microLife, 2024, **5**, uqae011
DOI: 10.1093/femsml/uqae011
Advance access publication date: 20 May 2024
Research Article

Time-series metaproteogenomics of a high-CO_2 aquifer reveals active viruses with fluctuating abundances and broad host ranges

Carrie Julia Moore[1], Till L. V. Bornemann[1,2], Perla Abigail Figueroa-Gonzalez[1], Sarah P. Esser[1], Cristina Moraru[1], André Rodrigues Soares[1,2], Tjorven Hinzke[3,4], Anke Trautwein-Schult[5], Sandra Maaß[5], Dörte Becher[5], Joern Starke[1], Julia Plewka[1], Lousia Rothe[2], Alexander J. Probst[1,2,*]

[1]Environmental Metagenomics, Research Centre One Health Ruhr of the University Alliance Ruhr, Faculty of Chemistry, University Duisburg-Essen, 45141 Essen, Germany
[2]Centre of Water and Environmental Research (ZWU), University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, Germany
[3]Department for Microbial Physiology and Molecular Biology, Institute of Microbiology, University of Greifswald, 17489 Greifswald, Germany
[4]Department of Pathogen Evolution, Helmholtz Institute for One Health, 17489 Greifswald, Germany
[5]Microbial Proteomics, Institute of Microbiology, University of Greifswald, 17489 Greifswald, Germany
*Corresponding author. Environmental Metagenomics, Research Centre One Health Ruhr of the University Alliance Ruhr, Faculty of Chemistry, University Duisburg-Essen, 45141 Essen, Germany. Tel: +49 (201) 183-7080; Fax: +49 (201) 183-6603; E-mail: alexander.probst@uni-due.de
Editor: [David Prangishvili]

Abstract

Ecosystems subject to mantle degassing are of particular interest for understanding global biogeochemistry, as their microbiomes are shaped by prolonged exposure to high CO_2 and have recently been suggested to be highly active. While the genetic diversity of bacteria and archaea in these deep biosphere systems have been studied extensively, little is known about how viruses impact these microbial communities. Here, we show that the viral community in a high-CO_2 cold-water geyser (Wallender Born, Germany) undergoes substantial fluctuations over a period of 12 days, although the corresponding prokaryotic community remains stable, indicating a newly observed "infect to keep in check" strategy that maintains prokaryotic community structure. We characterized the viral community using metagenomics and metaproteomics, revealing 8 654 viral operational taxonomic units (vOTUs). CRISPR spacer-to-protospacer matching linked 278 vOTUs to 32 hosts, with many vOTUs sharing hosts from different families. High levels of viral structural proteins present in the metaproteome (several structurally annotated based on AlphaFold models) indicate active virion production at the time of sampling. Viral genomes expressed many proteins involved in DNA metabolism and manipulation, and encoded for auxiliary metabolic genes, which likely bolster phosphate and sulfur metabolism of their hosts. The active viral community encodes genes to facilitate acquisition and transformation of host nutrients, and appears to consist of many nutrient-demanding members, based on abundant virion proteins. These findings indicate viruses are inextricably linked to the biogeochemical cycling in this high-CO_2 environment and substantially contribute to prokaryotic community stability in the deep biosphere hotspots.

Keywords: metaproteogenomics; aquifer; prokaryotic viruses; high-CO_2; subsurface; time-series

Background

Terrestrial subsurface environments contribute 12%–20% to global biomass, yet remain poorly characterized compared to marine and soil biomes (Bar-On et al. 2018, Soares et al. 2023). A growing body of research suggests that terrestrial subsurface environments are biodiversity hotspots harboring prokaryotes with diverse metabolic capabilities that contribute to planetary biogeochemical cycling (Anantharaman et al. 2016, Flemming et al. 2016, Probst et al. 2017). Ecosystems subject to mantle degassing are of particular interest for understanding global biogeochemistry; these microbiomes are shaped by prolonged exposure to volcanic gasses, such as CO_2. Investigations into prokaryotic replication measures in environments exposed to mantle degassing demonstrated that subsurface communities present similar replication rates as surface dwellers, indicating that mantle degassing fuels community growth (Bornemann et al. 2022). Studies on subsurface viruses indicate high levels of diversity, such as in old fracking wells dominated by *Halanaerobium* spp., where over a thousand unique viral populations have been recovered (Daly et al. 2018). Despite this, little is known about viruses in high-CO_2 subsurface environments.

Subsurface viromes contain primarily viruses with little similarity to other ecosystems or isolates (Holmfeldt et al. 2021). Prokaryotic hosts of these viruses are not well understood either. For example, across multiple sampling sites in the Fennoscandian Shield bedrock, many viruses infected a broad range of hosts, many of which lacked previous descriptions of viral interactions. Dynamics in this bedrock subsurface environment suggested the presence of a slow-motion "boom and bust" cycle, where viral activity facilitates nutrient recycling, allowing for increased microbial activity in this subsurface environment (Holmfeldt et al. 2021). Viral-host interactions appeared to follow the kill-the-winner model, with the dominant microbial community shifting over the 8-month observation period, presumably due to viral

Received 22 December 2023; **revised** 5 April 2024; **accepted** 18 May 2024

predation (Holmfeldt et al. 2021). Kill-the-winner dynamics have also been observed in other subsurface environments, such as that of the sulfidic aquifer Mühlbacher Schwefelquelle (Regensburg, Germany), where subsurface viruses target the highly abundant primary producers, i.e. *Ca.* Altiarchaeum hamixonexum (Rahlff et al. 2021, Esser et al. 2023, Turzynski et al. 2023). Little is known about the diversity and activity of viruses in high-CO_2 environments. One of the few studies investigating viruses in this type of environments found huge phage genomes up to 415 kbp in length in metagenomes from Crystal Geyser (Utah, USA) (Al-Shayeb et al. 2020). The genomes of these huge phages were annotated to encode for structural proteins, tRNAs, and proteins involved in DNA synthesis, DNA replication, and protein folding (Al-Shayeb et al. 2020). Insights into the hosts of viruses from high-CO_2 environments can be hampered by the lack of CRISPR-Cas systems in many uncultured bacterial lineages (Burstein et al. 2016). Previous work on the Geyser Wallender Born, the ecosystem-of-interest in this study, identified *Gammaproteobacteria* to be dominant in the community. This class of organism, among others identified in the ecosystem, have well documented reports of CRISPR-Cas systems in groundwater ecosystems and global databases, indicating a potential for identifying the associated viral community of the major players of this ecosystem (Burstein et al. 2016, Figueroa-Gonzalez et al. 2023).

In this study, we characterize the viral community in a subsurface environment subject to mantle degassing using temporally resolved metagenomes and metaproteomes collected from filter fractions of the cold-water geyser Wallender Born in the Volcanic Eifel region in Germany. Driven by a build-up of CO_2 from mantle degassing, this geyser regularly discharges groundwater fluids containing complex microbial communities, allowing easy sampling access to this subsurface ecosystem (Figueroa-Gonzalez et al. 2023). We sequentially filtered subsurface fluids over a twelve-day sampling period to provide a highly resolved metaproteogenomic time series, and identified strain-level viral infection histories in this environment using genome-resolved metagenomics and spacer-to-protospacer matching. Abundance of the viral genomes demonstrated fluctuations in 0.1-µm fractions across the time series. By contrast, most prokaryotic metagenome-assembled genomes (MAGs) demonstrated relatively stable coverage values in the 0.2-µm fraction across time. Viral genomes encoded auxiliary metabolic genes putatively involved in phosphate and sulfur metabolism of their hosts, though none were found actively expressed in the metaproteomes. The viral protein profiles of all viruses identified in the metaproteomes revealed temporal changes in virion protein presence. Virion proteins, specifically capsid proteins, were the most abundant viral proteins, but we also detected diverse nonstructural proteins from different stages of the viral life cycle. These findings indicate that an active viral community with demanding nutrient requirements for structural protein production exists in this high CO_2 environment.

Methods

Sampling

The time series sample was collected from the geyser Wallender Born (Volcanic Eifel region, Germany) over the course of 12 days (October 14th-25th, 2020), with 2–3 samples collected each day. Groundwater pushed to the surface during eruptions was collected in DNA-free containers and immediately filtered through pairs of 0.2-µm filters. Flow-through from each pair of 0.2-µm filters was then filtered through a single 0.1-µm filter. One additional bulk sample was collected by filtering water directly onto a 0.1-µm filter. A schematic of the sampling process is available in Figueroa-Gonzalez et al. (2023). Filters were stored in sterile Falcon tubes on dry ice during the course of sampling, then transferred to a −80°C freezer before processing.

DNA extraction and sequencing

Biomass from the pairs of 0.2-µm filters were pooled to create a single comparable sample for each 0.1-µm filter. This resulted in 44 samples, 22 from each filter size. One-third of the biomass from each 0.1-µm filter and from each combined set of 0.2-µm filters was used for DNA extraction, one-third was reserved for protein extraction, and the final third was used for lipid extraction when possible. 30 of these samples contained enough biomass for DNA extractions, which were performed with the DNeasy PowerMax Soil DNA Extraction kit (Qiagen, 12988–10) followed by further concentration via ethanol precipitations with a glycogen carrier. The Westburg NGS DNA Library Prep Kit (cat. No. WB 9096) was used to prepare DNA libraries, which were sequenced on an Illumina NextSeq500 (2 × 150 bp paired-end reads). 20 Gbps sequencing depth per sample was reached for 28 of the samples, though some samples were sequenced up to three times and concatenated to produce the aspired sequencing depth. Due to the low biomass of this environment, filters did not provide enough biomass to allow for sample replicates.

Read processing, assembly and binning

Metagenomic processing was performed on an Ubuntu server 20.04.4 LTS (GNU/Linux5.4.0–124-generic x86_64) with 40 cores and 1.5 TB RAM. Illumina adaptors and sequencing artifacts were removed from paired-end raw reads using BBduk v37.09 (Bushnell, https://sourceforge.net/projects/bbmap), trimmed by Sickle v1.33 (https://github.com/najoshi/sickle) and dereplicated by dedupe (Bushnell, https://sourceforge.net/projects/bbmap). Trimmed paired-end reads were assembled first using MetaViralSPAdes v3.15.2 (Antipov et al. 2020) to assemble viral scaffolds and plasmids. Reads that did not map onto the assembled scaffolds from MetaViralSPAdes (mapping with Bowtie2 v2.3.5.1, –sensitive mode) were further assembled using MetaSPAdes v3.15.2 (Nurk et al. 2017). Differential coverage of assembled scaffolds with minimum lengths of 1000 bp was calculated by cross-mapping reads from all metagenomes to the assemblies using Bowtie2 (Langmead and Salzberg 2012). Binning was performed using CONCOCT v1.1.0 (Alneberg et al. 2014), MaxBin2 v2.2.7 (Wu et al. 2016), and abawaca v1.0.0 (https://github.com/CK7/abawaca). In the case of MaxBin2, scaffolds were binned once with gene marker set 107 (bacterial) and once with gene marker set 40 (bacterial and archaeal). When binning with abawaca, input scaffolds were prepared with esomWrapper.pl (https://github.com/tetramerFreqs/Binning/tree/master/esomWrapper.pl, -min 3 kbp/5 kbp, -max 5 kbp/10 kbp) for two separate binning runs. DAS Tool v1.1.2 (Sieber et al. 2018) was used to select the best bin representatives for curation. Manual curation to reduce contamination was performed using uBin v0.9.20 (Bornemann et al. 2023). All scaffolds identified as viral in the "viruses.fna" output file from CheckV v0.7.0 (Nayfach et al. 2021), which excludes those identified as prophages, were removed from the curated bins (viral workflow described below). CheckM1 v1.1.3 (Parks et al. 2015) with database 2015_01_16 was used to determine final completeness and contamination of the microbial bins. dRep v3.2.2 (Olm et al. 2017) was used to dereplicate MAGs at 99% gANI (strain-level). Bins with

completeness greater than 70% completeness and less than 10% contamination (medium and high quality) (The Genome Standards Consortium et al. 2017) as determined by CheckM1 from both pre- and post- dereplication were used in further analysis.

Viral detection

Viral scaffolds were identified from assembled scaffolds using VirSorter2 v2.2.3 (–high-confidence-only) (Guo et al. 2021), VIBRANT v1.2.1 (Kieft et al. 2020), and DeepVirFinder v1.0 (Ren et al. 2020). CheckV end_to_end was used to assess completeness of identified viral scaffolds (Table S4). Scaffolds of 25% or greater completeness without any CheckV warnings, with lengths greater than 3000 bp proceeded to further analysis.

Spacer-to-protospacer matching

CRISPR systems were extracted from all medium and high quality MAGs using CRISPRCasFinder v4.2.19 with flags specifying a maximum 10% mismatches (2–3 mismatches in a typical direct repeat) between direct repeats and a better detection of truncated repeats (Couvin et al. 2018). Only repeats and spacers from evidence level 4 CRISPR arrays were used in the following analysis, which is important because it ensures CRISPR arrays have at least four spacers as well as neighboring Cas proteins, ensuring a high level of confidence. Spacers from the detected evidence level 4 CRISPR systems were collected. Additional spacers were extracted from the trimmed paired-end reads using the repeats identified by CRISPRCasFinder as input for metaCRAST (-q -d 3 -l 60 -r) (Moller and Liang 2017). Post-hoc clustering of extracted spacers from both metaCRAST and CRISPRCasFinder was performed at 90% using CD-HIT v4.8.1 (Li and Godzik 2006). Size distribution of spacer sequences was calculated, and upper and lower length quartiles were removed, resulting in filtered spacers of 24-70 bp in length. Spacers were aligned against identified viral scaffolds with blastn v2.9.0 (-short algorithm) (Altschul et al. 1990) and results were filtered with a 80% sequence similarity (alignment length × identity/query length) threshold using an in-house script. Phage genomes with only 1–2 spacer matches were manually inspected to remove any false positives that occurred due to a spacer consisting mainly of single nucleotide polymorphisms. Viral-host matches are available in Table S5.

Taxonomy of prokaryotes and viruses

Taxonomy was assigned to MAGs using GTDB-Tk v2.1 with the classify_wf workflow against database version r207 (Chaumeil et al. 2022). Taxonomy was assigned to viral contigs using geNomad v1.5.1 with database v1.3 (end-to-end –cleanup –splits 8) (Camargo et al. 2023).

Determining genome coverage

Coverage of MAGs was determined by first mapping the reads of each metagenome to the dereplicated MAGs using Bowtie2 (–sensitive mode, which was used in all following mentions of mapping).

Coverage of viral scaffolds with host matches was performed by first clustering scaffolds with VIRIDIC v1.0_r3.6 (Moraru et al. 2020) at 95% similarity clustering, for genus and species level clusters, respectively. Coverage was determined by mapping reads from each metagenome to the longest genome representative for each species-level cluster (vOTU) using Bowtie2 and filter to allow 5% mismatches, with the higher mismatch tolerance accounting for the higher mutation rate in viruses compared to prokaryotes. Breadth (i.e. number of nucleotide positions with coverage of at least 1) was calculated for each genome in each sample, and coverage was set to 0 when breadth was less than 75%. Coverage was normalized between samples as follows (Table S3A and S3B with coverage for prokaryotes and viruses respectively):

$$(\text{coverage}/\#\text{BPi}) * \max(\#\text{BP}))$$

In this manuscript, we describe abundance of organisms in two ways: relative abundance (reflective of the reads mapped to each organism in each sample) and % relative abundance (relative abundance calculated as a % of 100% in each sample). Relative abundance allows us to measure absolute changes in abundance between samples, whereas % relative abundance gives us information about how abundant organisms are in relation to each other in each sample.

Protein prediction and annotation

Open reading frames (ORFs) of all metagenomic assemblies were predicted by Prodigal 2.6.3 in meta-mode (Hyatt et al. 2010). Annotations of ORFs were performed against FunTaxDB 1.3 (Uniref100 functions and taxonomy, downloaded 2022–08-11) (Bornemann et al. 2023). Additional annotation of identified viral proteins were performed against the PHROG hidden Markov models (HMMs) v4 (Terzian et al. 2021) using hmmscan (-id 100 -diff 0 -p 50 -z 1 -Z 600) (HMMER Software Suite) keeping only hits with e-value <0.01 and selecting the hit with the highest bit-score. Structural predictions were performed using AlphaFold v2.2.0 (–max_template_date=2022–01-01) (Jumper et al. 2021, Varadi et al. 2022). The ranked_0.pdb file, which contains the structure with the highest confidence, was queried against the FoldSeek server (accessed August 2023) (Van Kempen et al. 2023) and DALI server (accessed July 2023) (Holm 2020; Holm et al. 2023) for structural matches. Consensus between databases, or a high degree of confidence from a single database (Z>20 for DALI, e-value<10^{-7} for FoldSeek), was used to select protein annotation.

Protein extraction

Metaproteomics workflow and schematic for these samples is also available in Figueroa-Gonzalez et al. (2023) and the methods were developed from Deusch and Seifert (2015). Frozen biomass remaining on the filters after DNA extraction was distributed to generate three comparative metaproteomics datasets: time, size, and bulk. In the time dataset, 2.25 g from each pair of 0.2-µm filters was collected (total 4.5 g per time point), resulting in a highly resolved temporal metaproteome of the microbial fraction. For the size dataset, biomass from sets of 3–4 0.1-µm filters were pooled to generate 4.5 g of biomass for an early, middle, and late time series representation. Associated 0.2-µm filters were also pooled in this manner to generate equivalent biomass for comparative early, middle, and late samples. Lastly, biomass to generate the bulk dataset was taken 1) from the bulk filter and 2) from combining remaining all 0.2-µm filters not already included in the size samples because the corresponding 0.1-µm filters had too low biomass.

With sample biomass distributions in mind, the protein extractions were performed by first placing small pieces of filters into low protein binding reaction tubes. One volume of resuspension solution 1 (50 mM Tris-HCl pH 7.5, 0.1 mg/ml chloramphenicol, 1 mM PMSF [phenylmethanesulfonyl fluoride]) was used to cover the filters, then samples were vortexed to resuspend biomass. Next, 1.5 volumes of resuspension solution 2 (20 mM Tris-HCl pH 7.5, 2% SDS) were added to the samples prior to a 10 minute incubation at 60°C with shaking. After incubation, samples were

re-equilibrated to room temperature then mixed with 5 volumes of DNAse (1 µg/ml DNAse I). Lysis of samples was performed via ultrasonication on ice for 6 minutes (amplitude 51–60%; cycle 0.5) followed by incubation at 37°C for 10 min with vigorous shaking. Samples were centrifuged to pellet the cell debris, and proteins in the supernatants were precipitated by adding precooled trichloroacetic acid (final v/v concentration 20% TCA). Incubation at 4°C for 30 min in an overhead inverter allowing for the precipitation. Precipitated proteins were pelleted, washed with precooled acetone, then dried. Protein pellets were resuspended in 2x SDS sampling solution (0.125 M Tris-HCl pH 6.8, 4% [w/v] SDS, 20% [v/v] glycerol, 10% (v/v) ß- mercaptoethanol), incubated in an ultrasonic bath for 15 min, then heated at 95°C for 5 min. Next, samples were centrifuged, supernatant saved, and remaining pellet was once again treated with 2x SDS sampling solution. Supernatants and pellets (the two protein fractions) were separated by SDS-PAGE in SDS Gels (Criterion TGX 4–20%, 12+2 wells, Bio-Rad). Coomassie staining was used to verify protein presence, then proteins were digested in the gel, which involved excising each gel lane into ten equidistant pieces, destaining, washing, and digesting with trypsin (Bonn et al. 2014). Digested peptides were eluted into water via an ultrasonic bath and desalted using C18 ZipTip columns (Merck) according to the manufacturer's guidelines.

Proteomics database generation

Open reading frames (ORFs) predicted by Prodigal (Hyatt et al. 2010) were pooled across all metagenomes. ORFs that had been binned were predicted with the default translation codes 11 or 4, except for cases where a binned scaffold was assigned to a Gracilibacteria bin, where the ORFs were predicted with translation code 25. ORFs not assigned to a MAG were predicted with prodigal in -meta mode. ORFs were uniquified to produce database representatives with usearch -fastx_uniques (Edgar 2010). Each database entry was assigned a 6-digit "Accession" number to be compatible with Mass spectrometry spectra identification software. Related metadata for each accession (e.g. Bin of origin, functional and taxonomic annotation, etc) was stored in a separate file (available in SI of Figueroa-Gonzalez et al. 2023).

Mass spectrometry and data processing

Trypsinized peptides were separated with an EASY-nLC 1200 liquid chromatography system and then subjected to MS/MS analyses on a LTQ Orbitrap Elite instrument (ThermoFisher Scientific). The chromatography column used was a self-packed analytical column (OD 360 µm, ID 100 µm, length 20 cm) filled with 3 µm diameter C18 particles (Maisch), and peptides were eluted by a binary nonlinear gradient of 5 to 99% acetonitrile in 0.1% (v/v) acetic acid over 82 min with a flow rate of 300 nl/min. Eluted peptides were measured in full scan mode by the Orbitrap at a 60 000 resolution. The 20 most abundant precursor ions were subjected to collision-induced dissociation to produce structure-informative fragments. Measured MS/MS spectra were searched against a forward-reverse database produced from the metagenome protein database, which contained 15 269 672 unique entries. Mascot (Matrix Science; version 2.7.0.1) software was used to perform this database search and was given the following parameters under the assumption of trypsin enzyme digestion: 10 ppm tolerance for parent and 0.5 Da tolerance for fragment ions, up to two missed cleavages, methionine oxidation as a variable modification. The database search results were merged and validated using Scaffold v5.1.2 (Proteome Software Inc.). An additional X!Tandem database search (default settings) was performed during the creation of the scaffold file for further validation of peptide and protein identifications. The false discovery rate (FDR) was set to 5% and protein groups were required to have at least 1 unique peptide for assignment.

Spectral count normalization

Raw spectral counts were normalized across the samples by normalizing to protein length, and by total spectral count per sample as previously described (Zybailov et al. 2006). Data will be available through the ProteomeXchange Consortium via the PRIDE (Perez-Riverol et al. 2022) partner repository (dataset identifier PXD042980).

Host and non-host metabolic activity statistics

Metagenome-based genome relative abundance and metaproteome-based genome relative abundances from the 0.2-µm fractions were transformed into percentages. Median and mean % relative abundances were calculated for each genome. Ratios were generated by dividing median metaproteome-based genome % relative abundances by metagenome-based genome % relative abundances, and likewise for mean. Median was determined to be the appropriate metric, as it is outlier resistant and no ratios produced 0 results. A Welch two-sample t-test (R Core Team 2022) was used to compare the median ratios of hosts and non-hosts within the Gammaproteobacteria class, but no significance was found.

Viral protein identification

ORFs predicted on viral scaffolds (identified via the viral detection workflow) were matched to their accession numbers, then these viral accession numbers were searched for in the time, bulk, and size metaproteomes in RStudio (2022.12.0+353) (R Core Team 2022).

VirClust annotation

FASTA sequences of selected viral genomes were uploaded to the VirClust webserver in July 2023 https://rhea.icbm.uni-oldenburg.de/virclust/ (Moraru 2021). Annotations were performed using all available databases, merged and the consensus annotation was manually established.

Visualization

Results were tidied and visualized in RStudio using ggplot2 (Wickham 2016), tidyverse (Wickham et al. 2019), broom (Robinson 2014), ggalluvial (Brunson 2020), gggenes (github.com/wilkox/gggenes/tree/master), ggstream (github.com/davidsjoberg/ggstream) and patchwork (https://github.com/thomasp85/patchwork) packages.

Results and discussion

A stable prokaryotic community with fluctuating associated viral abundance over twelve consecutive days in Geyser Wallender Born

Viral genomes were retrieved from 0.67 Tbps of metagenomic sequencing produced from a 12-day time series of 0.1-µm and 0.2-µm size-fractioned eruption water samples. We identified 16 825 scaffolds to be of viral origin, with 25% or more viral genome completeness by CheckV (Table S1) (Nayfach et al. 2021). Viral scaffold clustering at 95% similarity using VIRIDIC (Moraru et al. 2020) produced 8 654 vOTUs. GeNomad (Camargo et al. 2023)

taxonomy assignment determined three viral realms to be present in the metagenomes: *Duplodnaviria*, *Monodnaviria*, and *Varidnaviria* (Table S2). Among these realms, 97.22% of viral scaffolds were assigned to *Duplodnaviria*, class *Caudoviricetes*. Another 1.03% % of the scaffolds were assigned to the *Monodnaviria* realm, 0.64% of the scaffolds were assigned to *Varidnaviria*, and 1.23% of the scaffolds remained unclassified. The relative abundance of the viral community compared to the prokaryotic community is depicted in Fig. 1, where the viral community demonstrates abundance variability over the time series, compared to the relatively stable microbial community. Further visualizations of stability (or lack thereof) are available in the Supplemental Materials, where we investigated the range of abundances for each organism class (Fig. S1).

Since viruses are inert without a host to hijack, analysis of the associated prokaryotes in this environment was essential. Previous work elaborates on the 751 MAGs retrieved from the samples, which clustered into 123 strains after dereplication (Figueroa-Gonzalez et al. 2023). Briefly, iron-oxidizing *Gallionella spp.* dominate the community, with a single MAG of *Gallionella spp.* (MAG name: Nitrosomondales_51_490) recruiting 30% of the reads mapped to the microbial community in the 0.2-µm fraction (Table S3A) (Figueroa-Gonzalez et al. 2023). At the same time, 55 MAGs contribute less than 0.1% relative abundance each (for differentiation of relative abundance versus % relative abundance please see Methods section of this manuscript). Overall, the abundance of prokaryotic community members, especially of those who are the most dominant, is relatively stable across time.

Relative abundance values of the entire identified viral community displayed large fluctuations in the 0.1-µm fraction over the twelve days (refer to Fig. S2 for a non-log10 transformed representation of prokaryotic-linked viral abundances where fluctuations are drastic). In contrast, viral communities had much lower abundances with less drastic fluctuations across time in the 0.2-µm fraction (Fig. 1, Fig. S2). Although we dedicated 0.67 Tbps of sequencing to our samples, some of the observed fluctuations in the 0.1-µm fraction might be a result of generally low read recruitment of viral scaffolds in the samples paired with our stringent cutoffs of 75% breadth in read coverage to call a virus present and to avoid false positives. In addition, few eukaryotic sequences, which can tamper with assembly and lead to fragmented scaffolds with low read recruitment, were present in the sample (Fig. S3). To conclude, the changes in the relative abundances of viruses and prokaryotes across time as seen in our data are likely representative of the actual biome in the ecosystem (Fig. 1).

Although 0.2-µm and 0.1-µm filters typically correspond to "cellular" and "viral" fractions, we did not observe a stringent divide in the communities identified on the two filter fractions. For example, many prokaryotes, including the abundant *Gallionella spp.* MAGs were identified in both filter fractions, but with different abundance profiles. In the case of prokaryotes, there are several explanations for cells being found on both filter fractions. Over the course of sampling, filter pores become clogged and cells <0.1 µm could be caught on the larger filter fraction. Additionally, many cells that are themselves less than 0.1 µm live in flocs or attach to (in)organic matter, making their net size too large to pass through the 0.2-µm fraction. Conversely, large cells (>0.2 µm) may be observed on the 0.1-µm fraction as eDNA after lysis. Though we saw the most abundance and diversity of viruses on the 0.1-µm fractions as expected, viruses identified in the 0.2-µm fractions could be indicative of active intracellular infections.

There are several plausible explanations for the stable prokaryotic community and the fluctuating viral community. To begin the deliberation, it may be that strong external triggers, such as seismic activity impacting groundwater geochemistry, only caught via long-term sampling might cause fluctuations in the organismal part of the community. Fluctuations in viral abundance may reflect a faster turnover of viruses compared to longer-living ("living" in the case of viruses) prokaryotes. Additionally, infected prokaryotic cells can produce viral burst sizes many orders of magnitude larger than the original cellular population (Edwards et al. 2021), resulting in visible viral fluctuations, while the prokaryotic community is hardly impacted. Continual flushing of the system through eruptions, as well as grazing by protists, may prevent the accumulation of viral particles over time, resulting in the observed dips in viral abundance.

Spacer-to-protospacer matching suggests diverse generalist viruses in the terrestrial subsurface

We identified spacers in 99 MAGs (of the 751 MAGs in total), which corresponded to 32 strain-level MAGs after dereplication (see previous section). Using CrisprCasFinder (Couvin et al. 2018) and metaCRAST (Moller and Liang 2017), a total of 11 839 quality-filtered spacers were extracted either directly from MAGs or from associated repeats in the reads. Of these spacers, 630 were matched to protospacers on viral scaffolds using blast. Overall, 6% of our spacers had protospacer matches, which is comparable to the global average of 7% (Shmakov et al. 2017). Out of 16 825 viral scaffolds, 3 835 had protospacer matches to a host, which generated 278 species-level clusters (vOTUs) with VIRIDIC (Moraru et al. 2020). The majority of the spacer-to-protospacer matches were associated with a *Gallionella spp*. MAG, which, surprisingly, was a low abundance community member (1.82%±0.13 SD relative abundance), and not the *Gallionella spp.* MAG dominating the community (30.82%±0.85 SD) (Table S3A, Fig. S). The second-highest number of spacer-to-protospacer matches originated in a CRISPR-Cas system of a MAG identified as Bacteroidetes *GWF2-32-17 spp.*, which contributed only 0.16% relative abundance (±0.03 SD) to the overall community. Further viral-host matches were identified across the community structure—from high-abundant MAGs identified as *Gallionella spp.* and *Thiobacillaceae* bacterium *LSR1* to extremely low-abundant MAGs of *Nitrospirae*, *Leptospira ogonesis*, and *Bacteroidetes*.

Based on spacer-to-protospacer matches, many vOTUs apparently infect hosts across family level boundaries. Analyses of infection histories might be confounded by conjugative transfer of CRISPR systems between phylogenetically distinct hosts and non-hosts in densely populated ecosystems (Hwang et al. 2023). While the Wallender Born surely harbors biofilms, overall the water discharged from the geyser has a very low biomass compared to dense biofilms investigated in the aforementioned study. This likely lessens the effect of non-host spacer acquisition, rendering the spacer-to-protospacer matches herein reliable for determining host-virus interactions. Our results therefore indicate a putative generalist life strategy of the viruses, with most commonly observed cross-infections between *Gallionellaceae* and *Burkholderiaceae*. Two families within Bacteroidales were found to share an infection history with many viruses that were also predicted to infect *Gallionellaceae*, *Burkholderiaceae*, and several members of the rare biosphere. Such broad host ranges can obviously benefit viruses in diverse microbial communities by enabling access to more potential hosts, i.e. viral replication factories (Chevallereau et al. 2022). These generalist viruses, however, may incur evolutionary and ecological costs by not being well-adapted to all hosts, and thus may have lower replication rates (Chevallereau

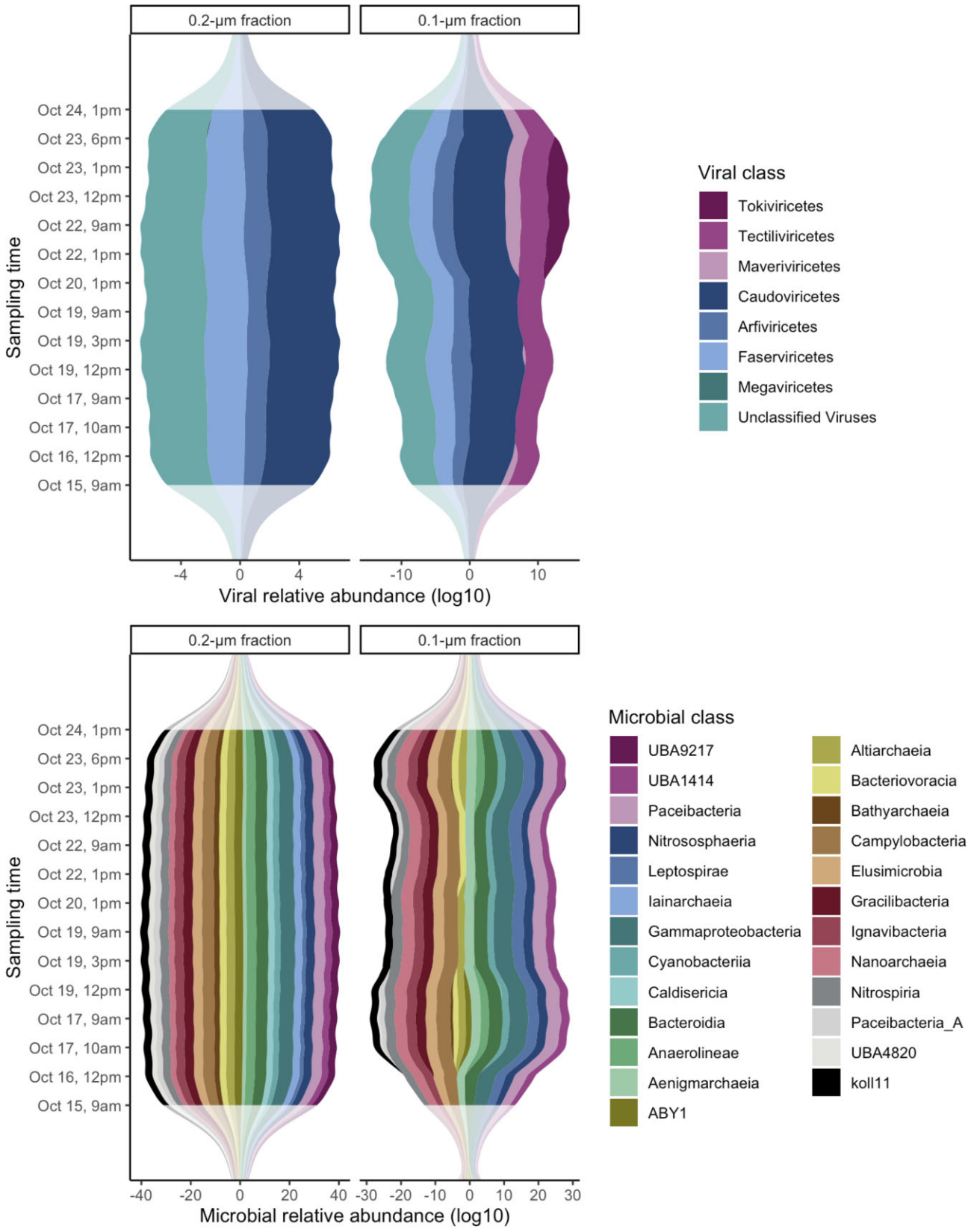

Figure 1. Stream plots visualizing major metagenome abundance trends of overall viral and prokaryotic communities in the two filter fractions. Abundance is represented by plotting the logarithm base 10-transformed, read-normalized coverage values for all genomes within the taxonomic classes. Gallionella spp. is classified as Gammaproteobacteria by the Genome Taxonomy Database, and is visible in this class above.

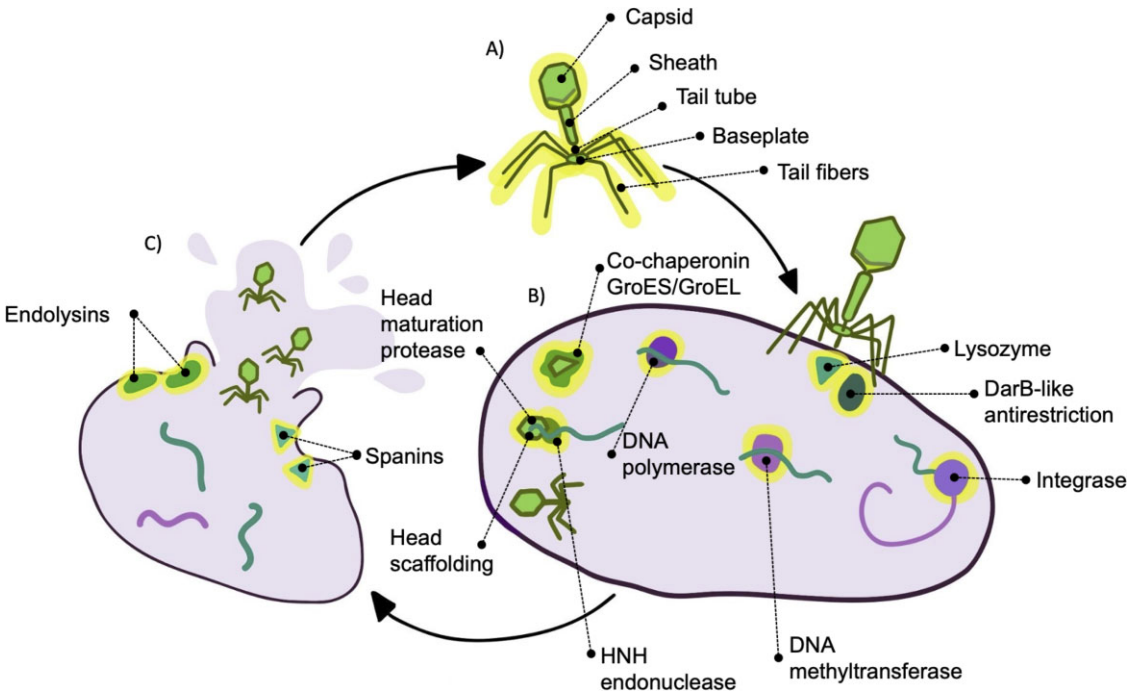

Figure 2. Conceptual sketch illustrating highly expressed viral proteins, indicated by text labels and glowing halos. Proteins required for different viral life cycle stages were identified in the proteomes. First, Panel A illustrates a free virion particle, with numerous structural proteins expressed. Next, Panel B illustrates cellactive proteins expressed during active infection. Proteins required for injecting viral genetic material into the host cell, replicating and manipulating nucleic material, and for assembling new virions were identified in the proteome. Lastly, Panel C shows proteins required for cell lysis that break down the cell wall and membrane were expressed.

et al. 2022). Phages reduce their host range when prolific hosts are available in the environment (Chevallereau et al. 2022). Generalist phages may be an indicator that the environment favors different organisms at different times, or that immunity of one host improves and necessitates the virus to target a new host, as may be the case with *Gallionellaceae* and *Burkholderiaceae*.

Though generalist viruses appear to play a large role in prokaryotic infection, we also observed putative specialist viruses. For example, we observed *Gallionellaceae* to be a family heavily targeted by viruses based on high spacer-to-protospacer matching, with many viruses infecting multiple genera within the family including *Gallionella spp* and *Sideroxydans spp.* These individual genera also appear to host specialist viruses, with several distinct vOTUs solely linked to one genus or the other (Table S4). These findings suggest diverse lifestyle strategies for subsurface viruses.

Viruses may play an important role in maintaining the community structure in groundwater discharged by Wallender Born, by targeting the most susceptible organisms in any given time window. Our observations lead us to believe this susceptibility varies enough at a strain level, that viral targeting of specific strains keeps their abundance low, allowing other strains to maintain high abundances. We investigated the possibility of kill-the-winner dynamics controlling the ecosystem, as has been hypothesized to control other subsurface environments (Holmfeldt et al. 2021). Kill-the-winner posits that fast-growing taxa are targeted by viruses, allowing for an overall balance in community structure to be maintained even with different growth rates (Thingstad and Lignell 1997). We used differences in median relative abundance of metagenome-based genome abundances versus metaproteome-based genome abundances as a proxy for metabolic activity, such that candidate "winners" would be disproportionately active in the proteome compared to their metagenome abundance (Fig. S5). We found that hosts were not significantly more metabolically active than non-hosts, suggesting that kill-the-winner does not appropriately describe the ecosystem, and other complex factors determine viral infection dynamics. As a result, we suggest "infect to keep in check" as an alternative model to describe viral control of this ecosystem, which posits that susceptible taxa carrying much of the viral infection load will be present in low abundances, allowing less susceptible taxa to dominate the ecosystem. This ecosystem violates a key assumption of kill-the-winner, that one phage has a single host and vice versa, which may play a role in this ecosystem not following strictly kill-the-winner dynamics, as we observe many generalist viruses infecting several hosts, as well as multiple viruses infecting single hosts. Highly abundant strains may have adapted their porins or other receptors to physically avoid viral infection, allowing them to maintain a large population size over the course of our 12-day sampling period, but possibly at the cost of reduced porin activity (Thingstad and Lignell 1997). In the case of the numerous *Gallionella spp.* strains found in this ecosystem, we observed CRISPR-Cas defense systems in all strains, but high-quality spacer matches to viral protospacers were only observed in the low abundance strains. High-abundance strains contained spacers with too much dissimilarity to viral protospacers to be included in our analysis, and this may be an indicator that the high-abundance strain was targeted by an ancestral virus, which has since mutated significantly and is no longer detectable in the environment.

Metaproteogenomics reveals many active stages of infection and abundant virion proteins

To investigate active viruses in the complex community with multiple infection histories, we leveraged previously published metaproteomics data (Figueroa-Gonzalez et al. 2023). Three metaproteomes were produced based on different pooling strate-

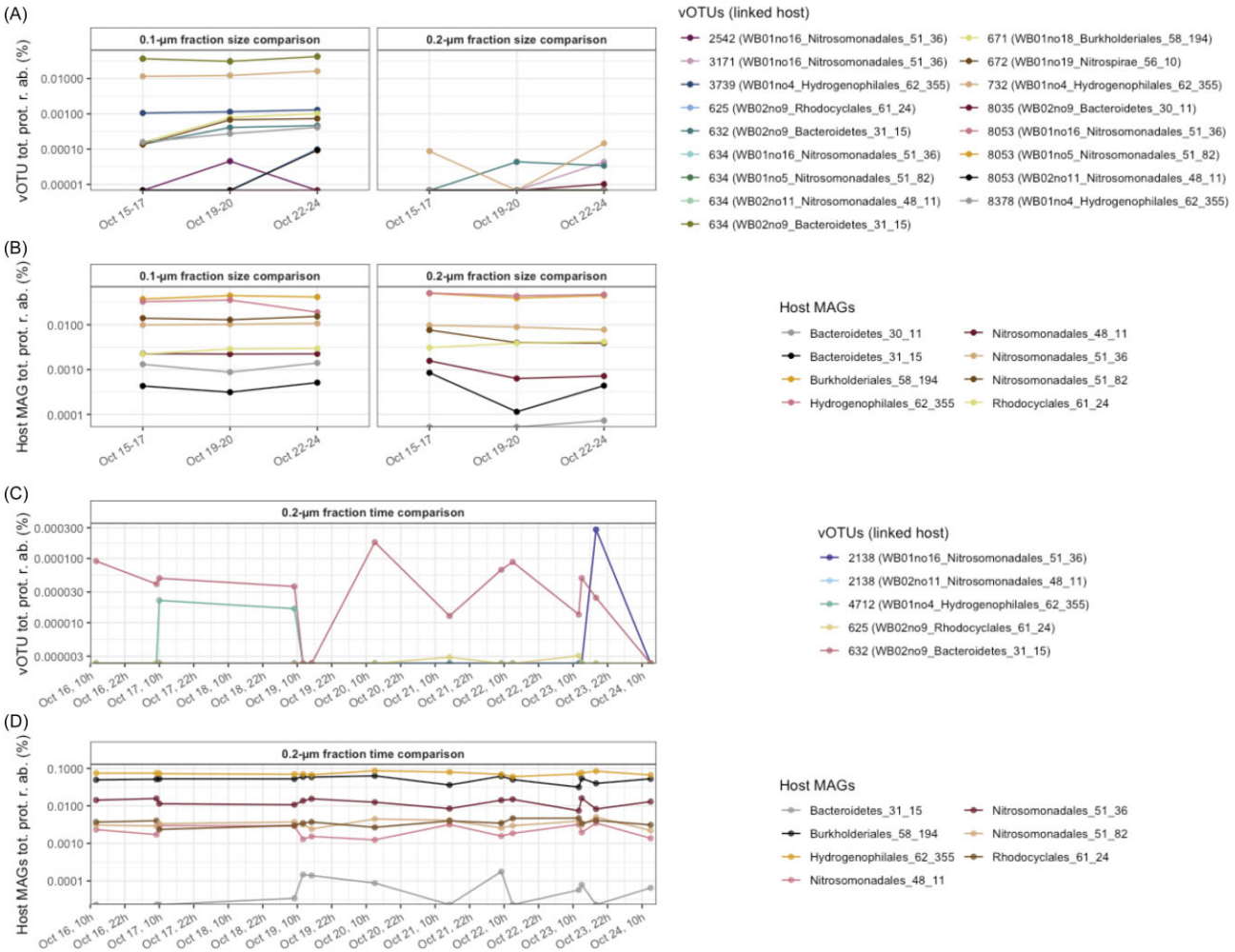

Figure 3. Percent relative abundance of normalized summed spectral counts of proteins from 14 active viruses and associated hosts. Panels A and B show changes in virus and host protein relative abundance respectively in size separated time series, demonstrating the lower presence of viral proteins in the 0.1-µm fractions. Panels C and D show changes in virus and host protein relative abundance respectively in a more time-resolved series in 0.2-µm fractions only, displaying dynamic variations in protein presence of linked hosts and viruses. Note that panels A and B have the same y-axis scale, but C and D differ in order to account for the much lower viral protein content on the 0.2-µm fractions. vOTU numbers in the legends are followed by the predicted host in parentheses in A and C, which can be matched to hosts in B and D.

gies on the biomass collected on the filters. The "bulk" metaproteome contains the proteins collected directly on 0.1-µm filters (a total of 1 sample). The "size" metaproteome contains pooled 0.1-µm filters from early, middle, and late stages of the time series, and likewise for 0.2-µm filters (a total of 6 samples in this metaproteome). The "time" metaproteome contains only proteins from 0.2-µm filters for 14 time points (total 14 samples).

Using these metaproteomes, we linked protein signatures to proteins of 103, 243, and 219 predicted viral genomes in bulk, size, and time samples, respectively. Most of the viral proteins were identified in the 0.1-µm fraction (220), while the 0.2-µm fraction included 114 viral proteins (both together constituting the size metaproteome). This finding is consistent with our expectations, since the 0.1-µm fraction is often referred to as the viral fraction because it catches many virions (Zhao et al. 2023). Bulk and time metaproteomes contained 110 and 231 proteins with matches to viral-encoded genes, respectively. Many of these identified proteins were uncharacterized or hypothetical (41 in the bulk, 37 in the size, 87 in time metaproteome), based on theUniref100-based FunTaxDB annotations (Bornemann et al. 2023). After PHROG HMM-based annotations, 7 proteins in the bulk, 23 proteins in the size, and 26 proteins in time metaproteome remained uncharacterized/hypothetical. Of the remaining unannotated proteins, 33 consisted of less than 2 000 amino acids and were thus of a reasonable size to predict structure with AlphaFold v2.2.0 (Jumper et al. 2021). Querying AlphaFold-predicted structures to DALI (Holm 2020) and Foldseek (Van Kempen et al. 2023) resulted in five additional annotations. Structure-based annotations included tail tube proteins, capsid proteins, and a DNA helicase. Overall, based on our annotations we showed expression of various viral proteins belonging to every life cycle stage, from host-takeover, DNA synthesis and replication proteins, to lysis and virion structural proteins (Fig. 2, Table S5).

Structural proteins were the primary contributors to the viral proteome, indicating abundant virions produced through lytic or chronic infections (Table S5). Total protein spectral counts contributed by virion proteins in the 0.1-µm filters were comparable or greater than the total protein spectral counts of their associated hosts found on both filter fractions, indicating significant nutrients acquired by cells are being partitioned into virions (Fig. 3). Given that virions require a higher nitrogen/phosphorus: carbon ratio than cellular counterparts (Jover et al. 2014), viruses may

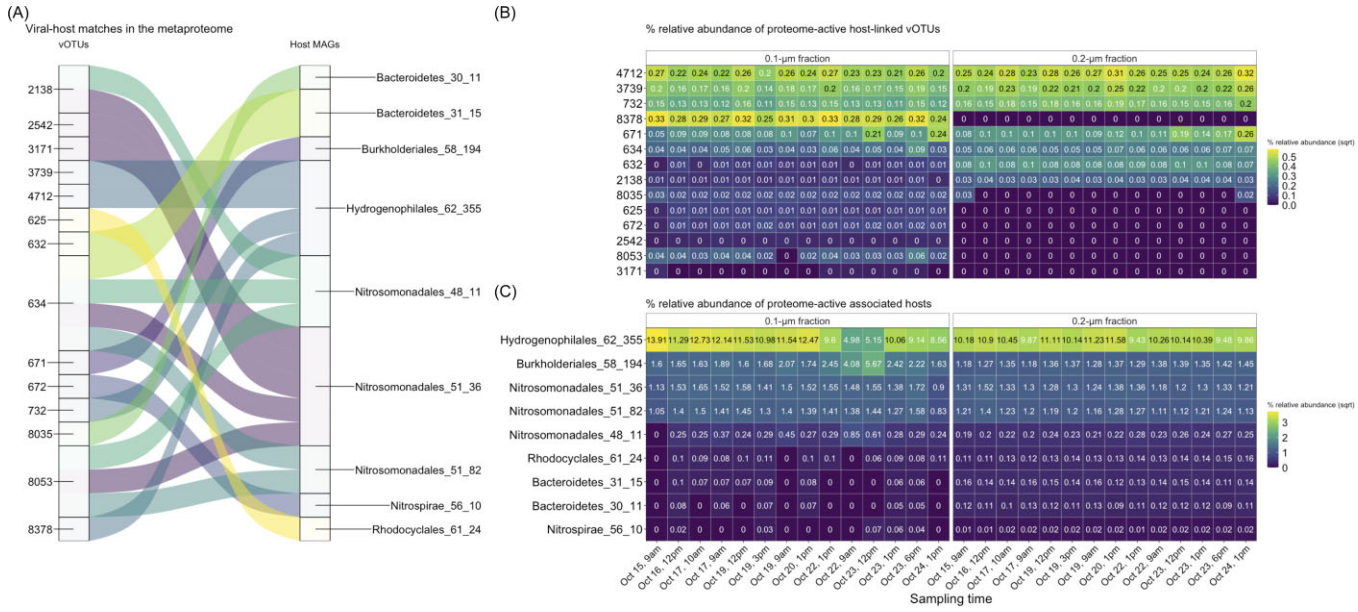

Figure 4. Panel A shows spacer-to-protospacer viral host links present in the metaproteome. Of all the viralhost associations, 14 viral genomes and nine hosts were found to have proteins in the expression profile. Each MAG was assigned a color, so that viruses infecting multiple MAGs can be identified by multiple colors flowing into a single vOTU. Percent relative abundance in the metagenome of associated vOTUs (Panel B) and hosts (Panel C) out of the entire viral and prokaryotic communities, respectively, are displayed across the time series.

sequester significant nutrients into their particles, likely having an important role in biogeochemical cycling in this environment.

Metaproteomics suggests viral control of *Gallionella* populations, active virion production, and presence of anti-restriction proteins in virions

Fourteen vOTUs with host matches were present in the metaproteomes along with the associated nine hosts (Fig. 4). Spectral counts between bulk, size, and time metaproteomes cannot be compared quantitatively due to different sample pooling. However, vOTU 634 and 732 contribute significant protein content within the 0.1-µm fractions of the size metaproteome (Fig. 3). vOTU 634 infects *Gallionella spp.* (Nitrosomondales, three organisms in Fig. 4) and Bacteroidetes bacterium GWF2-32-17 *spp.* (Bacteroidetes_31_15), while vOTU 732 infects *Thiobacillaceae* bacterium *LSR1* (Hydrogenphilales_62_355). Interestingly, vOTU 634, which makes the highest total protein contribution, has total spectral sums comparable to those of the associated *Gallionella spp.* hosts. While vOTU 634 does not have specific spacer matches to the most abundant *Gallionella spp.* strain, Nitrosomondales_51_490, it is observed to infect three separate strains (i.e. three different MAGs), suggesting some level of strain flexibility. Based on these read abundance-based observations, we conclude that the virus of vOTU 634 suppresses *Gallionella spp.* populations represented by three different MAGs helping Nitrosomondales_51_490 dominate the ecosystem. Since all MAGs encode the same metabolic functions (Figueroa-Gonzalez et al. 2023), vOTU 634 likely controls the population of *Gallionella spp.* in the aquifer accessible via geyser Wallender Born.

Annotation of the 14 viral genomes that show infection histories with hosts (Fig. 4) through VirClust's automatic query of multiple virus-specific databases allowed all but one of the expressed proteins to be confidently annotated. Expressed proteins were primarily structural proteins, present in both 0.1-µm and 0.2-µm filters (Fig. 5). Annotated intracellular proteins including recombinases, endonucleases, and ribonucleotide reductases were only present in 0.2-µm filters, except for the antirestriction enzyme which was only present in the 0.1-µm samples (Fig. 5). The 250-kbp *Caudoviricetes* linked to Bacteroidetes was found to express a RecA-like recombinase, HNH-like endonuclease, and capsid proteins. Taken together, this protein expression profile indicates active virion production via cleavage and packaging of viral genomes into proheads by HNH proteins, and maturation of capsid proteins.

A DarB-like antirestriction protein associated with a 60-kbp virus infecting *Gallionella spp.* was found in the 0.1-µm filter fraction. At first glance, an enzyme with intracellular activity would be expected only in the 0.2-µm fraction, but DarB is among several proteins known to be packaged into virion capsids. DarB is part of a six-protein antirestriction system that the P1 bacteriophage packages inside its capsid, so that this anti-host system can protect the phage genome at the first moment of host infection (Gonzales et al. 2022). Finding a DarB-like protein in the 0.1-µm filter fraction may indicate that packaging antirestriction proteins into the capsid is more widespread in environmental samples than previously recognized.

Viral genomes contribute putative auxiliary metabolic genes (AMGs) for heavy metal resistance, sulfate and phosphorus acquisition as well as transformation

In terms of the genomic potential of these active, host-linked viruses, several putative AMGs were identified with a putative role in improving host nutrient acquisition and cycling (Table S6). Phosphoadenosine phosphosulfate reductases were present in five of the 14 active viruses, and previous studies have identified phage homologues of this gene to participate in the assimilatory sulfate reduction pathway. This gene is more common in viruses in oligotrophic environments compared to other sulfur AMGs (Holmfeldt et al. 2021, Jacobson et al. 2021). AMGs supporting this pathway are crucial for sulfide incorporation into cysteine (Kieft et al. 2021). Moreover, vOTU 7 infecting Bacteroidetes en-

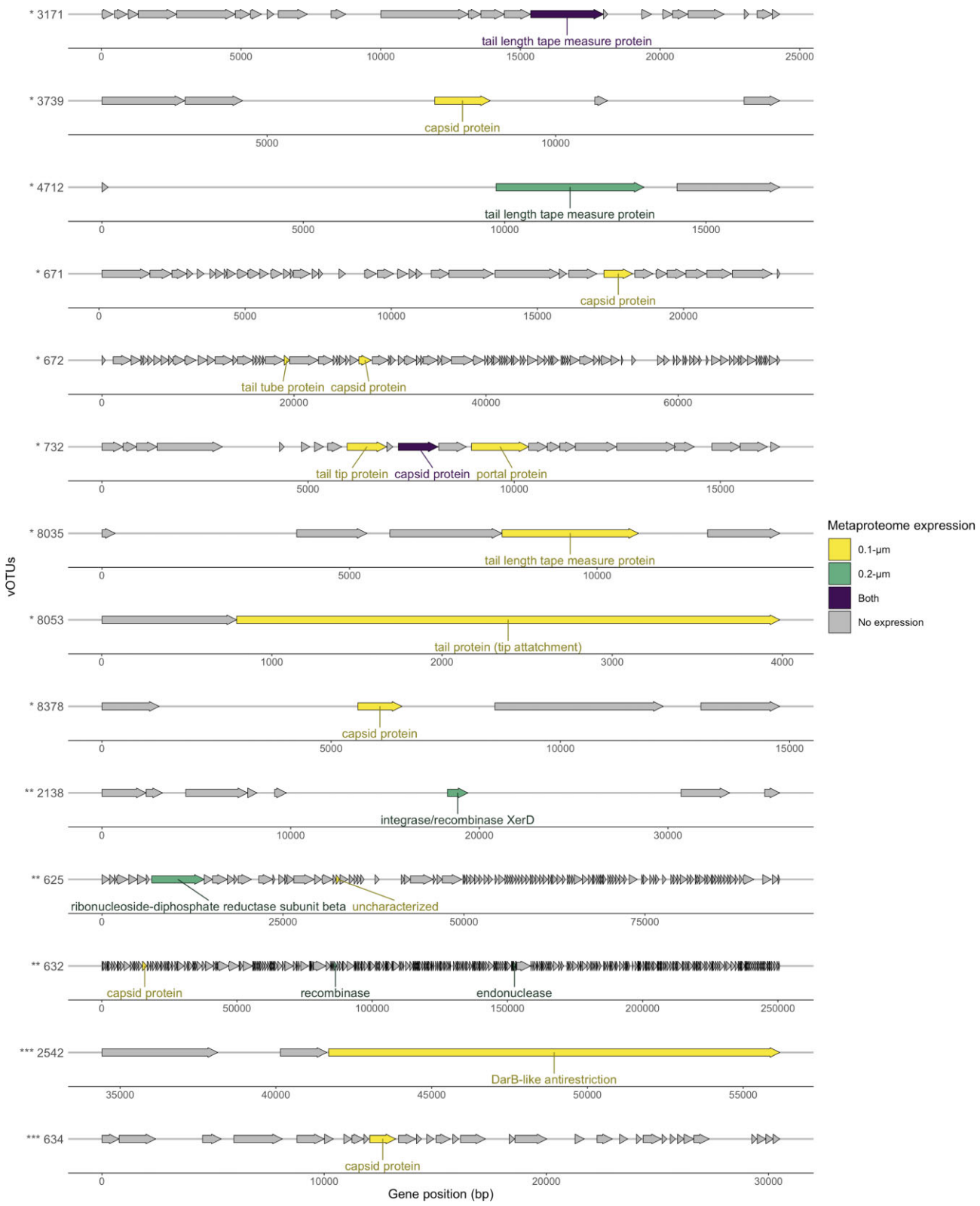

Figure 5. Predicted open reading frames on the 14 active viruses. Colored regions represent presence in the metaproteomes. A range of genome completeness are presented, indicated by asterisks (*low-quality, **medium-quality, ***high-quality; quality is defined by CheckV based on completeness scores).

codes a putative sulfate transport ATPase. These virus-encoded enzymes that support sulfur acquisition and assimilation have previously been hypothesized to support production of iron-sulfur proteins required for electron transport chains and redox processes (Jacobson et al. 2021).

In addition to sulfur metabolism, two of the active, host-linked viruses also possess putative AMGs involved in phosphate metabolism, namely a PhoH-like homologue detected on both viral genomes linked to Bacteroidetes. PhoH is a host derived phosphate-starvation inducible protein, which may enhance phosphorus uptake and transport in low-phosphorus environments (Yu et al. 2023). The corresponding gene product is not observed in the metaproteomes, so phages may only express this gene under specific conditions to reprogram the host to focus on phosphorus uptake, which is required for the extra DNA biosynthesis and NADPH (energy) production needed for phage replication. Overall, viruses have different stoichiometric compositions in comparison to cells, namely a lower carbon/nitrogen/phosphorus ratio (Jover et al. 2014); programmable increased phosphorus uptake may help compensate for the increased phosphorus requirement for producing virions.

Two of the viral genomes (infecting *Gallionella spp.* and Bacteroidetes) may also contribute to host heavy-metal resistance by carrying a tellurium resistance membrane protein (TerC homologue). Given that the Geyser Wallender Born is carbonated due to mantle degassing, tellurium may be entering the aquifer through degassing as a moderately volatile metal. Tellurium resistance genes have previously been reported on phage genomes isolated from mine drainage sites, and likely confer fitness to their hosts by protecting against tellurite which is highly toxic since it interferes with thiol inside of microbial cells (Zhang et al. 2023).

Conclusion

This study revealed an active viral community interacting with diverse hosts in the high-CO_2 Geyser Wallender Born. While viruses are nutrient-demanding members of this ecosystem as seen through a high expression of virion proteins, they encode genes to facilitate acquisition and transformation of nutrients by hosts. The viral metaproteome demonstrates viruses in many stages of infection, but is dominated by structural proteins, which likely renders viruses a significant nutrient sink in this oligotrophic environment.

Previous studies have observed viral-prokaryotic dynamics over long periods of time (i.e. many months), allowing the metagenomic surveys to observe prokaryotic community restructuring. The "halt and catch fire" strategy is suggested to be employed by organisms in environments where nutrients are scarce, where organisms must rapidly take advantage of sporadic nutrient influx to grow their population (Mehrshad et al. 2021). Adjacent to this, viruses interact with abundant organisms by facilitating slow motion "boom and burst" cycles, where booming organism growth is subject to heavy viral predation, thus recycling nutrients and making them available in the environment again (Holmfeldt et al. 2021). These life strategies describe the community restructuring and nutrient turnover in oligotrophic, subsurface environments over long periods, but are not appropriate to describe communities over small time scales. In our temporally-resolved metaproteomic survey in the geyser Wallender Born, we propose that a "infect to keep in check" strategy is more appropriate to describe the active, fluctuating viral infections that occur alongside a stable prokaryotic community over periods of days or weeks, with hosts that are not necessarily the winners that would be predicted by kill-the-winner. The "infect to keep in check" viral strategy may influence which organisms are dominant in the ecosystem over small time scales, thus controlling the primary active metabolic pathways and nutrient transformations in subsurface high-CO_2 environments.

Acknowledgements

We would like to thank Rashi Halder of the Luxembourg Centre for Systems Biomedicine (LCSB) for the careful sequencing of these low biomass samples, Ken Dreger for the excellent server administration and management, and Ines Pothmann, Maximiliane Ackers and Michelle Moll for the first class laboratory maintenance and logistics organization. We also thank Sebastian Grund for technical assistance in handling proteomics samples. We thank all former group members of the Group for Aquatic Microbial Ecology (GAME) for discussions.

Author contributions

T.L.V.B. and P.A.F.G. performed sampling and DNA extraction. T.L.V.B. took care of raw read processing and assembly. C.J.M., S.E., T.L.V.B., P.A.F.G., J.P., and J.S. contributed to the extensive metagenomic analyses through binning, CRISPR workflow supervision, viral prediction, and mapping. C.M. for viral workflow development and interpretation. A.T.-S., T.H., S.M., and D.B. performed protein extractions, proteomics, and interpretation. L.R. created scientific illustrations. A.J.P conceptualized the study. C.J.M. wrote and prepared the manuscript with support from A.J.P. and co-authors.

Supplementary data

Supplementary data is available at *FEMSML Journal* online.

Conflict of Interest: The authors declare no conflict of interest.

Funding

This study was funded by the Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen (Nachwuchsgruppe Dr. Alexander Probst, A.J.P), the NOVAC project of the German Research Foundation (grant number DFG PR1603/2–1, A.J.P), and Fulbright Germany (2020-21 Cohort, C.J.M). We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

References

Al-Shayeb B, Sachdeva R, Chen L-X *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* 2020;**578**:425–31.

Alneberg J, Bjarnason BS, de Bruijn I *et al.* Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014;**11**:1144–6.

Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.

Anantharaman K, Brown CT, Hug LA *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* 2016;**7**:13219.

Antipov D, Raiko M, Lapidus A *et al.* Metaviral SPAdes: assembly of viruses from metagenomic data. Robinson P (ed.). *Bioinformatics* 2020;**36**:4126–9.

Bar-On YM, Phillips R, Milo R. The biomass distribution on Earth. *Proc Natl Acad Sci USA* 2018;**115**:6506–11.

Bonn F, Bartel J, Büttner K *et al.* Picking vanished proteins from the void: how to collect and ship/share extremely dilute proteins in a reproducible and highly efficient manner. *Anal Chem* 2014;**86**:7421–7.

Bornemann TLV, Adam PS, Turzynski V *et al.* Genetic diversity in terrestrial subsurface ecosystems impacted by geological degassing. *Nat Commun* 2022;**13**:284.

Bornemann TLV, Esser SP, Stach TL *et al.* uBin: a manual refining tool for genomes from metagenomes. *Environ Microbiol* 2023;**25**:1077–83.

Brunson J. ggalluvial: layered grammar for alluvial plots. *J Open Source Softw* 2020;**5**:2017.

Burstein D, Sun CL, Brown CT *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat Commun* 2016;**7**:10613.

Camargo AP, Roux S, Schulz F *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2023. (28 May 2024, date last accessed).

Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Borgwardt K (ed.). *Bioinformatics* 2022;**38**:5315–6.

Chevallereau A, Pons BJ, van Houte S *et al.* Interactions between bacterial and phage communities in natural environments. *Nat Rev Micro* 2022;**20**:49–62.

Couvin D, Bernheim A, Toffano-Nioche C *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res* 2018;**46**:W246–51.

Daly RA, Roux S, Borton MA *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nat Microbiol* 2018;**4**:352–61.

Deusch S, Seifert J. Catching the tip of the iceberg—Evaluation of sample preparation protocols for metaproteomic studies of the rumen microbiota. *Proteomics* 2015;**15**:3590–5.

Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**:2460–1.

Edwards KF, Steward GF, Schvarcz CR. Making sense of virus size and the tradeoffs shaping viral fitness. Ostling A (ed.). *Ecol Lett* 2021;**24**:363–73.

Esser SP, Rahlff J, Zhao W *et al.* A predicted CRISPR-mediated symbiosis between uncultivated archaea. *Nat Microbiol* 2023. https://doi.org/10.1038/s41564-023-01439-2 (28 May 2024, date last accessed).

Figueroa-Gonzalez PA, Bornemann TLV, Hinzke T *et al.* Metaproteogenomics resolution of a high-CO_2 aquifer community suggests an active symbiotic lifestyle of groundwater Gracilibacteria. *bioRxiv* 2023. (28 May 2024, date last accessed).

Flemming HC, Wingender J, Szewzyk U *et al.* Biofilms: an emergent form of bacterial life. *Nat Rev Micro* 2016;**14**:563–75.

Gonzales MF, Piya DK, Koehler B *et al.* New insights into the structure and assembly of bacteriophage P1. *Viruses* 2022;**14**:678.

Guo J, Bolduc B, Zayed AA *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021;**9**:37.

Holm L, Laiho A, Törönen P *et al.* DALI shines a light on remote homologs: one hundred discoveries. *Protein Sci* 2023;**32**:e4519. https://doi.org/10.1002/pro.4519 (2 October 2023, date last accessed).

Holm L. Using Dali for protein structure comparison. In: Gáspári Z (ed.), *Structural Bioinformatics*. Vol **2112**. New York, NY: Springer US, 2020, 29–42.

Holmfeldt K, Nilsson E, Simone D *et al.* The Fennoscandian Shield deep terrestrial virosphere suggests slow motion 'boom and burst' cycles. *Commun Biol* 2021;**4**:307.

Hwang Y, Roux S, Coclet C *et al.* Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat Microbiol* 2023;**8**:946–57.

Hyatt D, Chen G-L, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf* 2010;**11**:119.

Jacobson TB, Callaghan MM, Amador-Noguez D. Hostile takeover: how viruses reprogram prokaryotic metabolism. *Annu Rev Microbiol* 2021;**75**:515–39.

Jover LF, Effler TC, Buchan A *et al.* The elemental composition of virus particles: implications for marine biogeochemical cycles. *Nat Rev Micro* 2014;**12**:519–28.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.

Kieft K, Breister AM, Huss P *et al.* Virus-associated organosulfur metabolism in human and environmental systems. *Cell Rep* 2021;**36**:109471.

Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;**8**:90.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**:357–9.

Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.

Mehrshad M, Lopez-Fernandez M, Sundh J *et al.* Energy efficiency and biological interactions define the core microbiome of deep oligotrophic groundwater. *Nat Commun* 2021;**12**:4253.

Moller AG, Liang C. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. *PeerJ* 2017;**5**:e3788.

Moraru C, Varsani A, Kropinski AM. VIRIDIC—A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses* 2020;**12**:1268.

Moraru C. VirClust—a tool for hierarchical clustering, core gene detection and annotation of (Prokaryotic) viruses. *Viruses* 2021;**13**:1007.

Nayfach S, Camargo AP, Schulz F *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021;**39**:578–85.

Nurk S, Meleshko D, Korobeynikov A *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;**27**:824–34.

Olm MR, Brown CT, Brooks B *et al.* dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* 2017;**11**:2864–8.

Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**:1043–55.

Perez-Riverol Y, Bai J, Bandla C *et al.* The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res* 2022;**50**:D543–52.

Probst AJ, Castelle CJ, Singh A *et al.* Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations: genomic resolution of a high-CO_2 subsurface community. *Environ Microbiol* 2017;**19**:459–74.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2022. https://www.R-project.org/ (1 May 2024, date last accessed).

Rahlff J, Turzynski V, Esser SP *et al.* Lytic archaeal viruses infect abundant primary producers in Earth's crust. *Nat Commun* 2021;**12**:4642.

Ren J, Song K, Deng C *et al.* Identifying viruses from metagenomic data using deep learning. *Quant Biol* 2020;**8**:64–77.

Robinson D. broom: an R package for converting statistical analysis objects into tidy data frames. 2014. https://doi.org/10.48550/ARXIV.1412.3565 (10 August 2023, date last accessed).

Shmakov SA, Sitnik V, Makarova KS *et al.* The CRISPR spacer space is dominated by sequences from species-specific mobilomes. Gilmore MS (ed.). *mBio* 2017;**8**:e01397–17.

Sieber CMK, Probst AJ, Sharrar A *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 2018;**3**:836–43.

Soares A, Edwards A, An D *et al.* A global perspective on bacterial diversity in the terrestrial deep subsurface. *Microbiology* 2023;**169**:001172. https://doi.org/10.1099/mic.0.001172 (1 October 2023, date last accessed).

Terzian P, Olo Ndela E, Galiez C *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics Bioinforma* 2021;**3**:lqab067.

The Genome Standards Consortium, Bowers RM, Kyrpides NC *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;**35**:725–31.

Thingstad T, Lignell R. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat Microb Ecol* 1997;**13**:19–27.

Turzynski V, Griesdorn L, Moraru C *et al.* Virus-host dynamics in archaeal groundwater biofilms and the associated bacterial community composition. *Viruses* 2023;**15**:910.

Van Kempen M, Kim SS, Tumescheit C *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2023;**42**:243–6. https://doi.org/10.1038/s41587-023-01773-0 (1 August 2023, date last accessed).

Varadi M, Anyango S, Deshpande M *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44.

Wickham H, Averick M, Bryan J *et al.* Welcome to the Tidyverse. *J Open Source Softw* 2019;**4**:1686.

Wickham H. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. 2016. Cham: Springer International Publishing : Imprint: Springer, 2016.

Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016;**32**:605–7.

Yu H, Xiong L, Li Y *et al.* Genetic diversity of virus auxiliary metabolism genes associated with phosphorus metabolism in Napahai plateau wetland. *Sci Rep* 2023;**13**:3250.

Zhang H, Huang J, Zeng W *et al.* Dissecting the metal resistance genes contributed by virome from mining-affected metal contaminated soils. *Front Environ Sci* 2023;**11**:1182673.

Zhao J, Wang Z, Li C *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* 2023;**86**:825–42.

Zybailov B, Mosley AL, Sardiu ME *et al.* Statistical analysis of membrane proteome expression changes in *saccharomyces c erevisiae*. *J Proteome Res* 2006;**5**:2339–47.

Received 22 December 2023; **revised** 5 April 2024; **accepted** 18 May 2024