Isolation of a phage with an arabinosylated genome from a cheese microbial community

We sampled cheese microbial communities to discover bacteriophages with unusual genome chemistries. We isolated 114 bacterial host strains and 17 phages, and identified one phage with a probable arabinose hypermodification of hydroxymethylcytosine.

Version 3, published Oct 12, 2023. Originally published Jul 18, 2023.

Arcadia Science DOI: 10.57844/arcadia-743p-ty94

Purpose

We sought to discover bacteriophages with novel DNA modifications. We began our search in cheese rind communities.

We first isolated a large panel of 114 bacterial host strains from cheese rinds, which we used to isolate 17 bacteriophages. We successfully screened a subset of these phages for DNA modification using HPLC and found one phage that modified its cytosines. When we sequenced the genome of the modified phage, we found genes associated with cytosine hydroxymethylation and generation of arabinose-UDP. We conclude that this phage likely performs arabinosylation of hydroxymethylcytosine.

Ultimately, after completing this work, we have decided not to continue our larger effort of studying phage nucleoside chemistries due to the technical challenges of doing this at the scale that would be required to reliably identify novel chemistries. We're sharing our approaches, findings, and genomic resources with the hope that others interested in similar questions will find it useful.

- This pub is part of the **project**, "Exploring bacteriophage nucleic acid chemistries." Visit the project narrative for more background and context.
- Access **raw metagenomic short reads** from our 11 starting cheese rind communities and one associated virome in the ENA under BioProject PRJEB57452; taxonomic and functional analysis is available on MGnify. **Raw metagenomic long reads** for WH 2M Hous in the ENA under run ERR11581409.
- *Serratia* phage 92A1 and *Arthrobacter* phage 1191A **genomes** are in GenBank and **16S rRNA gene sequencing data** for their hosts is on Zenodo.
- We've shared two new phage isolation **protocols**, "Enriching and isolating phages on agar plates" and "Enriching and isolating phages in liquid culture," on protocols.io.

We've put this effort on ice!

#HardToScale #TechnicalGap

We successfully identified a single phage with a genome modification using culture-based methods. However, this was a labor-intensive process with a low payoff-to-effort ratio, and also revealed some limitations of metagenomic sequencing to study phages with modified genomes. We would need a more scalable, efficient method to enable us to meet our goals of discovering novel DNA chemistries.

Learn more about the Icebox and the different reasons we ice projects.

Background and goals

The goal of this study is to identify new bacteriophage nucleic acid chemistries in microbial communities. We chose to start our search in cheese rinds. Cheese

rinds are experimentally tractable microbial communities composed of bacteria, viruses, and fungi that we can easily sample and study in the lab [1].

We harvested rinds and viromes from a set of 12 cheese communities and sequenced their metagenomes. By collecting this sequencing data up front, we hoped we could place any phages we isolated within the context of their community and also potentially use the community metagenomics to guide our isolation efforts. We launched an isolation and screening effort to first isolate putative phage hosts from cheese communities, and then used those hosts to isolate phages. Last, we used our HPLC-based workflow [2] to identify phages with non-standard DNA chemistries.

> **SHOW ME THE DATA**: Access our *Serratia* phage 92A1 and *Arthrobacter* phage 1191A **genomes** in GenBank and find **16S rRNA gene sequencing data** for their hosts on Zenodo. You can find **raw metagenomic short reads** from our 11 starting cheese rind communities and one associated virome in the ENA under BioProject PRJEB57452 (see MGnify for taxonomic/functional analysis) and **raw metagenomic long reads** for WH 2M Hous in the ENA under run ERR11581409.

The approach

We describe our methods in detail below, but provide a brief overview here. Using cheese as our model microbial community, we tried out a few different approaches to phage and host isolation. We also sequenced the cheese rind communities. We ultimately isolated 114 bacterial host strains, as well as 17 phages from cheese.

We were able to screen a subset of these 17 phages for DNA modification using HPLC analysis of phage nucleosides. We found two phages with unusual HPLC profiles: one phage that had consistently high levels of ribonucleosides in its genomic DNA prep, and another phage that had a clear modification of cytosine.

To learn more about these phages, we sequenced the genomes of both of these phages, and performed 16S rRNA gene sequencing on their hosts.

Skip straight to the results or read on for more methodological detail.

Cheese community sampling

We initially harvested cheese rind microbial communities (comm_1–comm_12) and their paired viromes (vir_1–vir_12) from 12 cheeses. In parallel, we stocked the cheese rinds and viromes for downstream bacteria and phage isolation. We harvested DNA from all the communities and viromes for sequencing, but only one of our viromes (vir_1, from comm_1) produced measurable amounts of DNA according to Nanodrop quantification. All of our microbial communities produced enough DNA to sequence, with the exception of comm_9, which we did not move forward with (see Table 1 for details on cheese type/origin and the sequencing we performed). As part of a related effort, we also generated paired short- and long-read sequencing datasets for time series of cheese communities [3]. We included one of those samples in our isolation efforts (WH 2M Hous).

See the following step-by-step protocols for more details:

- Harvesting and stocking cheese rind community samples
- High-molecular-weight DNA extraction from cheese rind microbial communities
- Virome harvesting from cheese microbiomes
- Virome DNA extraction with phenol-chloroform

Sample name	Cheese type	Sequencing
vir_1	Bloomy rind, goat milk, France	Illumina short-read, virome fraction of comm_1
comm_1	Bloomy rind, goat milk, France	Illumina short-read, whole community
comm_2	Bloomy rind, cow milk, France	Illumina short-read, whole community
comm_3	Washed rind, cow milk, CA USA	Illumina short-read, whole community
comm_4	Washed rind, cow milk, VT USA	Illumina short-read, whole community
comm_5	Washed rind, cow milk, France	Illumina short-read, whole community
comm_6	Washed rind, sheep milk, Italy	Illumina short-read, whole community
comm_7	Washed rind, cow milk, Italy	Illumina short-read, whole community
comm_8	Natural rind, sheep milk, France	Illumina short-read, whole community
comm_10	Natural rind, sheep milk, France	Illumina short-read, whole community
comm_11	Natural rind, sheep milk, France	Illumina short-read, whole community
comm_12	Natural rind, cow milk, France	Illumina short-read, whole community
WH 2M Hous	Washed rind, cow milk, VT USA	Native ONT, whole community [3]

Table 1. **Cheeses that we harvested and sequenced**.

Community metagenomics

As described previously [3], we sent whole community DNA for comm_1–comm_12 and vir_1 to Novogene for 2×150 bp paired-end sequencing on the Illumina NovaSeq platform. We uploaded raw reads to the ENA under the BioProject PRJEB57452. We then re-downloaded and processed these reads using our Arcadia-Science/metagenomics workflow [4], and uploaded full assemblies to the same BioProject. We prepared the WH 2M Hous library with ONT kit SQK-LSK112 and ran it on a full R10.4 flow cell on a Nanopore ONT GridION. This run is part of a larger dataset of paired long- and short-read cheese metagenomes, under BioProject PRJEB58160. Check out [3] for more info on this effort.

Host isolation (manual)

For a subset of cheese communities 1–12, we isolated hosts by resuspending a small scraping of frozen banked rind in PBS + 0.05% Tween, and plating out serial dilutions for single colony picking on PCAMS media. See our full protocol here. While we initially plated the bacterial hosts on PCAMS media, we restruck and purified them on LB media and found no growth difference. For simplicity, we used LB media to grow these strains moving forward. We isolated 6–12 strains per cheese.

Host isolation (high-throughput)

For community WH 2M Hous, we tried using a robotic culturing system called the Prospector in collaboration with Isolation Bio (formerly known as GALT). We provided Isolation Bio with a sample of WH 2M Hous with approximately 2×10^9 bacterial CFUs (colony-forming units) per mL based on CFU dilution plating and counting of the rind cell suspension. Isolation Bio then did a 1:20,000 dilution of this cell suspension and loaded the dilution into a Prospector Array containing LB + 100 µg/mL cycloheximide (anti-fungal). At 116 hours of aerobic growth in the array at room temperature, they transferred bacterial isolates from nanowells that had detected growth based on a change in resazurin signal to arrayed 96-well plates using a 1× chip-chip-plate standard transfer protocol. Isolation Bio consolidated the final 235 successful bacterial isolates into three 96-well glycerol stock plates. We then pinned glycerol stock plates onto LB agar in single-well OmniTray plates and grew bacteria at room temperature for three days to see colony morphologies. We found that this isolation method strongly enriched for one species or strain (likely a *Brachybacterium,* see Figure 1, A), so to rebalance our strain library, we selected 33 isolates that by eye appeared unique. In working with this panel of 33 isolates, we found that seven of them were actually mixed colony types that we had to further purify by streak plating. We suspect that by looking by eye for unique morphologies initially, we enriched for mixed colony types because they looked different than the other pure colonies. After purification, we ended up with a final panel of 42 isolates.

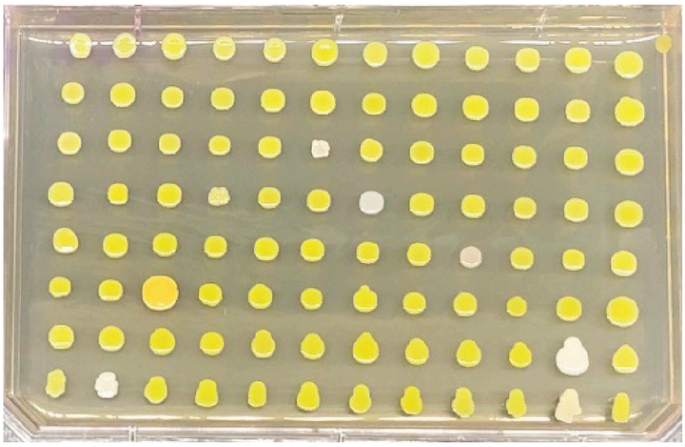

Figure 1. **A representative image of isolates from high-throughput isolation**.
A pinned 96-well plate of isolates from high-throughput isolation of WH 2M Hous community. A single morphology of smooth, opaque, and greenish-yellow colonies dominated this isolation effort. These characteristics are consistent with *Brachybacterium,* which is present but does not make up a large proportion of the WH cheese communities.

Host 16S rRNA gene sequencing

We extracted DNA from 500 µl of bacterial culture using the Omega EZNA kit with no modifications. We used 5 ng of DNA as input into a PCR reaction using Q5 Master Mix (NEB M0492S), with an annealing temperature of 56 °C and an extension time of 30 s. We used the primer pair below to generate a 1465 bp product. We checked the PCR product using agarose gel electrophoresis, and purified using the Zymo DNA Clean and Concentrator (D4033). Primordium Labs sequenced the amplicons, and we analyzed the data by BLASTn against the NCBI nr database on the NCBI BLAST web portal.

Primers (5′ to 3′)

16S_27F: AGAGTTTGATCCTGGCTCAG

16S_1492R: GGTTACCTTGTTACGACTT

Phage isolation

We tried three different approaches to phage isolation: 1) Using frozen banked viromes with paired bacterial isolates; 2) Screening freshly prepared viral extracts from resampled cheeses against previously isolated host strains; and 3) Using

rapid, high-throughput isolation to generate a panel of isolates and screen against a paired virome.

We started our phage isolation efforts for bacterial isolates from washed-rind communities (comm_3–comm_7). Using frozen banked viromes from communities 1–12, we tried both pooled liquid enrichment (see protocol) and spotting of individual virome samples onto individual host strains on solid plates (see protocol). Based on plaque morphology and host strain, we observed substantial overlap between the phages recovered from the liquid and plate-based isolations. Out of the nine phages isolated here, we only recovered one of them (103-1A) from the solid plate isolation but not the liquid isolation. Because our spotting protocol allows us to track which phages came from which virome, we observed that while we screened viromes for communities 1–12, almost all of our phage isolates originated from the same community as their hosts (Table 2). Only one phage (111-1A) came from an unmatched virome source — it was isolated on a host strain from comm_6 but originated from the viral extract of comm_4 (virome 4).

> **TRY IT**: Our full protocols, "Enriching and isolating phages on agar plates" and "Enriching and isolating phages in liquid culture," are available on protocols.io.

Community	Number host strains screened	Number phages isolated	Phage designation
comm_3	10	0	-
comm_4	12	4 (all from virome 4)	87-3b, 88-1A, *Serratia* phage 92A1, 93-1A1
comm_5	12	3 (all from virome 5)	97-1A, 103-1A, 106-1A,
comm_6	8	1 (from virome 4)	111-1A
comm_7	7	1 (from virome 7)	*Arthrobacter* phage 1191A

Table 2. **Summary of phage isolation effort for washed-rind cheese isolates using banked frozen viromes**.

Since we were using frozen banked viromes as our starting material for phage isolation, we wondered if our lack of success in isolating phages for some communities (like comm_3) indicated that the storage process was inactivating our phages. We decided to test whether using fresh, unfrozen viral extracts could increase our chance of success. We obtained new wheels of the previously sampled cheese communities six months after our initial sampling, and harvested fresh phage extracts. We pooled these fresh viral extracts and did a liquid enrichment with our initial panel of 49 strains from washed-rind cheeses, augmented by 15 strains from natural-rind cheeses (comm_10, comm_11) and eight strains from a bloomy-rind cheese (comm_1). Despite using fresh samples and an expanded pool of hosts, we did not recover substantially more phages (Table 3). This may have been because we did not prepare the viral samples from the same wheel of cheese from which we isolated the host.

Community	Number host strains screened	Number phages isolated	Phage designation
comm_1	8	1	141-2A1
comm_3	10	0	
comm_4	12	0	
comm_5	12	0	
comm_6	8	0	
comm_7	7	1	123-4A1
comm_10	8	0	
comm_11	7	3	134-1A, 135-1A, 135-2B

Table 3. **Summary of phage isolation effort with expanded host panel and fresh phage extracts**.

Finally, we tried one more round of phage isolation using WH 2M Hous. We used rapid, high-throughput isolation to obtain a panel of isolated strains against which we then screened a fresh phage extract from the same community. We used both liquid enrichment and plate spotting to isolate phages. Here, we screened 42 hosts and initially isolated five unique phages (by plaque morphology). However, two of these phages could not be stably propagated and were ultimately lost. In total, we recovered three new phages from this final isolation effort (Table 4).

Community	Number host strains screened	Number phages isolated	Phage designation
WH 2M Hous	42	3	152-1, 169-1, 175-1

Table 4. Summary of WH 2M Hous phage isolation effort.

Phage DNA extraction

We amplified phages to a high titer before doing PEG precipitation and phenol-chloroform DNA extraction (see our full protocols for phage amplification and PEG precipitation and phenol-chloroform DNA extraction). All phages used liquid amplification, with the exception of phage 88-1A and *Serratia* phage 92A1, which we amplified using the double agar overlay method on solid plates. Note that the phenol-chloroform DNA extraction protocol ends with a step to digest the DNA down to single nucleosides for chemical analysis. For applications that required intact DNA, like whole-genome sequencing, we omitted this final digestion step.

Phage nucleoside analysis with HPLC

We digested phenol-chloroform-extracted DNA down to single nucleosides using the NEB Nucleoside Digestion kit (see our full protocol for details). We used a short gradient to quickly screen and a long gradient to precisely resolve peaks. The short run is 10 minutes, and uses an isocratic gradient at 100% 20 mM ammonium acetate pH 5.4 and 20% MeOH. The long run is 90 minutes, and uses an isocratic gradient at 100% 20 mM ammonium acetate pH 5.4 and 1% MeOH. We ran each sample in triplicate. More details about our HPLC methods can be found in this protocol.

Arthrobacter phage 1191A RNase digest

We digested 1 µg of phage DNA with RNaseA at a final concentration of 10 µg/mL at 25 °C for 2 h, in the presence of 10 mM EDTA. We prepared an RNaseA-minus control reaction in parallel. These two samples underwent column cleanup using the Zymo DNA Clean and Concentrator kit (D4033) to remove digested nucleosides. We recovered ~500 ng of DNA for each sample in a volume of 25 µl. We then analyzed these samples via HPLC with a 60-minute run to assess their nucleoside content.

Phage genome sequencing and annotation

We extracted genomic DNA from *Serratia* phage 92A1 and *Arthrobacter* phage 1191A with phenol-chloroform, and generated libraries using the Illumina DNA Prep kit (20018704) with an input of 250 ng per phage. We sequenced samples with 2×150 bp reads on an Illumina MiniSeq machine. We adapter-trimmed reads and quality-controlled using fastp (version 0.23.2) [5], then assembled using SPAdes (version 3.15.5) [6] using the `-isolate` flag. *Arthrobacter* phage 1191A assembled as one contig with end repeats and a coverage of 19,273. *Serratia* phage 92A1 assembled as one contig with end repeats and a coverage of 738.

We used Pharokka (version 1.3.2) for genome annotation [7]. Within Pharroka, PHANOTATE predicts coding sequences [8], tRNAscan-SE 2.0 predicts tRNAs [9], Aragorn predicts tmRNAs [10], and CRT predicts CRISPR RNAs [11]. Then MMseqs2 [12] functionally annotates the genes using the PHROGS database [13], VFDB [14], and CARD [15]. Mash [16] matches contigs to their closest hit in the INPHARED database [17] and pyCirclize [18] creates final plots. To augment this annotation, we also used the HHpred web server to predict functions for specific genes of interest. We opened the *Serratia* phage 92A1 genome between *rIIA* and *rIIB*, in accordance with related T4-like genomes, resulting in a linear DNA molecule of 174,432 bp. The *Arthrobacter* phage 1191A genome was opened between *terL* and *terS*, resulting in a linear DNA molecule of 39,310 bp. We made a global comparison of the phage RB69 genome and *Serratia* phage 92A1 genome using the VipTree (v3.6) web server [19]. We used clinker (v0.0.27) to compare genomic neighborhoods of interest [20].

> **SHOW ME THE DATA**: Access our *Serratia* phage 92A1 and *Arthrobacter* phage 1191A **genomes** in GenBank and find **16S rRNA gene sequencing data** for their hosts on Zenodo.

Phage abundance analysis across metagenomes

To assess the abundance of isolated phage genomes in each metagenome community, we used two approaches. First, we used sourmash `compare` to determine the maximum containment between each phage genome and each

metagenome community. Sourmash uses k-mer sketches to estimate similarity or containment between sequencing samples [21]. We first used sourmash (version 4.8.2) `sketch` to sketch the phage genomes, using k size of 51 and a scale value of 1000. Using these sketches as well as those for the metagenome communities output by the Arcadia-Science/metagenomics Nextflow pipeline, we used sourmash `compare` to estimate the maximum containment of each phage in each community. We also used read mapping to measure how many read pairs matched the isolate phage genomes in communities of interest. To do this, we first trimmed and cleaned the reads using fastp (version 0.23.2) [5] and then aligned the reads to the isolated genomes using Bowtie 2 (version 2.5.1) [22]. We report the number of paired reads that aligned concordantly to the genome exactly one time as the number of read pairs mapped.

The results

Two out of eight phages we screened have an unusual HPLC nucleoside profile

We isolated 114 bacterial strains from nine cheese rind microbial communities and used this strain library to isolate 17 phages. Our goal was to use this panel of phages to get a sense of how common DNA modification is in a set of phages isolated from similar environments and to hopefully identify novel nucleoside chemistries. To do this, we grew our phages to high titer and extracted DNA to analyze by HPLC for nucleoside modifications. We were only able to harvest sufficient quantities of DNA for eight out of the 17 phages using the same standard conditions of growth on LB media supplemented with 1 mM $MgSO_4$ and 1 mM $CaCl_2$. While we could have spent significantly more time optimizing growth conditions for each phage-host pair, we decided to move forward with the eight phages for which we could easily isolate large quantities of DNA for analysis.

When we ran digested nucleosides from the eight phages on HPLC, we noticed two phages with non-standard features in their nucleoside profile (Table 5). The phage we've named "*Arthrobacter* phage 1191A" had a high level of C, G, and A RNA nucleosides in the isolated DNA fraction compared to all the other phages. A

phage we've named "*Serratia* phage 92A1" was missing the typical peak for the nucleoside dC and instead had an unknown peak.

Phage	Nucleosides observed by HPLC
88-1A	dC, dG, dT, dA
Serratia phage 92A1	dG, dT, dA, unknown peak (no dC)
97-1A	dC, dG, dT, dA
106-1A	dC, dG, dT, dA
Arthrobacter phage 1191A	dC, dG, dT, dA, C, G, A
141-2A1	dC, dG, dT, dA
169-1	dC, dG, dT, dA
175-1	dC, dG, dT, dA

Table 5. **Summary of HPLC screen of phage nucleoside content**.

Phages are named based on an internal numbering system.

Arthrobacter phage 1191A may package an RNA molecule

We investigated the two phages with non-standard HPLC nucleoside profiles further. *Arthrobacter* phage 1191A had what appeared to be RNA contamination in its DNA when we ran its digested nucleosides on the HPLC (Figure 2, A). This was surprising, because before DNA extraction, the phages undergo DNase and RNase treatment to degrade un-encapsidated nucleic acids. The phages we prepared and analyzed in parallel (97-1A, 106-1A, 141-2A1, 169-1, 175-1) did not have any RNA in their DNA samples, indicating that in most cases our methods worked as designed (Figure 2, B). We grew more *Arthrobacter* phage 1191A and re-prepped its DNA twice more, and saw the same RNA signal in the phage DNA both times.

We hypothesized that the RNA was either coming from an RNase-resistant un-encapsidated RNA, high levels of incorporation of ribonucleosides into the phage genome, or an RNA molecule that was packaged inside the capsid. We digested the prepped phage DNA with another RNaseA treatment followed by a spin column cleanup, and found that this greatly decreased the RNA signal, potentially consistent with a packaged RNA molecule (Figure 2, C). Since we only saw G, C, and A ribonucleosides in the HPLC trace (Figure 2, A), we assume that the RNA

molecule is not a standard mRNA molecule, but could be a small, GC-rich, structured RNA.

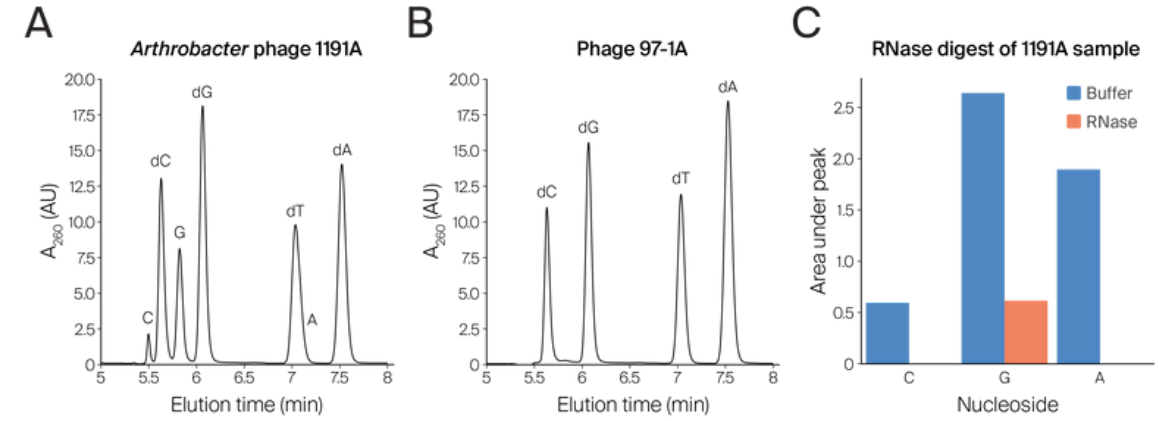

Figure 2. **RNA signal in *Arthrobacter* phage 1191A DNA preps**.

(A) *Arthrobacter* phage 1191A HPLC profile, using 100 ng of digested nucleosides. Peaks are annotated based on elution times of nucleoside standards that we ran in parallel. In this short 10-minute run, the adenine ribonucleoside appears as a slight shoulder on the thymidine peak.

(B) HPLC trace of 97-1A, a representative example of the phages prepped in parallel that did not have RNA contamination.

(C) Areas under each ribonucleoside HPLC peak in *Arthrobacter* phage 1191A DNA that we treated with RNase or a buffer control. We analyzed 60 ng of digested nucleosides with a 60-minute gradient to fully resolve individual nucleoside peaks for quantification.

Next, we sequenced and annotated the phage genome to see if it encoded any ncRNAs, such as tRNAs or CRISPR RNAs, which might be obvious candidates for a phage to package. The phage genome was 39,281 bp long with a GC content of 55% and 71 protein-coding genes. We did not observe any ncRNAs. We also used 16S rRNA gene sequencing of the phage host and found that the 16S amplicon had 99% identity over 100% of the 16S region of *Arthrobacter bergerei* (also called *Glutamicibacter bergerei*), a common cheese strain. We've shared the 16S sequence of our isolate through Zenodo (DOI: 10.5281/zenodo.8132984). We deposited this phage genome in GenBank (accession number OR088901), and also are sharing the annotated genome through Zenodo.

SHOW ME THE DATA: Access our *Serratia* phage 92A1 and *Arthrobacter* phage 1191A **genomes** in GenBank and find **16S sequencing data** for their hosts on Zenodo. You can find **raw**

metagenomic short reads from our 11 starting cheese rind communities and one associated virome in the ENA under BioProject PRJEB57452 (see MGnify for taxonomic/functional analysis) and **raw metagenomic long reads** for WH 2M Hous in the ENA under run ERR11581409.

Overall, we consider this to be a very preliminary result. It is possible that the RNA signal comes from a stable exogenous non-encapsidated RNA molecule that simply required multiple RNase treatments to be fully removed. Because we're not following up on this line of inquiry, we have decided to simply release these observations with the hope that others may find them useful.

Serratia phage 92A1 likely uses arabinose to hypermodify hydroxymethylcytosine in its genome

When we ran *Serratia* phage 92A1 nucleosides on the HPLC, we observed that there was no peak that matched the retention time for dC (~13 min), and instead observed a new peak that eluted just after 20 min (Figure 3, A and B). This is consistent with the phage modifying 100% of its cytosines.

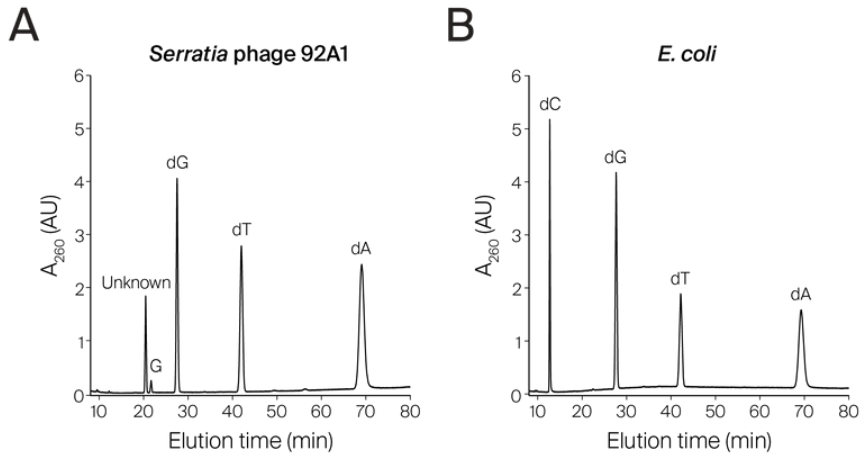

Figure 3. **Figure 3**. **HPLC analysis reveals *Serratia* phage 92A1 uses an unknown nucleoside**.
(A) HPLC trace of *Serratia* phage 92A1 nucleosides.
(B) HPLC trace of *E. coli* bacterial nucleosides.
Peaks are annotated based on elution times of nucleoside standards that we ran in parallel.

We re-ran the phage DNA with a panel of common cytosine modifications such as methylcytosine, hydroxymethylcytosine, and phage T4 DNA, which contains alpha and beta linked glucosyl-methylcytosines. The mystery peak did not match any of these modified nucleosides (Table 6).

Nucleoside	Elution time (minutes)
Hydroxymethyl deoxycytidine	13.306
-glucosyl-methyl deoxycytidine	19.952
Unknown peak	**20.517**
β-glucosyl-methyl deoxycytidine	28.511
Methyl deoxycytidine	26.353

Table 6. Elution times of known cytosine modifications relative to the unknown cytosine modification.

Next, we sequenced the phage genome to see if we could figure out what the modification might be based on the phage's genes. The phage was 174,432 bp with a GC content of 39%. It is predicted to encode 294 protein-coding genes, and six tRNAs. We deposited this phage genome in GenBank (accession number OR088902), and we're also sharing the annotated genome through Zenodo (DOI: 10.5281/zenodo.8132984). We used 16S rRNA gene sequencing to identify the phage host, and found that the 16S amplicon has a 99.9% identity across 100% of the 16S region of *Serratia proteamaculans*, which is a common cheese bacterium. We've shared the 16S sequence of our isolate through Zenodo.

Conservation of DNA modification genes in *Serratia* phage 92A1 and Phage RB69

We looked at the phage gene annotations for genes potentially involved in nucleoside modification. We immediately noticed an arabinose 5-phosphate isomerase downstream of DNA polymerase, as well as a gene annotated as a thymidylate synthetase, a common annotation of phage dUMP hydroxymethylases and dCMP hydroxymethyltransferases [23].

We compared the *Serratia* phage 92A1 genome to T4-like *E. coli* phage RB69, which is known to use arabinose to hypermodify hydroxymethylcytosines in its genome [24]. Whole-genome alignment showed that *Serratia* phage 92A1 and

phage RB69 are related, with mostly syntenic genomes (Figure 4, A), suggesting that *Serratia* phage 92A1 may use the same genome modification as RB69.

We next looked at the genes related to DNA modification across these two phage genomes. Phage RB69 is thought to use the same pathway as phage T4 to generate hydroxymethyl deoxycytidine triphosphate (hmdCTP) from dCTP via the activity of a dCTPase, a dCMP hydroxymethyltransferase, and a hydroxymethyl-dCMP kinase [23][24]. This nucleotide is then incorporated into the genome by phage DNA polymerase. Next, a separate pathway would be used to prepare the arabinose donor molecule. It appears that this pathway is less well-understood, but is hypothesized to result in the generation of UDP-arabinose [24]. Ultimately, a glycosyltransferase would be required to transfer the arabinose onto the hydroxymethylated cytosines in the phage genome (Figure 4, B).

We found strong conservation of genes related to DNA modification between the two phage genomes. The *Serratia* phage 92A1 gene annotated as a thymidylate synthetase is a homolog of the RB69 dCMP hydroxymethyltransferase (Figure 4, C). Both phages also encode a dCTPase and a nucleotide kinase (annotated as a thymidylate kinase) that could potentially act with the dCMP hydroxymethyltransferase to generate hmdCTP. *Serratia* phage 92A1 also encodes its own DNA polymerase. We conclude that *Serratia* phage 92A1 is producing hmdCTP and incorporating that nucleotide into its genome during DNA replication.

The next step in this modification pathway is the generation of an arabinose donor (putatively UDP-arabinose), which would be used to hypermodify the hydroxymethylcytosine in the phage genome [24]. Downstream of the DNA polymerase, we observed conservation of genes likely involved in the generation of an arabinose-UDP (Figure 4, C). These phages have a conserved arabinose 5-phosphate isomerase and another conserved phosphoheptose isomerase (Figure 4, C). They may be involved in the generation of arabinose-5-phosphate from another pentose phospho-sugar, potentially ribulose-5-phosphate. These two sugar-phosphate isomerases flank a gene encoding a conserved hypothetical protein in these phages (Figure 4, C). More careful analysis of the *Serratia* phage 92A1 hypothetical protein with the HHpred web server revealed that it is a multi-domain protein with a sugar phosphate nucleotidyl-transferase domain

(probability: 99.73%, E-value: $3.6e^{-16}$) and a kinase domain (probability: 99.41%, E-value: $1.6e^{-12}$). This gene appears to be split into two genes in the RB69 genome, though follow-up by others has shown that the annotated split is due to a sequencing error [24]. We propose that the sugar phosphate nucleotidyl-transferase domain performs a uridyltransfer reaction with the arabinose-5-phosphate molecule to form UDP-arabinose. It is unclear exactly what role the kinase domain plays in the UDP-arabinose biosynthesis pathway, but our best guess is that it participates in the generation of phosphorylated arabinose.

This genetic neighborhood also contains a gene annotated as a peptidase, specifically a U32 domain peptidase (HHpred probability: 99.94%, E-value: $2.9e^{-25}$). This gene piqued our interest because it is the most closely related protein that *Serratia* phage 92A1 and RB69 share, with >85% amino acid identity across 99% of the protein (see the red region of the Figure 4, A dotplot, and the darkest linkage in the gene-level alignment in Figure 4, C). We initially were unsure about the significance of this strong conservation, and confused as to why a peptidase would be in a gene cluster related to DNA modification. Upon further reading, we learned that U32 peptidases often have unexpected functions, including nucleic acid modification. For example, RhlA is an *E. coli* U32 peptidase involved in hydroxylation of a cytosine in the 23S rRNA molecule [25], and TrhP is another *E. coli* U32 peptidase involved in tRNA hydroxylation [26]. We're excited by the possibility that this phage U32 peptidase may have a novel role in bacteriophage DNA modification, though we have no idea what that role might be.

The final step of this genome modification pathway requires a glycosyltransferase to use the arabinose-UDP donor to transfer an arabinose molecule to the genomically incorporated hydroxymethylcytosines. The paper that originally described the arabinosylation of the RB69 genome identified a putative glycosyltransferase (RB69ORF003c, colored purple in Figure 4, D) that they hypothesized might be responsible for catalyzing that reaction [24]. However, this gene was not conserved in *Serratia* phage 92A1 and there were no other obvious glycosyltransferases in that genetic neighborhood (Figure 4, D). Future research — including a dedicated bioinformatic search for glycosyltransferases in the *Serratia* phage 92A1 genome as well as biochemical characterization of the RB69ORF003c protein — will be needed to clarify the genes responsible for the arabinose transfer onto hydroxymethylated DNA.

In summary, we conclude that *Serratia* phage 92A1 likely uses the same arabinose hypermodification of hydroxymethylated cytosines as phage RB69. While it seems likely that both phages use a T4-like pathway to generate hydroxymethyl deoxycytidine triphosphate (hmdCTP) and incorporate it into the genome, there are several big unknowns about how these phages go about using arabinose to hypermodify those genomic hydroxymethylcytosines.

One question is how the arabinose-UDP donor molecule is generated, and we think it will be especially interesting to determine the role of the highly conserved U32 peptidase in this process. It's also unclear which glycosyltransferase transfers the arabinose moiety onto the hydroxymethylcytosine in the genome. Comparative analysis of phage genomes known to make the same genome modification will probably help with figuring out which genes are responsible.

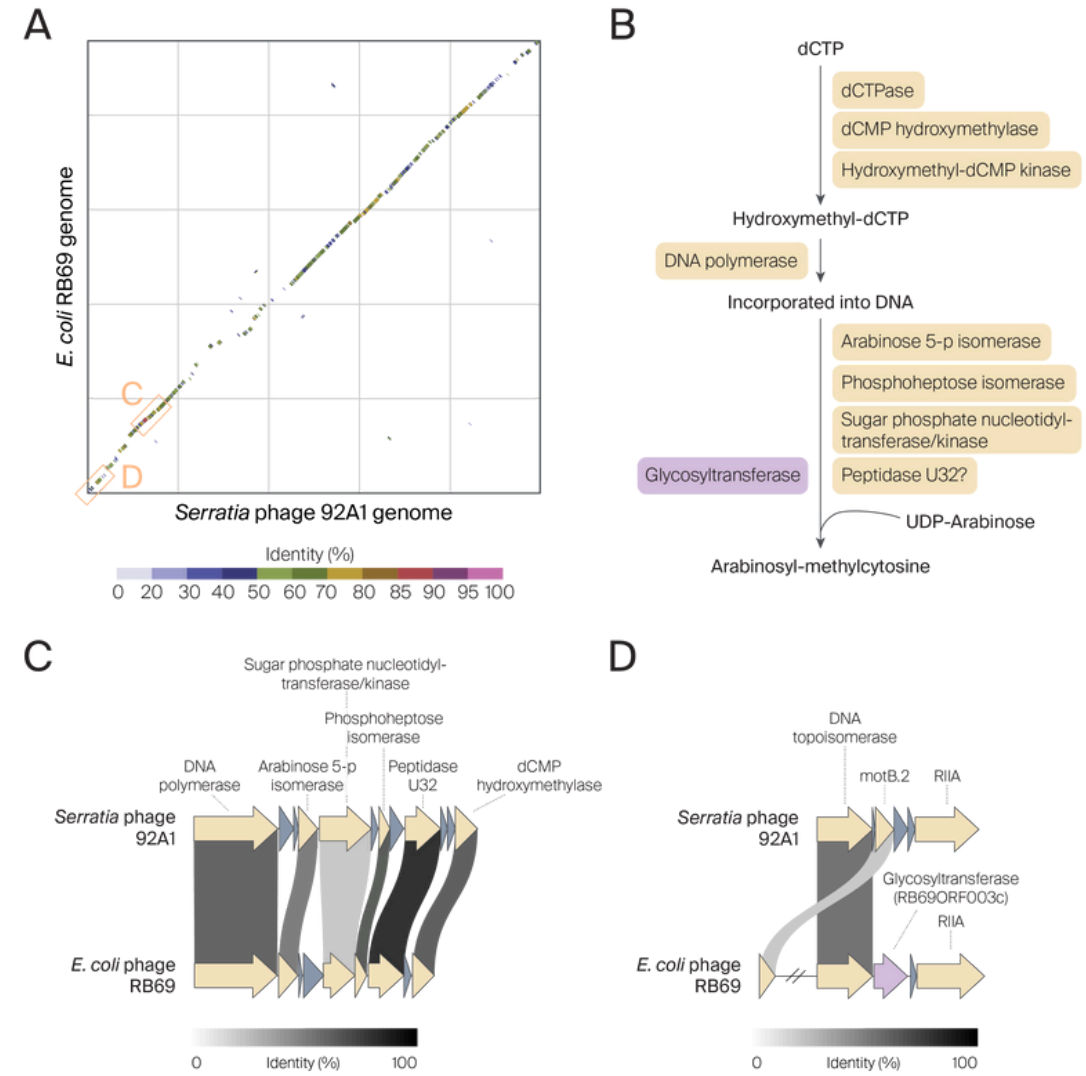

Figure 4. **Genomic basis of nucleoside modification in *Serratia* phage and Phage RB69**.

(A) Dotplot comparison of the *Serratia* phage and Phage RB69 genomes. Regions are colored by their similarity. The genomic regions explored in C and D are boxed and labeled.

(B) Pathway for hypermodification of hydroxymethylcytosine with arabinose, and the enzymes involved in each step. Genes that are conserved in RB69 and *Serratia* phage 92A1 are in yellow. A glycosyltransferase (purple) is necessary to complete the modification, but the candidate transferase in RB69 is not conserved in *Serratia* phage 92A1.

(C) A gene cluster putatively responsible for generating UDP-arabinose encoded between DNA polymerase and dCMP hydroxymethylase.

(D) Genetic neighborhood of RB69 ORF003c (purple), the glycosyltransferase hypothesized to link arabinose to hydroxymethylated cytosines.

In both C and D, conserved genes are in yellow, and genes that are not conserved are in blue-grey. The ribbons linking genes indicate shared amino acid identity, and the darkness of the ribbons indicates percent identity.

Low abundance of modified phage in community metagenomes

DNA modification evolved in phages to protect their genomes from degradation by bacterial immune systems [27]. This presumably increases phage fitness, potentially resulting in higher densities and wider distributions of modified phages compared to unmodified phages.

We wondered if modified *Serratia* phage 92A1 would be widely distributed and/or highly abundant. We included *Arthrobacter* phage 1191A as a "control," unmodified phage and used sourmash to identify all the communities (comm_1 through comm_12) that had k-mer matches to either phage genome.

We were surprised to find that *Serratia* phage 92A1 only had matches in comm_4, the community we originally isolated it from (Figure 5). In comparison, *Arthrobacter* phage 1191A had matches in comm_7, the community we isolated it from, as well as communities 4, 6, 8, and 11. We were also surprised to see that *Serratia* phage 92A1 was at extremely low abundance in comm_4. Read-mapping of comm_4 metagenome against the *Serratia* phage 92A1 genome revealed that only 17 out of 27,761,010 read pairs concordantly mapped to *Serratia* phage 92A1. This is an extremely low signal, and means that without isolation and sequencing of the phage genome, we likely would not have detected *Serratia* phage 92A1 in the community.

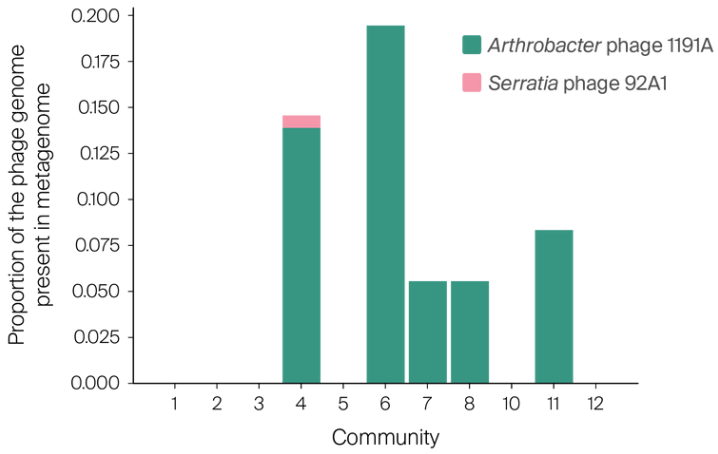

Figure 5. **Representation of isolated phage genomes in short-read cheese community metagenomes**.

We used sourmash to evaluate how much of each phage genome was represented, or "contained" within each cheese community short-read metagenome. Fractions of the *Arthrobacter* phage 1191A genome are represented in five community metagenomes, including comm_7, from which it was originally isolated. Only a very small fraction (corresponding to 17 read pairs) of the *Serratia* phage 92A1 genome was represented in the comm_4 metagenome, which is the community it was isolated from.

A potential explanation for this is that isolation can enrich for even very low-abundance phages, making it a more sensitive discovery method for individual phage genotypes than community-wide sequencing. Isolation entails first making a concentrated extract of phages from cheese communities, and then using host strains to selectively amplify individual phage genotypes. Theoretically, isolation can pull out and amplify even as low as one single phage particle from a complex community. It is possible that this is what happened here, though it is unclear why *Serratia* phage 92A1 would be at such low abundance when it was co-isolated with a sensitive bacterial host.

Another explanation may come from the fact that some DNA modifications can reduce the efficiency at which phage genomes are sequenced [28]. While we had no trouble preparing sequencing libraries of purified *Serratia* phage 92A1 DNA, it is possible that DNA modification may impact its apparent abundance in the metagenome relative to organisms with standard DNA chemistries. It remains an open question as to how much DNA modification biases metagenomic recovery of phage genomes. New sequencing approaches such as REMoDE [29] and mEnrich-

seq [30] that specifically target modified DNA for sequencing will be valuable tools in characterizing the full diversity of phage DNA in microbial communities.

Key takeaways

We set out to discover phages with novel DNA modifications in microbial communities. We began our search with cheese communities because they are easy to sample, safe to handle, and substantially derisked as an experimentally tractable "model" microbial community.

We cast a wide initial net by isolating 114 bacterial strains and screened these against paired virome extracts as well as freshly sampled extracts from related cheese communities. We isolated 17 phages, one of which had an obvious DNA modification that is likely an arabinose hypermodification of genomically incorporated hydroxymethylated cytosine. Using bioinformatic analysis, we propose a set of candidate genes potentially involved in generating the modification in this phage, including a potentially novel role for a U32 peptidase domain protein.

Next steps

Overall, while our isolation efforts were technically successful, this process was very labor-intensive and had an extremely low recovery rate of modified phages. We've concluded that we'd need to use high-throughput sampling techniques to discover phage genomes with novel chemistries in microbial communities if we were to pursue this further.

We were hoping that we could use metagenomics to guide phage isolation, for instance by prioritizing communities and host strains based on metagenomic identification of phages encoding potential marker genes for genome modification in our cheese samples. However, our single modified phage genome was barely detectable in our short-read metagenomic datasets, making it clear that metagenomic measurements aren't necessarily predictive of culturing outcomes. This may be especially true for phages with unusual DNA chemistry. We have also

tried modification-aware Nanopore sequencing and LC-MS/MS of communities to search for interesting chemistries to enable a broader scan of nucleoside diversity in communities, but encountered substantial technical challenges [2]. Ultimately, we've decided to ramp down our phage nucleoside discovery effort.

Acknowledgements

Thank you to Novogene for metagenomics sequencing, Isolation Bio for high-throughput bacterial isolation, and Primoridium labs for 16S amplicon sequencing.

Contributors (A–Z)

- **Januka Athukoralage**: Resources
- **Adair L. Borges**: Conceptualization, Formal Analysis, Investigation, Supervision, Writing
- **Ben Braverman**: Resources
- **Rachel J. Dutton**: Supervision
- **Megan L. Hochstrasser**: Editing, Visualization
- **Elizabeth A. McDaniel**: Software
- **David G. Mets**: Resources
- **Atanas Radkov**: Formal Analysis, Investigation
- **Taylor Reiter**: Data Curation, Formal Analysis
- **Emily C.P. Weiss**: Investigation

References

1. Wolfe B, Button J, Santarelli M, Dutton R. (2014). Cheese Rind Communities Provide Tractable Systems for In Situ and In Vitro Studies of Microbial Diversity. https://doi.org/10.1016/j.cell.2014.05.041
2. Borges AL, Radkov A, Thuy-Boun PS. (2022). A workflow to isolate phage DNA and identify nucleosides by HPLC and mass spectrometry. https://doi.org/10.57844/arcadia-1ey9-j808
3. Borges AL, Dutton RJ, McDaniel EA, Reiter T, Weiss EC. (2023). Paired long- and short-read metagenomics of cheese rind microbial communities at multiple time points. https://doi.org/10.57844/arcadia-0zvp-xz86

4. Dutton RJ, McDaniel EA. (2023). Quickly preprocessing and profiling microbial community sequencing data with a Nextflow workflow for metagenomics. https://doi.org/10.57844/arcadia-7etp-pj24
5. Chen S, Zhou Y, Chen Y, Gu J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. https://doi.org/10.1093/bioinformatics/bty560
6. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. https://doi.org/10.1089/cmb.2012.0021
7. Bouras G, Nepal R, Houtak G, Psaltis AJ, Wormald P-J, Vreugde S. (2022). Pharokka: a fast scalable bacteriophage annotation tool. https://doi.org/10.1093/bioinformatics/btac776
8. McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA. (2019). PHANOTATE: a novel approach to gene identification in phage genomes. https://doi.org/10.1093/bioinformatics/btz265
9. Chan P, Lin B, Mak A, Lowe T. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. https://doi.org/10.1093/nar/gkab688
10. Laslett D. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. https://doi.org/10.1093/nar/gkh152
11. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. https://doi.org/10.1186/1471-2105-8-209
12. Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. https://doi.org/10.1038/nbt.3988
13. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio R, Mom R, Toussaint A, Petit M-A, Enault F. (2021). PHROG: families of prokaryotic virus proteins clustered using remote homology. https://doi.org/10.1093/nargab/lqab067
14. Chen L. (2004). VFDB: a reference database for bacterial virulence factors. https://doi.org/10.1093/nar/gki008
15. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran H-K, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman

FSL, Hsiao WWL, Domselaar GV, McArthur AG. (2019). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. https://doi.org/10.1093/nar/gkz935

16. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. (2016). Mash: fast genome and metagenome distance estimation using MinHash. https://doi.org/10.1186/s13059-016-0997-x
17. Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, Stekel DJ, Hobman J, Jones MA, Millard A. (2021). INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. https://doi.org/10.1089/phage.2021.0007
18. https://github.com/moshi4/pycirclize
19. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. (2017). ViPTree: the viral proteomic tree server. https://doi.org/10.1093/bioinformatics/btx157
20. Gilchrist CLM, Chooi Y-H. (2021). clinker & clustermap.js: automatic generation of gene cluster comparison figures. https://doi.org/10.1093/bioinformatics/btab007
21. Titus Brown C, Irber L. (2016). sourmash: a library for MinHash sketching of DNA. https://doi.org/10.21105/joss.00027
22. Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. https://doi.org/10.1038/nmeth.1923
23. Weigele P, Raleigh EA. (2016). Biosynthesis and Function of Modified Bases in Bacteria and Their Viruses. https://doi.org/10.1021/acs.chemrev.6b00114
24. Thomas JA, Orwenyo J, Wang L-X, Black LW. (2018). The Odd “RB” Phage—Identification of Arabinosylation as a New Epigenetic Modification of DNA in T4-Like Phage RB69. https://doi.org/10.3390/v10060313
25. Kimura S, Sakai Y, Ishiguro K, Suzuki T. (2017). Biogenesis and iron-dependency of ribosomal RNA hydroxylation. https://doi.org/10.1093/nar/gkx969
26. Sakai Y, Kimura S, Suzuki T. (2019). Dual pathways of tRNA hydroxylation ensure efficient translation by expanding decoding capability. https://doi.org/10.1038/s41467-019-10750-8
27. Samson JE, Magadán AH, Sabri M, Moineau S. (2013). Revenge of the phages: defeating bacterial defences. https://doi.org/10.1038/nrmicro3096

28. Leskinen K, Pajunen MI, Vilanova MVG-R, Kiljunen S, Nelson A, Smith D, Skurnik M. (2020). YerA41, a Yersinia ruckeri Bacteriophage: Determination of a Non-Sequencable DNA Bacteriophage Genome via RNA-Sequencing. https://doi.org/10.3390/v12060620

29. Enam SU, Cherry JL, Leonard SR, Zheludev IN, Lipman DJ, Fire AZ. (2023). Restriction Endonuclease-Based Modification-Dependent Enrichment (REMoDE) of DNA for Metagenomic Sequencing. https://doi.org/10.1128/aem.01670-22

30. Cao L, Kong Y, Fan Y, Ni M, Tourancheau A, Ksiezarek M, Mead EA, Koo T, Gitman M, Zhang X-S, Fang G. (2022). mEnrich-seq: Methylation-guided enrichment sequencing of bacterial taxa of interest from microbiome. https://doi.org/10.1101/2022.11.07.515285