Exploring the Archaeal Virosphere by Metagenomics

Yifan Zhou, Yongjie Wang, David Prangishvili, Mart Krupovic

Zhou et al., Methods Mol Biol. 2024;2732:1-22. doi: 10.1007/978-1-0716-3515-5_1.

Exploring the archaeal virosphere by metagenomics

Yifan Zhou[1,2], Yongjie Wang[3,4,5], David Prangishvili[1,6], Mart Krupovic[1]*

[1]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, 75015 Paris, France
[2]Sorbonne Université, Collège Doctoral, F-75005 Paris, France
[3]College of Food Science and Technology, Shanghai Ocean University, Shanghai, China
[4]Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China
[5]Laboratory of Quality and Safety Risk Assessment for Aquatic Products on Storage and Preservation (Shanghai), Ministry of Agriculture, China
[6]Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia.

* - for correspondence
E-mail: mart.krupovic@pasteur.fr

Abstract
During the past decade, environmental research has demonstrated that archaea are abundant and widespread in nature, and play important ecological roles at a global scale. Currently, however, the majority of archaeal lineages cannot be cultivated under laboratory conditions and are known exclusively or nearly exclusively through metagenomics. A similar trend extends to the archaeal virosphere, where isolated representatives are available for a handful of model archaeal virus-host systems. Viral metagenomics provides an alternative way to circumvent the limitations of culture-based virus discovery and offers insight into the diversity, distribution and environmental impact of uncultured archaeal viruses. Presently, metagenomics approaches have been successfully applied to exploring the viromes associated with various lineages of extremophilic and mesophilic archaea, including Asgard archaea (Asgardarchaeota), ANME-1 archaea (Methanophagales), thaumarchaea (Nitrososphaeria), altiarchaea (Altiarchaeota) and marine group II archaea (Poseidoniales). Here, we provide an overview of methods widely used in archaeal virus metagenomics, covering metavirome preparation, genome annotation, phylogenetic and phylogenomic analyses, and archaeal host assignment. We hope that this summary will contribute to further exploration and characterization of the enigmatic archaeal virome lurking in diverse environments.

Key words: archaeal viruses, Archaea, metagenomics, hyperthermophiles, major capsid protein, CRISPR spacers, host prediction, virome

1. Introduction

Recent advances in high-throughput genome sequencing and computational approaches have transformed our appreciation of the diversity, ubiquity and importance of archaea in natural environments [1-9]. Similar to bacteriophages [10-15], archaeal viruses represent one of the major factors controlling the diversity and metabolic activity of archaeal populations [16, 17]. Although culture-based approaches are revealing extraordinary morphological and genomic diversity of archaeal viruses isolated from extreme geothermal and hypersaline environments [18-23], only a handful of virus isolates infecting mesophilic archaea have been described thus far [24, 25], limiting our appreciation of their diversity and ecological impacts. Nevertheless, the culture-based virus discovery efforts are increasingly complemented by culture-independent metagenomic approaches. For instance, metagenomics has uncovered a number of family-level groups of viruses and mobile genetic elements associated with Asgard archaea [26-29], a prominent phylum Asgardarchaeota widely considered to represent the ancestors of eukaryotes [1, 2]; ANME-1 clade (order Methanophagales), a group of methane oxidizing archaea implicated in modulation of greenhouse gas emission [30]; methanogenic archaea [31, 32]; ubiquitous marine archaea of the order Poseidoniales [33-35] and ammonia oxidizing thaumarchaea [33, 36-38]; as well as Altiarchaeota, abundant primary producers in subsurface ecosystems [39]. These studies have provided precious insights into the parts of the archaeal virosphere which are currently inaccessible through classical culture-based techniques. In this chapter, we provide an overview of protocols and practices used in archaeal virus metagenomics, covering metavirome preparation, genome annotation, phylogenetic and phylogenomic analyses, abundance profiling and archaeal host assignment.

2. Materials

2.1 Bioinformatics tools

Software packages and bioinformatic tools commonly used for the analysis of archaeal virus genomes and proteins are listed in Table 1.

Table 1. Software and web servers

Tool	Available at	Function	Reference
Fastp	github.com/OpenGene/fastp	Reads quality control	[40]
Trimmomatic	github.com/usadellab/Trimmomatic	Reads quality control	[41]
Megahit	github.com/voutcn/megahit	Sequence assembly	[42]
metaSPAdes	cab.spbu.ru/software/spades	Sequence assembly	[43]
Blast+	ftp.ncbi.nlm.nih.gov/blast/executables	Sequence homology search	[44]
VirSorter2	github.com/jiarong/VirSorter2	Viral sequence identification	[45]
VIBRANT	github.com/AnantharamanLab/VIBRANT	Viral sequence identification	[46]
DeepVirFinder	github.com/jessieren/DeepVirFinder	Viral sequence identification	[47]
Cenote-Taker 2	github.com/mtisza1/Cenote-Taker2	Viral sequence identification	[48]
Geneious Prime	Biomatters, Inc.	Sequence extension and reads mapping	N/A
ContigExtender	github.com/dengzac/contig-extender	Sequence extension	[49]
CheckV	bitbucket.org/berkeleylab/CheckV	Virus genome completeness assessment	[50]
CRISPRCasFinder	crisprcas.i2bc.paris-saclay.fr/CrisprCasFinder/Index	CRISPR spacer extraction	[51]
CRISPRDetect	github.com/davidchyou/CRISPRDetect_2.4	CRISPR spacer extraction	[52]
CD-HIT	cd-hit.org	Sequence redundancy removal	[53]
tRNAscan-SE	github.com/UCSC-LoweLab/tRNAscan-SE	tRNA gene detection	[54]
Batch CD-Search	www.ncbi.nlm.nih.gov	Protein annotation	[55]
eggNOG-mapper	eggnog-mapper.embl.de	Protein annotation	[56, 57]
DRAM-v	github.com/WrightonLabCSU/DRAM	Viral AMGs detection	[58]
HostG	github.com/KennthShang/HostG	Host prediction	[59]
MArVD2	bitbucket.org/MAVERICLab/marvd2	Archaeal virus identification	N/A
WIsH	github.com/soedinglab/wish	Host prediction	[60]
PHISDetector	github.com/HIT-ImmunologyLab/PHISDetector	Host prediction	[61]
RaFAH	sourceforge.net/projects/rafah	Host prediction	[62]
iPHoP	bitbucket.org/srouxjgi/iphop	Host prediction	N/A
Pharokka	github.com/gbouras13/pharokka	Genome annotation	[63]
HHsearch	github.com/soedinglab/hh-suite	Genome annotation	[64]
VIRFAM	biodev.cea.fr/virfam	Genome annotation	[65]
T-Coffee	tcoffee.crg.eu	Sequence alignment	[66]
MUSCLE	www.drive5.com/muscle	Sequence alignment	[67]
PROMALS3D	prodata.swmed.edu/promals3d	Sequence alignment	[68]
trimAl	trimal.cgenomics.org/	Alignment trimming	[69]
PhyML	www.atgc-montpellier.fr/phyml	Phylogeny construction	[70]
FastTree	www.microbesonline.org/fasttree	Phylogeny construction	[71]
IQ-TREE	www.iqtree.org	Phylogeny construction	[72]
iTOL	itol.embl.de	Tree visualization	[73]
Evolview	www.evolgenius.info/evolview	Tree visualization	[74]
vConTACT2	bitbucket.org/MAVERICLab/vcontact2	Gene-sharing network construction	[75]
compareM	github.com/dparks1134/CompareM	comparative genomic analyses	N/A
Easyfig	easyfig.sourceforge.net	Genome map construction	[76]
Clinker	github.com/gamcil/clinker	Genome map construction	[77]
Cytoscape	cytoscape.org	Network visualization	[78]
ViPTree	www.genome.jp/viptree/	Comparative genomic analyses	[79]
R package	www.r-project.org/	Heatmap construction	N/A
Bowtie2	bowtie-bio.sourceforge.net/bowtie2	Reads mapping	[80]

3. Methods

Archaea typically represent an abundant or even dominant component of the microbial communities in geothermal (e.g., terrestrial hot spring), hypersaline (e.g., crystallizer pond) and certain marine (e.g., estuary, sediments) environments [81-85]. Not surprisingly, archaeal viruses are also prevalent in such archaea-dominated ecosystems. Thus, to exemplify the protocols for exploration of the archaeal virome, in this chapter, we focus on samples originating from hot springs, hypersaline ponds and marine environments. Generally, preparation of the viral DNA for archaeal virus metagenomics is not different from that used for other prokaryotic DNA viruses and, given that there are many excellent dedicated studies on this topic (e.g.,[86-91]), this part will be discussed only briefly. Notably, all currently known archaeal viruses contain DNA genomes [92], thus protocols for RNA virus discovery will not be considered. For environmental samples with low archaeal abundance, whenever possible, an archaeal enrichment step [93] might be considered prior to collecting the viral DNA for the downstream metagenomics studies.

3.1. Concentration of viral particles and DNA purification (adapted from [88-91, 94, 95])

For environmental samples with low cellular (and viral) density, large initial volumes, in the range of tens or hundreds of liters, are typically processed. For instance, in the case of marine samples, ~100 L of the water sample can be filtered through a 5 μm pore size filter to remove large inorganic particles and eukaryotic cells (Figure 1). Next, the prefiltered water sample is further passed through two successive

0.45 μm and 0.22 μm pore size filter-columns to remove the remaining cells, mainly archaea and bacteria. The virus particles can be concentrated by pumping the cell-free water through a 50 kDa cutoff tangential-flow filtration (TFF) cassette until the retentate reaches a volume of ~500 mL (*see* **Note 1** and **Note 2**). Finally, the viral preparation can be further concentrated to ~20 mL using 30 kDa cutoff ultrafiltration cups or ultracentrifugation.

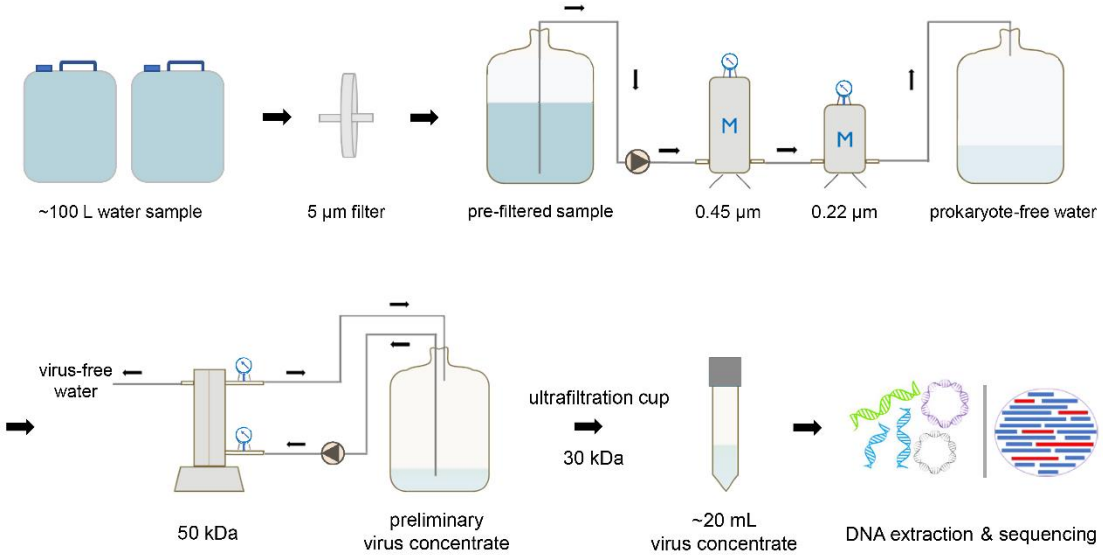

Figure 1. A workflow for the virus metagenome preparation.

Chemical flocculation is an alternative approach which has been successfully applied for concentration of viruses from seawater samples. This method depends on iron (III) chloride to precipitate viruses which are recovered by filtration onto large-pore size membranes and then resuspended using a buffer containing magnesium and a reductant (ascorbic acid or oxalic acid) at slightly acid pH [87, 96].

Next, total nucleic acids are extracted from the virus concentrate and prepared for metagenomic sequencing (*see* **Note 3** and **Note 4**). Generally, libraries are subjected to 150-bp, 250-bp, or 300-bp paired-end sequencing on Illumina platforms such as HiSeq X Ten, NovaSeq 6000 and MiSeq, respectively. Furthermore, the long-read sequencing, for example, provided by PacBio (Pacific Biosciences) or Nanopore (Oxford Nanopore Technologies), is becoming increasingly popular and affordable, providing single-molecule long-read sequences which can cover the entire virus genomes or considerably facilitate the assembly of large virus genomes [97].

3.2 Quality control of sequencing reads

Sequencing adapters and low-quality reads (<Q20) are trimmed off by using quality control tools such as Fastp and Trimmomatic to obtain a clean metagenomic dataset (*see* **Note 5**).
Example:

```
'fastp -i input_R1.fastq.gz -o clean_R1.fastq.gz -I input_R2.fastq.gz -O clean_R2.fastq.gz'
```

3.3 Sequence assembly

In the next step, clean reads are assembled to generate metagenomic contigs. The assembly is one of the most memory-consuming steps. A server with at least 200 GB of RAM is required. It is advisable to use more than one sequence assembler, as they produce slightly different results. Tools such as Megahit and metaSpades are applied to assemble the clean reads into metagenomic contigs. Contigs with sequence length over 1 kb are retained for the subsequent analyses.
Examples:

```
'megahit -1 clean_R1.fastq.gz -2 clean_R2.fastq.gz -o output_assembly -t 48'
```

```
'spades.py --meta -1 clean_R1.fastq.gz -2 clean_R2.fastq.gz -t 48 -m 360 -o output_assembly'
```

3.4 Identification of viral contigs

It is not possible to completely eliminate prokaryotic DNA from environmental metaviromes, because cellular DNA can be packaged into virus capsids and be transferred from one host to another by a process known as general transduction [98]. Alternatively, cellular DNA can be protected from DNase digestion within virus-sized extracellular membrane vesicles, which are known to be produced by many archaeal species [99-102]. To assess the extent of contamination with prokaryotic DNA, the 16S ribosomal RNA gene sequences can be downloaded from the SILVA database (latest release, www.arb-silva.de/) and used as queries to search against the assembled metagenomic sequences. Depending on the results, instead of using all sequences derived from the 'virus metagenome', it is recommended to implement a step of viral sequence sorting.

Tools such as VirSorter2, VIBRANT, DeepVirFinder or Cenote-Taker 2 (here and elsewhere, see Table 1 for references) can be used to identify and extract viral sequences from the assembled contigs (*see* **Note 6**).
Examples:

'virsorter run -w vir_outputVS2 -i Assembly_contigs.fasta --include-groups dsDNAphage --high-confidence-only'

'python run_cenote-taker2.py -c Assembly_contigs.fasta -r vir_outputCT2 -p True -m 256 -t 48 --exact_dtrs True'

3.5 Sequences extension and sequence quality check

The viral contigs can be extended by recruiting the sequencing reads that overlap with the edges of the de novo assembled contigs [103, 104]. To this end, all identified viral sequences can be pooled together and used as seed sequences to perform the reference assembly. Geneious Prime 'map to reference' function can be used to compare all the metagenomic reads to the seed sequences. If there are significant matches (i.e., ≥30 bp overlap and ≥95% overlap identity), the reads will be assembled to the corresponding seed sequence, yielding longer viral contigs. This procedure can be repeated until the seed sequence ceases to extend (*see* **Note 7**). Alternatively, sequence extension can be performed using ContigExtender.
Example:

'extender_wrapper.py --m1 clean_R1.fastq --m2 clean_R2.fastq --out VirContigExtension --min-overlap-length 30 --stop-length 55 --threads 48 vir_contigs.fasta'

'Repeat Finder' (Geneious Prime plugin) can be used to detect direct terminal repeats (DTR) and inverted terminal repeats (ITR) on the extended viral contigs. If the contig contains either DTR or ITR regions, the sequence can be considered to represent a complete circular or linear virus genome, respectively. Note that for formal classification, the International Committee on Taxonomy of Viruses (ICTV) only considers viruses with complete genome sequences [105] (*see* **Note 8**).

CheckV can also be used to estimate genome completeness by examining the terminal repeats and similarity of the viral contigs to related viral genomes.
Example:

'checkv end_to_end viral_contigs.fasta checkv_output -t 48 -d ~/checkv_database'

Note that due to its inherent dependency on viral reference genomes, CheckV shows inadequate performance when assessing the completeness of truly novel viral genomes which have only distant relatives or no relatives in the reference database.

3.6 Assignment of archaeal hosts

Arguably, the most challenging step in archaeal virus metagenomics is differentiation between viruses infecting bacteria and archaea and accurate assignment of viruses to their archaeal hosts. Below we introduce several host assignment approaches which have been previously applied to identify archaeal viruses in metagenomic datasets.

3.6.1 CRISPR spacer targeting

CRISPR-Cas is an adaptive immune system encoded by most archaeal species [106]. The spacer sequences stored in CRISPR arrays represent the immune memory of past encounters with foreign mobile genetic elements. Therefore, by matching CRISPR spacers to the corresponding protospacer sequences in the viral genomes, it is possible to identify archaeal virus-host pairs in metagenomic dataset. See an

example of archaeal CRISPR spacers matching an archaeal virus genome in Figure 2. This method is currently by far the most reliable among host prediction approaches. However, it should be noted that in the case of archaeal viruses, the accuracy of host prediction is typically limited to the level of family [107] or even order [30]. For instance, in the case of rudiviruses, viruses infecting *Saccharolobus* and *Metallosphaera* species were targeted by spacers from CRISPR arrays of *Metallosphaera* and *Saccharolobus*, respectively [107].

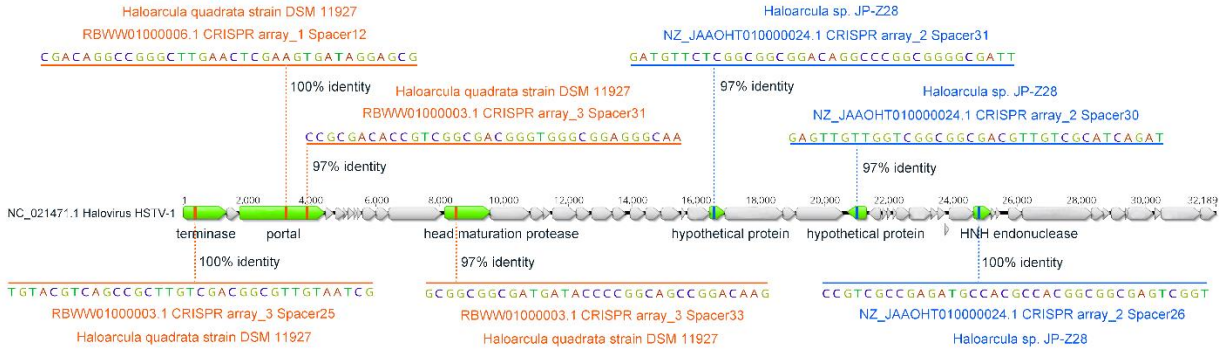

Figure 2. Seven CRISPR spacers from two *Haloarcula* species match the genome of Haloarcula sinaiiensis tailed virus 1 (HSTV-1).

To assemble the database of CRISPR spacers, all archaeal genomic sequences can be downloaded from the Genome Taxonomy Database (GTDB) or any other genome database. The retrieved archaeal genomes can then be analyzed using CRISPRDetect or CRISPRCasFinder to detect CRISPR arrays and extract the archaeal CRISPR spacer sequences.
Example:

'CRISPRCasFinder -in archaeal_seqs.fasta -out CRISPR_results -md 10 -t 15 -mr 25 -xr 55 -ms 25 -xs 55 -pm 0.7 -px 2.5 -s 55 -fl 120 -cpuM 48'

CD-HIT can be used to remove redundant spacers with 100% sequence identity cutoff. In order to maintain the targeting specificity, it is recommended to discard spacers with sequence length shorter than 25 bp.
Example:
'cd-hit -i archaeal_spacer.fasta -o unique_archaeal_spacers.fasta -c 1 -M 36000 -T 12 -d 0'

Next, a BLASTn database should be prepared from the extracted viral contigs:

'makeblastdb -i viral_contigs.fasta -out nt_db/viral_contigs -dbtype nucl'

The archaeal spacers can then be used as queries to search against the viral contig database (*see* **Note 9** and **Note 10**). Hits with at least 95% coverage and 95% identity are considered as valid protospacers and link the protospacer-containing viral contigs to the corresponding archaeal hosts.

'blastn -query unique_archaeal_spacers.fasta -db nt_db/viral_contigs -out ArSpacers_vs_VirContigs.xls -outfmt 6 -word_size 7 -dust no -qcov_hsp_perc 95 -perc_identity 95 -num_threads 48'

More than one spacer hit increases the reliability of the host prediction. The sequences of the CRISPR-targeted viral contigs should be retrieved for the subsequent analyses.

3.6.2 tRNA gene matching
tRNAs are an integral part of the cellular protein translation system. Certain archaeal viruses occasionally also encode host-derived tRNA genes [18, 108]. Therefore, the tRNA gene match between a virus and an archaeal genome is highly suggestive of a virus-host relationship. See an example of tRNA gene match between an archaeal virus and its host in Figure 3.

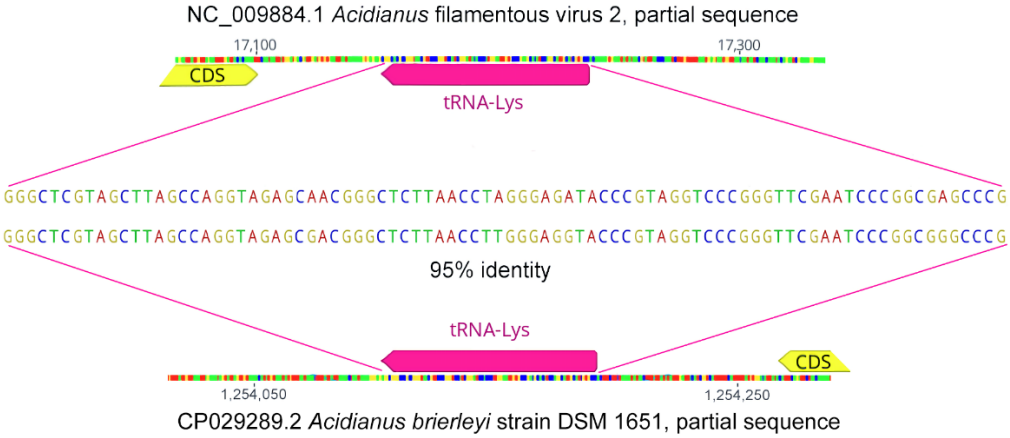

Figure 3. The tRNA-Lys gene encoded by the genome of Acidianus filamentous virus 2 shares 95% identity with a counterpart from the genome of *Acidianus brierleyi* strain DSM 1651.

The tRNA genes encoded by viral contigs can be predicted using tRNAscan-SE 2.0 with '-A' option (archaeal mode):

'tRNAscan-SE viral_contigs.fasta -A -o viral_tRNA.results -a viral_tRNA.fasta'

The viral tRNA gene sequences should then be extracted and used as queries in BLASTn search against the GTDB archaeal sequences (or other archaeal sequence database with available taxonomy information).

'makeblastdb -in archaeal_seqs.fasta -out nt_db/archaeal_seqs -dbtype nucl'

'blastn -query viral_tRNA.fasta -db nt_db/archaeal_seqs -out Vir_tRNA_vs_ArSeqs.xls -outfmt 6 -word_size 16 -dust no -qcov_hsp_perc 95 -perc_identity 95 -num_threads 48'

Hits with at least 95% coverage and 95% identity can be considered as genuine matches, providing host assignments for the tRNA-encoding viral contigs (*see* **Note 11**).

3.6.3 Matching of the viral and host attachment site(s)

Many archaeal viruses with circular dsDNA genomes encode integrases of the tyrosine recombinase superfamily and are commonly found as integrated proviruses within archaeal genomes [109-114]. The integration involves integrase-mediated homologous recombination between identical sequences of variable lengths on the viral and cellular genomes, known as the viral and archaeal attachment sites, attV and attA, respectively [115]. In the viral genome, the attV is typically located in the vicinity of or within the integrase gene, whereas the most common attA site occupies the 3'-proximal regions of the tRNA genes, although cases of integration into 5'-distal regions of tRNA genes, intergenic regions as well as protein-coding genes have been also described [112, 113]. The identity between the viral and archaeal attachment sites can be used for the host assignment for temperate viruses which integrate into the genome of their host [116]. See two examples of attachment site sharing between archaeal viruses and their respective hosts in Figure 4.

To identify the attachment sites within the viral genomes, the integrase genes have to be identified first. This can be done using batch protein annotation tools, such as Batch CD-Search or eggNOG-mapper v2. Once identified, the sequences of the viral integrase genes along with the upstream and downstream non-coding regions (≤ 500 bp) can be extracted and used as queries in BLASTn searches against archaeal sequences to detect the potential attachment sites.

'blastn -query integrase_region.fasta -db nt_db/archaeal_seqs -out Vir_integrase_vs_ArSeqs.xls -outfmt 6 -word_size 6 -dust no -perc_identity 95 -num_threads 48'

Hits with at least 95% coverage and 95% identity can be considered as genuine matches (*see* **Note 12**).

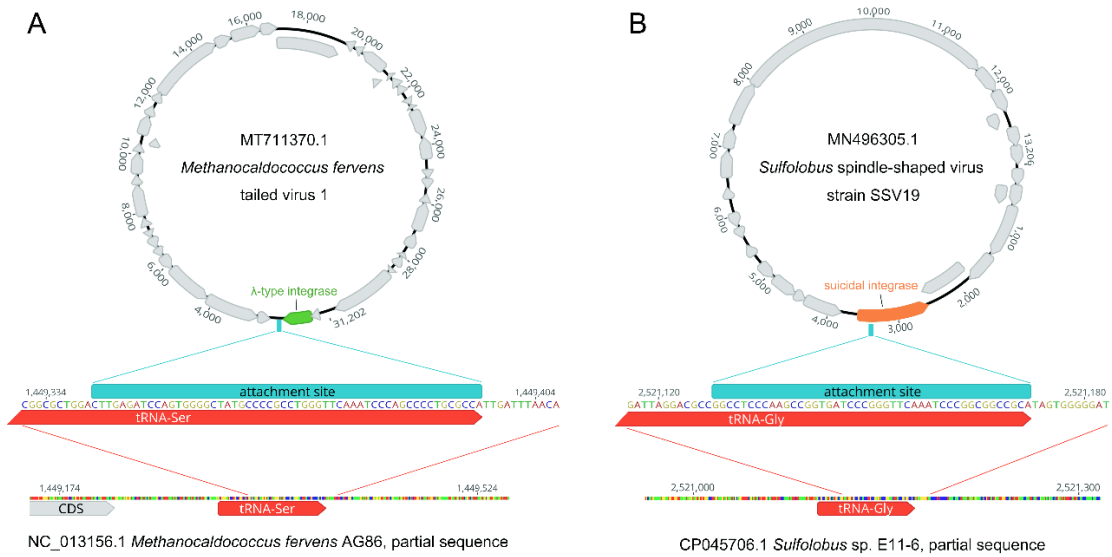

Figure 4. Examples of attachment sites located next to (A) and within (B) the integrase genes. A. The attachment site of Methanocaldococcus fervens tailed virus 1 is located next to the gene encoding the λ-type integrase of the tyrosine recombinase superfamily (green) and is identical to the 3'-proximal region of the host's tRNA-Ser gene (red). B. The attachment site of Sulfolobus spindle-shaped virus 19 (SSV19) is located within its integrase gene (orange) and is identical to the 3'-proximal region of the host's tRNA-Gly gene (red).

3.6.4 MCP homology searches

The ability to form virions distinguishes viruses from other types of mobile genetic elements and cellular organisms [117]. The major capsid proteins (MCP) are highly diverse, with some structurally related MCPs being widespread in viruses infecting hosts from different domains of life and others being specific to particular domains, including archaea [118]. Regardless, at the sequence level, MCPs are typically virus family or order specific. Therefore, homology searches using MCP as a signature protein has become one of the most commonly used approaches for the identification of new prokaryotic viruses in metagenomic dataset [33, 36, 37, 109, 119].

To assemble a database of archaeal virus MCPs, all archaeal virus protein sequences can be downloaded from the NCBI virus database (www.ncbi.nlm.nih.gov/labs/virus/vssi/#/) and the MCP sequences extracted. Once the MCP database is assembled, open reading frames in the metagenomic viral contigs can be predicted using the software Prodigal with '-p meta' option. The corresponding *in silico* translated protein sequences can then be used as BLASTp queries to search against the archaeal virus reference MCP database.

```
'prodigal -i viral_contigs.fasta -a ORFs_viral_contigs.fasta -p meta'

'makeblastdb -in Ref_ArVir_MCPs.fasta -out prot_db/ Ref_ArVir_MCPs -dbtype prot'

'blastp -query ORFs_viral_contigs.fasta -db prot_db/Ref_ArVir_MCPs -out
viral_contigs_vs_Ref_ArVir_MCPs.blastp -evalue 1e-5 -num_threads 48'
```

BLASTp hits with E-value $<10^{-5}$, ≥30% identity, ≥50% coverage, and ≥100 bit score can be considered significant. The corresponding viral contigs should be extracted for further analyses. More accurate host assignment can be inferred from phylogenetic analysis of the MCP sequences from the reference viruses and viruses assembled from metagenomes.

3.6.5 Host-specific AMGs

It is increasingly recognized that viruses can modulate the metabolic processes within the infected prokaryotes by expressing auxiliary metabolic genes (AMGs) [12]. In order to efficiently tinker with the host metabolism, the virus-encoded versions of the metabolic genes are usually recognizably similar to the host homologs [37, 38]. This property offers a possibility to trace potential virus-host pairs from environmental samples by analyzing the virus-encoded AMGs. Figure 5 shows an example of high sequence similarity between the ammonia monooxygenase subunit C (AmoC) encoded by a virus and an archaeon, *Nitrosopumilus maritimus*.

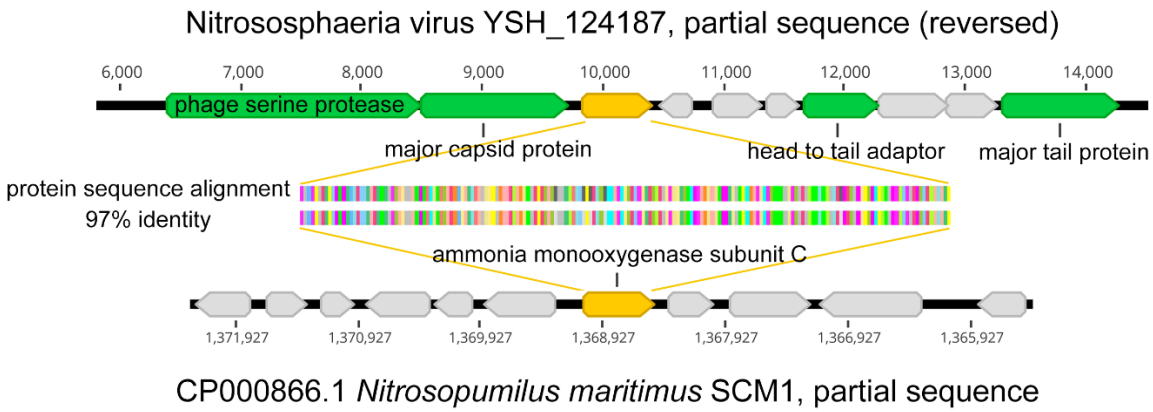

Figure 5. The ammonia monooxygenase subunit C encoded by the Nitrososphaeria virus YSH_124187 [33] shares 97% protein sequence identity with the counterpart encoded by *Nitrosopumilus maritimus* SCM1.

Viral AMGs can be predicted using the software DRAM-v, the mode of DRAM for viral sequences (*see* **Note 13**).

'DRAM-v.py annotate -i viral-combined-for-dramv.fa -v viral-affi-contigs-for-dramv.tab -o dramv_annotate --skip_trnascan --threads 48'

'DRAM-v.py distill -i dramv_annotate/annotations.tsv -o dramv-distill'

Viral AMGs should then be extracted and used as BLASTp queries to search against the archaeal proteome. The archaeal proteome database can be constructed using the following commands:

'prodigal -i archaeal_seqs.fasta -a ORFs_archaeal_seqs.fasta -p meta'

'makeblastdb -in ORFs_archaeal_seqs.fasta -out prot_db/archaeal_proteome'

Next, the viral AMGs should be assigned to specific archaeal species using the following command:

'blastp -query viral_AMGs.fasta -db prot_db/archaeal_proteome -out Vir_AMGs_vs_archaeal_proteome.blastp -evalue 1e-5 -num_threads 48'

Hits with at least 90% identity and 90% coverage are considered particularly indicative of a virus-host relationship, whereas hits with lower sequence identity should be evaluated more cautiously.

3.6.6 Host prediction tools (*see* **Note 14**)
A number of host prediction tools have been recently developed for prokaryotic viruses, e.g., HostG, WiSH, PHISDetector, RaFAH, iPHoP, and others. These tools rely on certain databases and yield results with host prediction with associated confidence scores.
Example (HostG):

'python run_Speed_up.py --contigs viral_contigs.fasta --len 1000 --t 0'

Example (iPHoP):

'iphop predict --fa_file viral_contigs.fasta --db_dir ~/iPHoP_db --out_dir iphop_prediction --num_threads 48'

In addition, the tool MArVD2 has been specifically developed for the identification of archaeal viruses from a set of viral contigs.

'MArVD2.py -i viral_contigs.fasta -o marvd_prediction --db-pvog ~/AllvogHMMprofiles.hmm --db-nr ~/nr.faa --db-accession2tax ~/prot.accession2taxid --marine-jackhmmer-db ~/pVOG_prots_ref_marine_pVOG.faa --viral-refseq-txt ~/viruses.txt --pvog-dir ~/pVOGs --cpu-count 48 --load-model ~/rf_model.pkl'

3.7. Genome annotation

All viral contigs can be automatically annotated using such tools as Batch CD-Search, eggNOG-mapper or Pharokka. However, for comprehensive functional annotation of the complete or near-complete archaeal virus genomes, it is advisable to use sensitive hidden Markov model (HMM) profile–profile comparisons with HHsearch v3.3.0 against the publicly available databases: CDD, Pfam, Protein Data Bank (PDB), uniprot_sprot_vir70, and PHROG. In the case of *Caudoviricetes,* viral structural proteins can be also predicted using VIRFAM.

3.8 Phylogenetic & phylogenomic analyses

Relationships between related viruses can be assessed using different methods, including single gene phylogenies, network or phylogenomic analyses as briefly detailed below.

3.8.1 *Phylogenetic analysis*

For the purpose of understanding the relationship between evolutionarily related viruses, a protein conserved in a given group of viruses has to be selected. The commonly used viral hallmark proteins include MCP, portal protein or large subunit of the terminase and other genome packaging ATPases. The protein sequences can be aligned using tools such as MUSCLE, T-COFFEE or PROMALS3D, followed by removal of highly divergent, uninformative positions, e.g., using trimAl. Maximum likelihood phylogenetic trees can be constructed using IQ-tree or PhyML. Both IQ-tree and PhyML can select the amino acid substitution model best fitting the given dataset. In the case of very large datasets, approximate maximum likelihood trees can be calculated using FastTree. The phylogenies can be annotated and visualized using iTOL v5 or Evolview v3.

3.8.2 *Gene-sharing networks*

The relationships between the identified archaeal viruses and other known prokaryotic viruses can be assessed using network analysis. vConTACT2 can be used to generate the gene-sharing networks with the latest prokaryotic virus database. Given that nodes (viral genomes) are connected only when they share three gene families, ideally, the length of input viral sequences should exceed 10 kb. Alternatively, the relationships between the viral genomes can be explored using bipartite networks, which include two types of nodes, viral genomes (type 1 nodes) connected through shared gene families (type 2 nodes) [120]. The resulting networks can be visualized using Cytoscape.

3.8.3 *Viral proteomic tree*

A virus proteomic tree is a dendrogram that represents global genomic relationships between viral sequences calculated from comparison of all protein sequences encoded by a given set of viruses. The viral proteomic tree generally corresponds well with the established virus taxonomy [33, 35]. The virus proteomic tree can be calculated for any given dataset using ViPTree, which can be either locally installed or run through the web server (*see* **Note 15**).

3.8.4 *Estimation of the orthologous protein fraction*

For classification purposes, it is useful to know the fraction of genes shared with other viruses (i.e., the degree of relatedness between viruses). For example, members of head-tailed archaeal viruses of the same family in the class *Caudoviricetes* generally share ~20–50% of orthologous genes, while viruses from different families share less than 10% [18]. The fraction of orthologous proteins can be estimated using the CompareM software toolkit with the following command:

```
'comparem aai_wf -e 0.0001 -p 30 -a 50 -c 4 ~/seqs output_results'
```

Based on the results (*see* **Note 16**), taxonomic classification can be tentatively assigned to the sequenced viruses according to the established taxon-specific demarcation criteria. It is important to note that depending on the virus group, different parameters can be adopted, resulting in different estimates of the orthologous fractions.

3.9 Abundance and distribution

To gain ecological and evolutionary insights into how archaeal viruses interact with their hosts and environments, it is important to explore the distribution and abundance of these viruses in different ecosystems. This information can be obtained by recruiting sequencing reads from metagenomes to the identified archaeal virus genomes, using read mappers, such as Bowtie2 or Geneious.
Example (Bowtie2):

```
'bowtie2-build ref_virus.fasta ref_virus'
```

```
'bowtie2 -x ref_virus -1 clean_R1.fastq -2 clean_R2.fastq -S results.sam'
```

The relative abundance of viruses in any particular sample or environment can be estimated by mapping the sequence reads from a metavirome to the reference genomes and expressed as **R**eads recruited **P**er **K**b of genome per **G**b of metagenome (RPKG). This way, the sequencing depth is normalized and is comparable to the distribution and abundance of the viruses in different environments.

4. Notes

1. The sample is passed through the TFF system by using a peristaltic pump and the pressure of the flow within the system should always be kept below 10 p.s.i. (~62 kPa) to avoid disruption of the viral particles [89].
2. One should consider the pH of the water sample to choose the corresponding filters. Ideally, all filtration operations should be conducted in a cold room to minimize the enzymatic degradation of virus particles and/or nucleic acids. Reservoirs, e.g., 500 mL flasks or bottles and 10-, 25- and 100-liter plastic containers, should be sterilized before use.
3. Filter the obtained viral concentrate through a 0.22 μm pore size filter to remove any potentially remaining prokaryotes. The absence of cells in the virus concentrate and the number of purified virus particles can be verified by fluorescence microscopy after staining of an aliquot of the sample with nucleic-acid-staining fluorescent dyes, such SYBR Green or DAPI (4′,6-diamidino-2-phenylindole), as described previously [121-123].
4. The majority of "free" DNA should be filtered out during the process of virus concentration. Nevertheless, an additional step of DNase treatment prior to extraction of the viral DNA from the virus particles can be performed to further reduce the contamination of the preparation with cellular DNA [124].
5. The result of reads quality control can be visualized as graphical and statistical reports by FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/).
6. For tools that are reference-based, it is necessary to apply the most updated databases. For more reliable identification of viral contigs, at least 2 different tools can be used in parallel.
7. It is advisable to manually check the mapping of assembled reads to avoid the possible mis-assembly.
8. For reporting sequences of uncultivated virus genomes, the genome quality is one of the requirements by the Minimum Information about an Uncultivated Virus Genome (MIUViG), with other qualifiers being information about virus origin, assembly tool, virus identification software, genome type, taxonomic classification, biogeographic distribution and *in silico* host prediction [125]. For recommendations on official classification of uncultivated viruses, see [126].
9. Since the default BLASTn parameters are not optimal for short sequences (e.g., 30 bp), it is recommended to use a word size of 7 and dust filtering turned off to identify the targets of CRISPR spacers [127, 128].
10. Although highly reliable, the CRISPR-targeting host assignment approach is dependent on the richness of the CRISPR spacer database available for a particular host organism. To further enrich the spacer database, one may consider extracting spacers from taxonomically unclassified sequences by identifying group-specific CRISPRs (e.g., Asgard-archaea-specific CRISPRs [26]). In parallel or alternatively, additional CRISPR spacers for the archaeal groups of interest can be recovered by amplifying CRISPR arrays with CRISPR-specific PCR primers from the environmental sample from which viral metagenome is being prepared (e.g., [129]).
11. The tRNA genes of unknown viruses (here, the extracted viral contigs) can be also searched against the tRNA genes of viruses for which the hosts have been assigned. The shared tRNA genes could indicate that the two viruses infected the same host.
12. Some attachment sites are as short as 8 bp [112], but such hits can hardly be considered significant without further validation. Thus, to avoid short random matches and to improve the specificity of this host assignment method, we recommend using matches with nucleotide alignment length not shorter than 25 bp.
13. Follow this Standard Operating Procedure www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-5qpvoyqebg4o/v3.
14. The performance of these host-prediction tools depends on the representation of the actual host organism within the initial training and reference dataset. Thus, it is important to make sure that the suspected host species were included in the training dataset. Otherwise, reasonable results are hardly to be expected. We recommend using these predictions only as supporting evidence to complement the host predictions inferred using other methods.
15. ViPTree web server also provides an informative genome map visualization useful for comparative genomics. Genome maps can be also compared using Easyfig or Clinker.
16. The output file of compareM can be converted into a matrix using tidyr (R package) and visualized using pheatmap (R package).

Acknowledgements
YW was supported by the National Natural Science Foundation of China (41376135, 31570112 and 41876195). The preparation of this chapter was supported by the Emergence(s) project MEMREMA from Ville de Paris to MK.

References

1. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH *et al*., (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature 541:353-358. doi:10.1038/nature21031
2. Liu Y, Makarova KS, Huang WC *et al*., (2021) Expanded diversity of Asgard archaea and their relationships with eukaryotes. Nature 593:553-557. doi:10.1038/s41586-021-03494-3
3. Baker BJ, De Anda V, Seitz KW *et al*., (2020) Diversity, ecology and evolution of Archaea. Nat Microbiol 5:887-900. doi:10.1038/s41564-020-0715-z
4. Arbab S, Ullah H, Khan MIU *et al*., (2022) Diversity and distribution of thermophilic microorganisms and their applications in biotechnology. Journal of basic microbiology 62:95-108. doi:10.1002/jobm.202100529
5. Danovaro R, Rastelli E, Corinaldesi C *et al*., (2017) Marine archaea and archaeal viruses under global change. F1000Res 6:1241. doi:10.12688/f1000research.11404.1
6. Offre P, Spang A, Schleper C (2013) Archaea in biogeochemical cycles. Annu Rev Microbiol 67:437-457. doi:10.1146/annurev-micro-092412-155614
7. Zou D, Liu H, Li M (2020) Community, Distribution, and Ecological Roles of Estuarine Archaea. Front Microbiol 11:2060. doi:10.3389/fmicb.2020.02060
8. Adam PS, Borrel G, Brochier-Armanet C *et al*., (2017) The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. ISME J 11:2407-2425. doi:10.1038/ismej.2017.122
9. Ogunrinola GA, Oyewale JO, Oshamika OO *et al*., (2020) The Human Microbiome and Its Impacts on Health. International journal of microbiology 2020:8045646. doi:10.1155/2020/8045646
10. Wigington CH, Sonderegger D, Brussaard CP *et al*., (2016) Re-examination of the relationship between marine virus and microbial cell abundances. Nat Microbiol 1:15024. doi:10.1038/nmicrobiol.2015.24
11. Suttle CA (2007) Marine viruses--major players in the global ecosystem. Nature Rev Microbiol 5:801-812. doi:10.1038/nrmicro1750
12. Breitbart M, Bonnain C, Malki K *et al*., (2018) Phage puppet masters of the marine microbial realm. Nat Microbiol 3:754-766. doi:10.1038/s41564-018-0166-y
13. Jurgensen SK, Roux S, Schwenck SM *et al*., (2022) Viral community analysis in a marine oxygen minimum zone indicates increased potential for viral manipulation of microbial physiological state. ISME J 16:972-982. doi:10.1038/s41396-021-01143-1
14. Kieft K, Zhou Z, Anderson RE *et al*., (2021) Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat Commun 12:3503. doi:10.1038/s41467-021-23698-5
15. Jacobson TB, Callaghan MM, Amador-Noguez D (2021) Hostile Takeover: How Viruses Reprogram Prokaryotic Metabolism. Annu Rev Microbiol 75:515-539. doi:10.1146/annurev-micro-060621-043448
16. Danovaro R, Dell'Anno A, Corinaldesi C *et al*., (2016) Virus-mediated archaeal hecatomb in the deep seafloor. Sci Adv 2:e1600492. doi:10.1126/sciadv.1600492
17. Lee S, Sieradzki ET, Nicol GW *et al*., (2023) Propagation of viral genomes by replicating ammonia-oxidising archaea during soil nitrification. ISME J 17:309-314 doi:10.1038/s41396-022-01341-5
18. Liu Y, Demina TA, Roux S *et al*., (2021) Diversity, taxonomy, and evolution of archaeal viruses of the class *Caudoviricetes*. PLoS Biol 19:e3001442. doi:10.1371/journal.pbio.3001442
19. Baquero DP, Liu Y, Wang F *et al*., (2020) Structure and assembly of archaeal viruses. Adv Virus Res 108:127-164. doi:10.1016/bs.aivir.2020.09.004
20. Munson-McGee JH, Snyder JC, Young MJ (2018) Archaeal Viruses from High-Temperature Environments. Genes 9 (3). doi:10.3390/genes9030128
21. Demina TA, Pietilä MK, Svirskaitė J *et al*., (2017) HCIV-1 and other tailless icosahedral internal membrane-containing viruses of the family Sphaerolipoviridae. Viruses 9 (2). doi:10.3390/v9020032
22. Aulitto M, Martinez-Alvarez L, Fusco S *et al*., (2022) Genomics, Transcriptomics, and Proteomics of SSV1 and Related Fusellovirus: A Minireview. Viruses 14 (10). doi:10.3390/v14102082
23. Luk AW, Williams TJ, Erdmann S *et al*., (2014) Viruses of haloarchaea. Life (Basel, Switzerland) 4:681-715. doi:10.3390/life4040681
24. Kim JG, Kim SJ, Cvirkaite-Krupovic V *et al*., (2019) Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. Proc Natl Acad Sci U S A 116:15645-15650. doi:10.1073/pnas.1905682116
25. Weidenbach K, Nickel L, Neve H *et al*., (2017) Methanosarcina Spherical Virus, a Novel Archaeal Lytic Virus Targeting Methanosarcina Strains. J Virol 91. doi:10.1128/jvi.00955-17
26. Medvedeva S, Sun J, Yutin N *et al*., (2022) Three families of Asgard archaeal viruses identified in

metagenome-assembled genomes. Nat Microbiol 7:962-973. doi:10.1038/s41564-022-01144-6
27. Rambo IM, Langwig MV, Leão P *et al*., (2022) Genomes of six viruses that infect Asgard archaea from deep-sea sediments. Nat Microbiol 7:953-961. doi:10.1038/s41564-022-01150-8
28. Tamarit D, Caceres EF, Krupovic M *et al*., (2022) A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal viruses. Nat Microbiol 7:948-952. doi:10.1038/s41564-022-01122-y
29. Wu F, Speth DR, Philosof A *et al*., (2022) Unique mobile elements and scalable gene flow at the prokaryote-eukaryote boundary revealed by circularized Asgard archaea genomes. Nat Microbiol 7:200-212. doi:10.1038/s41564-021-01039-y
30. Laso-Pérez R, Wu F, Crémière A *et al*., (2023) Evolutionary diversification of methanotrophic Ca. Methanophagales (ANME-1) and their expansive virome. Nat Microbiol 8:231-245. doi: 10.1038/s41564-022-01297-4.
31. Li R, Wang Y, Hu H *et al*., (2022) Metagenomic analysis reveals unexplored diversity of archaeal virome in the human gut. Nat Commun 13:7978. doi:10.1038/s41467-022-35735-y
32. Ngo VQH, Enault F, Midoux C *et al*., (2022) Diversity of novel archaeal viruses infecting methanogens discovered through coupling of stable isotope probing and metagenomics. Environ Microbiol 24:4853-4868. doi:10.1111/1462-2920.16120
33. Zhou Y, Zhou L, Yan S *et al*., (2023) Diverse viruses of marine archaea discovered using metagenomics. Environ Microbiol 25:367-382. doi: 10.1111/1462-2920.16287
34. Philosof A, Yutin N, Flores-Uribe J *et al*., (2017) Novel Abundant Oceanic Viruses of Uncultured Marine Group II Euryarchaeota. Curr Biol 27:1362-1368. doi:10.1016/j.cub.2017.03.052
35. Nishimura Y, Watai H, Honda T *et al*., (2017) Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. mSphere 2:e00359-16. doi:10.1128/mSphere.00359-16
36. López-Pérez M, Haro-Moreno JM, de la Torre JR *et al*., (2019) Novel Caudovirales associated with Marine Group I Thaumarchaeota assembled from metagenomes. Environ Microbiol 21:1980-1988. doi:10.1111/1462-2920.14462
37. Ahlgren NA, Fuchsman CA, Rocap G *et al*., (2019) Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. ISME J 13:618-631. doi:10.1038/s41396-018-0289-4
38. Roux S, Brum JR, Dutilh BE *et al*., (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature 537:689-693. doi:10.1038/nature19366
39. Rahlff J, Turzynski V, Esser SP *et al*., (2021) Lytic archaeal viruses infect abundant primary producers in Earth's crust. Nat Commun 12:4642. doi:10.1038/s41467-021-24803-4
40. Chen S, Zhou Y, Chen Y *et al*., (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884-i890. doi:10.1093/bioinformatics/bty560
41. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120. doi:10.1093/bioinformatics/btu170
42. Li D, Liu CM, Luo R *et al*., (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674-1676. doi:10.1093/bioinformatics/btv033
43. Nurk S, Meleshko D, Korobeynikov A *et al*., (2017) metaSPAdes: a new versatile metagenomic assembler. Genome research 27:824-834. doi:10.1101/gr.213959.116
44. Camacho C, Coulouris G, Avagyan V *et al*., (2009) BLAST+: architecture and applications. BMC bioinformatics 10:421. doi:10.1186/1471-2105-10-421
45. Guo J, Bolduc B, Zayed AA *et al*., (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9:37. doi:10.1186/s40168-020-00990-y
46. Kieft K, Zhou Z, Anantharaman K (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome 8:90. doi:10.1186/s40168-020-00867-0
47. Ren J, Song K, Deng C *et al*., (2020) Identifying viruses from metagenomic data using deep learning. Quantitative Biology 8:64-77. doi:10.1007/s40484-019-0187-4
48. Tisza MJ, Belford AK, Domínguez-Huerta G *et al*., (2021) Cenote-Taker 2 democratizes virus discovery and sequence annotation. Virus Evol 7:veaa100. doi:10.1093/ve/veaa100
49. Deng Z, Delwart E (2021) ContigExtender: a new approach to improving de novo sequence assembly for viral metagenomics data. BMC bioinformatics 22:119. doi:10.1186/s12859-021-04038-2
50. Nayfach S, Camargo AP, Schulz F *et al*., (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39:578-585. doi:10.1038/s41587-020-00774-7
51. Couvin D, Bernheim A, Toffano-Nioche C *et al*., (2018) CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res 46:W246-W251. doi:10.1093/nar/gky425
52. Biswas A, Staals RH, Morales SE *et al*., (2016) CRISPRDetect: A flexible algorithm to define CRISPR

arrays. BMC Genomics 17:356. doi:10.1186/s12864-016-2627-0
53. Fu L, Niu B, Zhu Z *et al*., (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150-3152. doi:10.1093/bioinformatics/bts565
54. Chan Patricia P, Lin Brian Y, Mak Allysia J *et al*., (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res 49:9077-9096. doi:10.1093/nar/gkab688
55. Lu S, Wang J, Chitsaz F *et al*., (2020) CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res 48:D265-d268. doi:10.1093/nar/gkz991
56. Cantalapiedra CP, Hernández-Plaza A, Letunic I *et al*., (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825-5829. doi:10.1093/molbev/msab293
57. Huerta-Cepas J, Szklarczyk D, Heller D *et al*., (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 47:D309-D314. doi:10.1093/nar/gky1085
58. Shaffer M, Borton MA, McGivern BB *et al*., (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res 48:8883-8900. doi:10.1093/nar/gkaa621
59. Shang J, Sun Y (2021) Predicting the hosts of prokaryotic viruses using GCN-based semi-supervised learning. BMC Biol 19:250. doi:10.1186/s12915-021-01180-4
60. Galiez C, Siebert M, Enault F *et al*., (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. Bioinformatics 33:3113-3114. doi:10.1093/bioinformatics/btx383
61. Zhou F, Gan R, Zhang F *et al*., (2022) PHISDetector: A Tool to Detect Diverse In Silico Phage-host Interaction Signals for Virome Studies. Genomics, proteomics & bioinformatics 20:508-523. doi:10.1016/j.gpb.2022.02.003
62. Coutinho FH, Zaragoza-Solas A, López-Pérez M *et al*., (2021) RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. Patterns 2:100274. doi:https://doi.org/10.1016/j.patter.2021.100274
63. Bouras G, Nepal R, Houtak G *et al*., (2023) Pharokka: a fast scalable bacteriophage annotation tool. Bioinformatics 39. doi:10.1093/bioinformatics/btac776
64. Steinegger M, Meier M, Mirdita M *et al*., (2019) HH-suite3 for fast remote homology detection and deep protein annotation. BMC bioinformatics 20:473. doi:10.1186/s12859-019-3019-7
65. Lopes A, Tavares P, Petit MA *et al*., (2014) Automated classification of tailed bacteriophages according to their neck organization. BMC Genomics 15:1027. doi:10.1186/1471-2164-15-1027
66. Di Tommaso P, Moretti S, Xenarios I *et al*., (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 39:W13-17. doi:10.1093/nar/gkr245
67. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. doi:10.1186/1471-2105-5-113
68. Pei J, Kim BH, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295-2300. doi:10.1093/nar/gkn072
69. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972-1973. doi:10.1093/bioinformatics/btp348
70. Guindon S, Dufayard JF, Lefort V *et al*., (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307-321. doi:10.1093/sysbio/syq010
71. Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one 5:e9490. doi:10.1371/journal.pone.0009490
72. Nguyen LT, Schmidt HA, von Haeseler A *et al*., (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-274. doi:10.1093/molbev/msu300
73. Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293-W296. doi:10.1093/nar/gkab301
74. Subramanian B, Gao S, Lercher MJ *et al*., (2019) Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. Nucleic Acids Res 47:W270-W275. doi:10.1093/nar/gkz357
75. Bin Jang H, Bolduc B, Zablocki O *et al*., (2019) Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37:632-639. doi:10.1038/s41587-019-0100-8
76. Sullivan MJ, Petty NK, Beatson SA (2011) Easyfig: a genome comparison visualizer. Bioinformatics 27:1009-1010. doi:10.1093/bioinformatics/btr039
77. Gilchrist CLM, Chooi YH (2021) Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. Bioinformatics. doi:10.1093/bioinformatics/btab007

78. Shannon P, Markiel A, Ozier O *et al*., (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498-2504. doi:10.1101/gr.1239303
79. Nishimura Y, Yoshida T, Kuronishi M *et al*., (2017) ViPTree: the viral proteomic tree server. Bioinformatics 33:2379-2380. doi:10.1093/bioinformatics/btx157
80. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357-359. doi:10.1038/nmeth.1923
81. Belilla J, Moreira D, Jardillier L *et al*., (2019) Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area. Nat Ecol Evol 3:1552-1561. doi:10.1038/s41559-019-1005-0
82. Xie W, Luo H, Murugapiran SK *et al*., (2018) Localized high abundance of Marine Group II archaea in the subtropical Pearl River Estuary: implications for their niche adaptation. Environ Microbiol 20:734-754. doi:10.1111/1462-2920.14004
83. Inskeep WP, Rusch DB, Jay ZJ *et al*., (2010) Metagenomes from high-temperature chemotrophic systems reveal geochemical controls on microbial community structure and function. PloS one 5:e9773. doi:10.1371/journal.pone.0009773
84. Kambourova M, Tomova I, Boyadzhieva I *et al*., (2016) Unusually High Archaeal Diversity in a Crystallizer Pond, Pomorie Salterns, Bulgaria, Revealed by Phylogenetic Analysis. Archaea (Vancouver, BC) 2016:7459679. doi:10.1155/2016/7459679
85. Oren A (2020) The microbiology of red brines. Adv Appl Microbiol 113:57-110. doi:10.1016/bs.aambs.2020.07.003
86. Hurwitz BL, Deng L, Poulos BT *et al*., (2013) Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. Environ Microbiol 15:1428-1440. doi:10.1111/j.1462-2920.2012.02836.x
87. John SG, Mendez CB, Deng L *et al*., (2011) A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ Microbiol Rep 3:195-202. doi:10.1111/j.1758-2229.2010.00208.x
88. Santos F, Yarza P, Parro V *et al*., (2010) The metavirome of a hypersaline environment. Environ Microbiol 12:2965-2976. doi:10.1111/j.1462-2920.2010.02273.x
89. Thurber RV, Haynes M, Breitbart M *et al*., (2009) Laboratory procedures to generate viral metagenomes. Nat Protocols 4:470-483. doi:10.1038/nprot.2009.10
90. Zablocki O, van Zyl LJ, Kirby B *et al*., (2017) Diversity of dsDNA Viruses in a South African Hot Spring Assessed by Metagenomics and Microscopy. Viruses 9. doi:10.3390/v9110348
91. Wu S, Zhou L, Zhou Y *et al*., (2020) Diverse and unique viruses discovered in the surface water of the East China Sea. BMC Genomics 21:441. doi:10.1186/s12864-020-06861-y
92. Koonin EV, Krupovic M, Agol VI (2021) The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? Microbiol Mol Biol Rev 85:e0005321. doi:10.1128/mmbr.00053-21
93. Liu Y, Brandt D, Ishino S *et al*., (2019) New archaeal viruses discovered by metagenomic analysis of viral communities in enrichment cultures. Environ Microbiol 21:2002-2014. doi:10.1111/1462-2920.14479
94. Adriaenssens EM, van Zyl LJ, Cowan DA *et al*., (2016) Metaviromics of Namib Desert Salt Pans: A Novel Lineage of Haloarchaeal Salterproviruses and a Rich Source of ssDNA Viruses. Viruses 8. doi:10.3390/v8010014
95. Schoenfeld T, Patterson M, Richardson PM *et al*., (2008) Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol 74:4164-4174. doi:10.1128/aem.02598-07
96. Poulos BT, John SG, Sullivan MB (2018) Iron Chloride Flocculation of Bacteriophages from Seawater. Methods Mol Biol 1681:49-57. doi:10.1007/978-1-4939-7343-9_4
97. Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics, proteomics & bioinformatics 13:278-289. doi:10.1016/j.gpb.2015.08.002
98. Chiang YN, Penadés JR, Chen J (2019) Genetic transduction by phages and chromosomal islands: The new and noncanonical. PLoS pathogens 15:e1007878. doi:10.1371/journal.ppat.1007878
99. Liu J, Soler N, Gorlas A *et al*., (2021) Extracellular membrane vesicles and nanotubes in Archaea. microLife 2:uqab007. doi:10.1093/femsml/uqab007
100. Gaudin M, Krupovic M, Marguet E *et al*., (2014) Extracellular membrane vesicles harbouring viral genomes. Environ Microbiol 16:1167-1175. doi:10.1111/1462-2920.12235
101. Choi DH, Kwon YM, Chiura HX *et al*., (2015) Extracellular Vesicles of the Hyperthermophilic Archaeon "Thermococcus onnurineus" NA1T. Appl Environ Microbiol 81:4591-4599. doi:10.1128/aem.00428-15
102. Liu J, Cvirkaite-Krupovic V, Commere PH *et al*., (2021) Archaeal extracellular vesicles are produced in an ESCRT-dependent manner and promote gene transfer and nutrient cycling in extreme environments. ISME J 15:2892-2905. doi:10.1038/s41396-021-00984-0
103. Simpson JT, Pop M (2015) The Theory and Practice of Genome Sequence Assembly. Annu Rev

Genomics Hum Genet 16:153-172. doi:10.1146/annurev-genom-090314-050032
104. Li Z, Chen Y, Mu D *et al*., (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Briefings in functional genomics 11:25-37. doi:10.1093/bfgp/elr035
105. Simmonds P, Adams MJ, Benko M *et al*., (2017) Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol 15:161-168. doi:10.1038/nrmicro.2016.177
106. Makarova KS, Wolf YI, Iranzo J *et al*., (2020) Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat Rev Microbiol 18:67-83. doi:10.1038/s41579-019-0299-x
107. Baquero DP, Contursi P, Piochi M *et al*., (2020) New virus isolates from Italian hydrothermal environments underscore the biogeographic pattern in archaeal virus communities. ISME J 14:1821-1833. doi:10.1038/s41396-020-0653-z
108. Sencilo A, Jacobs-Sera D, Russell DA *et al*., (2013) Snapshot of haloarchaeal tailed virus genomes. RNA Biol 10:803-816. doi:10.4161/rna.24045
109. Krupovic M, Forterre P, Bamford DH (2010) Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol 397:144-160. doi:10.1016/j.jmb.2010.01.037
110. Held NL, Whitaker RJ (2009) Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ Microbiol 11:457-466. doi:10.1111/j.1462-2920.2008.01784.x
111. Medvedeva S, Brandt D, Cvirkaite-Krupovic V *et al*., (2021) New insights into the diversity and evolution of the archaeal mobilome from three complete genomes of Saccharolobus shibatae. Environ Microbiol 23:4612-4630. doi:10.1111/1462-2920.15654
112. Krupovic M, Makarova KS, Wolf YI *et al*., (2019) Integrated mobile genetic elements in Thaumarchaeota. Environ Microbiol 21:2056-2078. doi:10.1111/1462-2920.14564
113. Krupovic M, Bamford DH (2008) Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology 375:292-300. doi:10.1016/j.virol.2008.01.043
114. Wang J, Liu Y, Liu Y *et al*., (2018) A novel family of tyrosine integrases encoded by the temperate pleolipovirus SNJ2. Nucleic Acids Res 46:2521-2536. doi:10.1093/nar/gky005
115. Badel C, Da Cunha V, Oberto J (2021) Archaeal tyrosine recombinases. FEMS microbiology reviews 45. doi:10.1093/femsre/fuab004
116. Mizuno CM, Rodriguez-Valera F, Kimes NE *et al*., (2013) Expanding the marine virosphere using metagenomics. PLoS Genetics 9:e1003987. doi:10.1371/journal.pgen.1003987
117. Koonin EV, Dolja VV, Krupovic M *et al*., (2021) Viruses Defined by the Position of the Virosphere within the Replicator Space. Microbiol Mol Biol Rev 85:e0019320. doi:10.1128/mmbr.00193-20
118. Krupovic M, Koonin EV (2017) Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A 114:E2401-e2410. doi:10.1073/pnas.1621061114
119. Zhou J, Zhang W, Yan S *et al*., (2013) Diversity of virophages in metagenomic data sets. Journal of virology 87:4225-4236. doi:10.1128/JVI.03398-12
120. Iranzo J, Koonin EV, Prangishvili D *et al*., (2016) Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. J Virol 90:11043-11055. doi:10.1128/jvi.01622-16
121. Patel A, Noble RT, Steele JA *et al*., (2007) Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. Nat Protoc 2:269-276. doi:10.1038/nprot.2007.6
122. Antón J, Llobet-Brossa E, Rodríguez-Valera F *et al*., (1999) Fluorescence in situ hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. Environ Microbiol 1:517-523. doi:10.1046/j.1462-2920.1999.00065.x
123. Rachel TN, Jed AF (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquatic Microbial Ecology 14:113-118
124. Allander T, Emerson SU, Engle RE *et al*., (2001) A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species. Proc Natl Acad Sci U S A 98:11609-11614. doi:10.1073/pnas.211424698
125. Roux S, Adriaenssens EM, Dutilh BE *et al*., (2019) Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol 37:29-37. doi:10.1038/nbt.4306
126. Dutilh BE, Varsani A, Tong Y *et al*., (2021) Perspective on taxonomic classification of uncultivated viruses. Curr Opin Virol 51:207-215. doi:10.1016/j.coviro.2021.10.011
127. Biswas A, Gagnon JN, Brouns SJ *et al*., (2013) CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. RNA Biol 10:817-827. doi:10.4161/rna.24046
128. Edwards RA, McNair K, Faust K *et al*., (2016) Computational approaches to predict bacteriophage-host relationships. FEsMS Microbiol Rev 40:258-272. doi:10.1093/femsre/fuv048
129. Medvedeva S, Liu Y, Koonin EV *et al*., (2019) Virus-borne mini-CRISPR arrays are involved in interviral conflicts. Nat Commun 10:5204. doi:10.1038/s41467-019-13205-2