the input files required for DRAM-v[94], and viral contigs with high viral scores (>0.5) were selected for AMGs annotation using DRAM-v. We checked the genomic content of viral contigs containing AMGs, and only the AMGs flanked by two viral genes or viral hallmark genes were further analyzed. For genomic context assessment, genome maps for viral contigs with AMGs were visualized based on DRAM-v and VirSorter2 annotations. Protein structural homology searches and prediction were performed using the Phyre2 web portal[95].

Viral proteomic tree generation

Complete and high-quality GWVC vOTUs in viral clusters prevalent in all geo-environmental zones were compared with complete viral genomes publicly available in NCBI Refseq to generate a viral proteomic tree using VIPtree[96]. In brief, a proteomic similarity score was calculated for each pair of genomes based on an all-versus-all tblastx similarity. A proteomic tree is generated by BIONJ based on the genomic distances, and iTOL (https://itol.embl.de/) was used to visualize and display the proteomic tree[97].

Phylogenetic tree generation

We constructed a concatenated protein phylogeny of Caudoviricetes as previously described[23,98]. The 77 marker proteins were identified from the GWVC vOTUs over high-quality and NCBI Refseq viral genomes using HMMER v3.1b1[61]. Specifically, HMMs for the 77 markers were used to search against protein sequences, and the best hits (highest bitscore) were selected. Only genomes containing at least three markers were retained. All marker alignments were individually trimmed using trimAl v.1.4 (parameter: -gt 0.5)[99] and concatenated by filling in gap positions where markers were absent. We further removed genomes with <5% alignment columns, leading to a final multiple sequence alignment of 7199 genomes (4238 GWVC vOTUs and 2961 Refseq viruses) with 23,268 columns. The Caudoviricetes phylogeny was inferred from the multiple sequence alignment using FastTree v2.7.1[100] under the WAG + G model. The midpoint-rooted tree was visualized using iToL, and the family/subfamily taxonomic annotations for the NCBI Refseq viral genomes are straightly from the Virus Metadata Resource of ICTV.

Sequences similar to AMGs were recruited from the 34,993 MAGs in this study and the NCBI nr database, based on the blastp[92] searching of the identified viral AMGs (threshold of 100 for bit score, 1e-5 for E-value). These sets of viral AMGs and related protein sequences were aligned using Muscle[101], and the alignments were manually curated to remove poorly aligned positions using Jalview[102]. Maximum-likelihood trees were computed using FastTree v2.7.1[100] and visualized using iTOL.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data (viral genomes in the GWVC, prokaryotic spacer sequences, spacer hit tables of host prediction) generated in this study and 77 markers used for the Caudoviricetes phylogeny have been deposited in the zenodo repository (https://doi.org/10.5281/zenodo.11230969). Sequence reads used in this study have been deposited in the NCBI database under accession code PRJNA858913. Source data are provided with this paper.

Code availability

The custom codes used for viral identification and host assignment are available in the zenodo repository (https://doi.org/10.5281/zenodo.12740384).

References

1. Zimmerman, A. E. et al. Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat. Rev. Microbiol.* **18**, 21–34 (2020).
2. Chevallereau, A., Pons, B. J., van Houte, S. & Westra, E. R. Interactions between bacterial and phage communities in natural environments. *Nat. Rev. Microbiol.* **20**, 49–62 (2022).
3. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nat. Rev. Microbiol.* **19**, 514–527 (2021).
4. Magnabosco, C. et al. The biomass and biodiversity of the continental subsurface. *Nat. Geosci.* **11**, 707–717 (2018).
5. Anantharaman, K. et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
6. Méheust, R. et al. Groundwater Elusimicrobia are metabolically diverse compared to gut microbiome Elusimicrobia and some have a novel nitrogenase paralog. *ISME J.* **14**, 2907–2922 (2020).
7. He, C. et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. *Nat. Microbiol.* **6**, 354–365 (2021).
8. Holmes, D. E. et al. Evidence of Geobacter-associated phage in a uranium-contaminated aquifer. *ISME J.* **9**, 333–346 (2015).
9. Holmfeldt, K. et al. The Fennoscandian Shield deep terrestrial virosphere suggests slow motion 'boom and burst' cycles. *Commun. Biol.* **4**, 307 (2021).
10. Kyle, J. E., Eydal, H. S. C., Ferris, F. G. & Pedersen, K. Viruses in granitic groundwater from 69 to 450 m depth of the Äspö hard rock laboratory, Sweden. *ISME J.* **2**, 571–574, (2008).
11. Eydal, H. S. C., Jägevall, S., Hermansson, M. & Pedersen, K. Bacteriophage lytic to Desulfovibrio aespoeensis isolated from deep groundwater. *ISME J.* **3**, 1139–1147, (2009).
12. Hylling, O. et al. Two novel bacteriophage genera from a groundwater reservoir highlight subsurface environments as underexplored biotopes in bacteriophage ecology. *Sci. Rep.* **10**, 11879 (2020).
13. Cai, L., Weinbauer, M. G., Xie, L. & Zhang, R. The smallest in the deepest: the enigmatic role of viruses in the deep biosphere. *Natl Sci. Rev.* **10**, nwad009 (2023).
14. Dion, M. B., Oechslin, F. & Moineau, S. Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* **18**, 125–138 (2020).
15. Breitbart, M., Bonnain, C., Malki, K. & Sawaya, N. A. Phage puppet masters of the marine microbial realm. *Nat. Microbiol.* **3**, 754–766 (2018).
16. Paez-Espino, D. et al. Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
17. Roux, S. et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
18. Schulz, F. et al. Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).
19. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109.e9 (2021).
20. Brum, J. R. et al. Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
21. Emerson, J. B. et al. Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870–880 (2018).
22. Starr, E. P., Nuccio, E. E., Pett-Ridge, J., Banfield, J. F. & Firestone, M. K. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc. Natl Acad. Sci. USA* **116**, 25900–25908 (2019).
23. Nayfach, S. et al. Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
24. Chen, Y., Wang, Y., Paez-Espino, D., Polz, M. F. & Zhang, T. Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. *Nat. Commun.* **12**, 5398 (2021).
25. Daly, R. A. et al. Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat. Microbiol.* **4**, 352–361 (2019).
26. Kothari, A. et al. Ecogenomics of groundwater phages suggests niche differentiation linked to specific environmental tolerance. *mSystems* **6**, e0053721 (2021).
27. Rahlff, J. et al. Lytic archaeal viruses infect abundant primary producers in Earth's crust. *Nat. Commun.* **12**, 4642 (2021).
28. Burstein, D. et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* **7**, 10613 (2016).
29. Zhong, S. et al. May microbial ecological baseline exist in continental groundwater? *Microbiome* **11**, 152 (2023).
30. 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).
31. Dombrowski, N., Lee, J.-H., Williams, T. A., Offre, P. & Spang, A. Genomic diversity, lifestyles and evolutionary origins of DPANN archaea. *FEMS Microbiol. Lett.* **366**, fnz008 (2019).
32. Dombrowski, N. et al. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

33. Makarova, K. S. et al. Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
34. Burstein, D. et al. New CRISPR–Cas systems from uncultivated microbes. *Nature* **542**, 237–241 (2017).
35. Li, Y.-X. et al. Deciphering Symbiotic Interactions of "*Candidatus* Aenigmarchaeota" with Inferred Horizontal Gene Transfers and Co-occurrence Networks. *mSystems* **6**, e00606–e00621 (2021).
36. Mende, D. R. et al. Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nat. Microbiol.* **2**, 1367–1373 (2017).
37. Grzymski, J. J. & Dussaq, A. M. The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J.* **6**, 71–80 (2011).
38. 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).
39. Bischoff, V. et al. Cobaviruses—a new globally distributed phage group infecting Rhodobacteraceae in marine ecosystems. *ISME J.* **13**, 1404–1421 (2019).
40. Koonin, E. V. et al. Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* **84**, e00061–19 (2020).
41. Roux, S. et al. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat. Microbiol.* **4**, 1895–1906 (2019).
42. Dudek, N. K. et al. Novel Microbial Diversity and Functional Potential in the Marine Mammal Oral Microbiome. *Curr. Biol.* **27**, 3752–3762.e6 (2017).
43. Li, S. et al. A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. *iScience* **25**, 104418 (2022).
44. Al-Shayeb, B. et al. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
45. Hwang, Y., Roux, S., Coclet, C., Krause, S. J. E. & Girguis, P. R. Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat. Microbiol.* **8**, 946–957 (2023).
46. Malki, K. et al. Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virol. J.* **12**, 164 (2015).
47. Castelle et al. Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Curr. Biol.* **25**, 690–701 (2015).
48. Barker, J. F. & Fritz, P. Carbon isotope fractionation during microbial methane oxidation. *Nature* **293**, 289–291 (1981).
49. Chen, L.-X. et al. Large freshwater phages with the potential to augment aerobic methane oxidation. *Nat. Microbiol.* **5**, 1504–1515 (2020).
50. Lee, S. et al. Methane-derived carbon flows into host–virus networks at different trophic levels in soil. *Proc. Natl Acad. Sci. USA* **118**, e2105124118 (2021).
51. Gazitúa, M. C. et al. Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. *ISME J.* **15**, 981–998 (2021).
52. Li, Z. et al. Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME J.* **15**, 2366–2378 (2021).
53. Kieft, K. et al. Virus-associated organosulfur metabolism in human and environmental systems. *Cell Rep.* **36**, 109471 (2021).
54. Liang, J.-L. et al. Hidden diversity and potential ecological function of phosphorus acquisition genes in widespread terrestrial bacteriophages. *Nat. Commun.* **15**, 2827 (2024).
55. Gao, S. et al. Patterns and ecological drivers of viral communities in acid mine drainage sediments across Southern China. *Nat. Commun.* **13**, 2389 (2022).
56. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
57. Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N. & Kyrpides, N. C. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.* **12**, 1673–1682 (2017).
58. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. MetaviralSPAdes: assembly of viruses from metagenomic data. *Bioinformatics* **36**, 4126–4129 (2020).
59. 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).
60. Fang, Z. et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* **8**, giz066 (2019).
61. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
62. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
63. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res* **42**, D222–D230 (2014).
64. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780, (2013).
65. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.*, (2023).
66. Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
67. Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).
68. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
69. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
70. 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).
71. 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).
72. 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).
73. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
74. 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).
75. Kavagutti, V. S., Andrei, A., Mehrshad, M., Salcher, M. M. & Ghai, R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. *Microbiome* **7**, 135 (2019).
76. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
77. Lu, C. et al. Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol.* **19**, 5 (2021).

78. Hockenberry, A. J. & Wilke, C. O. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. *PeerJ* **9**, e11396 (2021).
79. Roux, S. et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol.* **21**, e3002083 (2023).
80. Couvin, D. et al. CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* **46**, W246–W251 (2018).
81. Bin Jang, H. et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
82. Shannon, P. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13**, 2498–2504 (2003).
83. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359, (2012).
84. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
85. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
86. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–36, (2000).
87. Sayers, E. W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20–D26 (2022).
88. Suzek, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2014).
89. Cantarel, B. L. et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**, D233–D238 (2009).
90. Selengut, J. D. et al. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res* **35**, D260–D264 (2007).
91. Zhou, Z. et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* **10**, 33 (2022).
92. Pratama, A. A. et al. Expanding standards in viromics: in silico evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation. *PeerJ* **9**, e11447 (2021).
93. Guo, J. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
94. Shaffer, M. et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res* **48**, 8883–8900 (2020).
95. 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).
96. Nishimura, Y. et al. ViPTree: the viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).
97. 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).
98. Low, S. J., Džunková, M., Chaumeil, P.-A., Parks, D. H. & Hugenholtz, P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat. Microbiol.* **4**, 1306–1315 (2019).
99. 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).
100. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
101. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792–1797 (2004).
102. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

Acknowledgements

Financial support is from the National Natural Science Foundation of China under Grant No. 51721006 (J.R.N.), U2240205 (J.R.N.), and 423B2703 (Z.Z.W.). Special support from the Hydrologic Bureau of Ministry of Water Resources, China is appreciated. Sincere thanks are to Professor Alistair Borthwick from University of Edinburgh for his help in English editing. Bioinformatic support from the High-performance Computing Platform of Peking University and the open access policy of the Geocloud Database developed by China Geological Survey are also acknowledged.

Author contributions

J.R.N. designed the research. Z.Z.W., T.L. and Q.C. conducted the bioinformatic and statistical analysis with help from T.Y.C., J.Y.H., L.Y.S., B.X.W. and W.P.L. Z.Z.W. and J.R.N. wrote the paper. All authors read and approved the final manuscript.

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/s41467-024-51230-y.

Correspondence and requests for materials should be addressed to Jinren Ni.

Peer review information *Nature Communications* thanks Clément Cocle, and Akbar Adjie Pratama for their contribution to the peer review of this work. A peer review file is available.