PAPER

Microdiversity is higher in temperate than in virulent bacteriophages from soil environments

Thomas E.P. de Bruijn,[1,2,*] Hilje M. Doekes[2] and Anne Kupczok[1,*]

[1]Bioinformatics Group, Wageningen University and Research, Droevendaalsesteeg 1, 6708PB, The Netherlands and [2]Laboratory of Genetics, Wageningen University and Research, Droevendaalsesteeg 1, 6708PB, The Netherlands
[*]Corresponding author: thomas.debruijn@wur.nl, anne.kupczok@wur.nl.

Abstract

Bacteriophages are not only the most ubiquitous biological entity on earth, they also display remarkable genetic diversity across and within populations. While macrodiversity has been extensively studied, the drivers of microdiversity (intra-species genetic diversity) remain poorly understood, particularly in relation to phage lifestyle. The distinguishing ability of temperate phages to integrate themselves into the host genome has an unknown influence on the microdiversity present. This difference in microdiversity could impact the adaptability of phages to (a)biotic factors. To identify a possible association between microdiversity and lifestyle, we analysed 12 existing viromics datasets focusing on soil bacteriophages, including 41 412 viral genomes in total. We found that phages predicted to be temperate consistently exhibit significantly higher microdiversity than their virulent counterparts in eight of 12 datasets, whereas the remaining four datasets did not show a significant trend. The detected pattern holds across multiple quality thresholds and lifestyle prediction methods. These findings suggest that lysogeny may promote or preserve genetic variation within phage populations, with potential implications for phage-host coevolution and environmental adaptability.

Key words: bacteriophages, microdiversity, lysogeny, metagenomics, soil virome

1. Introduction

Viruses are known to be the most ubiquitous biological entities on Earth, exerting immense influence over prokaryotic and eukaryotic life. Bacteriophages for example shunt and shuttle nutrients throughout ecosystems by lysing up to 40% of marine bacterial populations each day [Gregory et al., 2019, Roux and Emerson, 2022, Jansson and Wu, 2023, Piedade et al., 2024]. The ubiquity of bacteriophages is reflected in—and perhaps due to—their morphological and genetic diversity [Dion et al., 2020, Roux and Coclet, 2025]. Advances in sequencing methods have enabled researchers to study this genetic diversity of bacteriophages at an unprecedented scale and precision [Gregory et al., 2019].

Recent research into microbiome biodiversity has unearthed the vast diversity of bacteriophages, especially in soil environments [Santos-Medellin et al., 2021, Roux and Emerson, 2022, Chen et al., 2025, Liao et al., 2025]. Historically, this biodiversity is described as macrodiversity or any of the many related terms (e.g. species richness, $\alpha/\beta/\gamma$-diversity)[Kimura et al., 2008]. The decreased costs and improved accuracy of DNA sequencing have also paved the way to study another type of biodiversity, the so-called microdiversity. Whereas macrodiversity describes between-species diversity, microdiversity is the genetic diversity *within* a species or population.

In microbial communities, both bacterial and phage microdiversity is involved in maintaining the possibility for long-term host-virus coexistence [Piel et al., 2022]. In bacterial populations, microdiversity is essential in providing population-level protection against changing environmental conditions [Larkin and Martiny, 2017, García-García et al., 2019]. The role microdiversity plays in the eco-evolutionary dynamics of phages cannot be uncoupled from those dynamics of the host system [Brüssow et al., 2004, Koskella and Brockhurst, 2014, Burmeister et al., 2016]. Bacteriophages and bacteria are involved in a complex coevolutionary arms race between infecting and preventing infection. Microdiversity in marine phages has been shown to be influenced by host availability, environmental stability and virus population dynamics [Coutinho et al., 2019, Gregory et al., 2019]. However, microdiversity in bacteriophages is still understudied compared to microdiversity in bacteria [Dion et al., 2020].

The interplay between host and phage gains an extra dimension when the lysogenic lifestyle of temperate phages is considered in contrast to the lytic lifestyle of virulent (or obligatory lytic) phages [Hobbs and Abedon, 2016, Wahl and Pattenden, 2017]. Both virulent and temperate phages can perform the lytic lifestyle: the phage infects the cell, starts replication of its genome and production of necessary structural proteins and finally induces cell lysis to release new virions. The lytic lifestyle thus includes the relatively error prone viral mutation rate [Peck and Lauring, 2018]. The lysogenic lifestyle of temperate phages involves integration of the viral genome into the host genome, instead of immediate production of virions and lysis. This turns the phage into a prophage and the host into a lysogen. The prophage can multiply with the lysogen when the cell divides, or excise itself from the host genome and return to the lytic lifestyle. Depending on biotic and abiotic factors, temperate and virulent phages can be under very different selection pressures. Some bacteria can transform into endospores, potentially protecting the prophage from extreme environmental conditions [Van Goethem et al., 2019]. Furthermore, prophages can exchange genetic material with superinfecting phages or other prophages [Moineau et al.,

© The Author 2026.

1994, Brüssow et al., 2004]. How the lifestyle of bacteriophages affects their genetic microdiversity is currently unknown.

Here we studied the association between phage lifestyle and microdiversity in *Caudoviricetes*, a well-studied class of double-stranded DNA viruses that can have either a virulent or temperate lifestyle [Roux and Coclet, 2025]. Twelve soil-borne bacteriophage datasets, selected for their recent publication and FAIR data (i.e. data was Findable, Accessible, Interoperable and Reusable), were re-analyzed. We focus on soil environments, as these tend to contain a more even ratio between virulent and temperate phages [Liang et al., 2020]. The 12 datasets together contained 537 397 viral operational taxonomic units (vOTUs, approximately species rank viral genomes), which were filtered for completeness, mapping quality and taxonomy, using a high-throughput analysis pipeline. A total of 41 412 high-quality *Caudoviricetes* vOTUs were assigned a virulent or temperate lifestyle and microdiversity values. We show that eight of the 12 datasets consistently indicate temperate phages as having a higher microdiversity than virulent phages.

2. Experimental procedures

2.1. Data collection

Datasets were selected from metagenome studies of soil-based bacteriophages catalogued in the NCBI PubMed database. Additional requirements were recent publishing (< 5 years, from December 2025) and data availability of raw reads and vOTUs. Using these criteria, 17 studies were selected for further analysis (Table 1). The vOTUs and raw reads were retrieved for these datasets. vOTUs were downloaded from their respective data repositories. Raw reads were retrieved from SRA by their BioProject ID (Table 1 for IDs) using the NCBI SRA-toolkit (github.com/ncbi/sra-tools) [Sayers et al., 2022]. There were slight differences how the original studies computed the vOTUs from assembled contigs. 14 studies followed the predominant standard for clustering of contigs into vOTUs: at least 95% average nucleotide identity over 85% alignment fraction [Gregory et al., 2019, Roux et al., 2019], whereas 3 studies used different parameters. All datasets were also re-clustered together using PSI-CD-HIT (v4.8.1, using the flags for alignment thresholds "-c 0.95 -G 1 -g 1 -aS 0.85 -aL 0.85") and the standard definition for vOTUs (at least 95% average nucleotide identity over 85% alignment fraction) to check for redundancy caused by different vOTU clustering definitions between datasets [Fu et al., 2012]. This re-clustering did not result in a meaningful difference in vOTU definition or dataset size (see Supplemental Information, "Re-clustering of vOTUs using standard thresholds").

2.2. vOTU quality assessment and mapping of reads

The files containing the vOTUs (for each dataset) were used as supplied by the authors of the respective publications. All vOTUs were analysed with CheckV (v1.0.3, "end_to_end" pipeline, database version 1.5) and filtered on at least 80% estimated genome completeness [Nayfach et al., 2021].

Raw reads of each library were mapped against the vOTUs of the respective study with Bowtie2 (v2.4.4), using the "very-sensitive" setting [Langmead and Salzberg, 2012]. Unmapped reads were discarded during the mapping process, using the Bowtie2 "–no-unal" flag. Bowtie2 SAM output files were converted to BAM files, sorted using SAMtools (v1.21) and merged within each dataset [Danecek et al., 2021]. Coverage of contigs by mapped reads was calculated using CoverM (v0.7.0, using the contig option, "mean" coverage calculation method) [Aroney et al., 2025]. vOTUs were filtered for a minimum coverage of 10x and breadth of coverage of at least 80%. These thresholds were necessary for downstream microdiversity analysis, primarily ensuring accurate (high coverage) microdiversity estimation of the full genome (high breadth). All vOTUs were also assigned taxonomic information using geNomad (v1.11.1, default settings) [Camargo et al., 2024].

2.3. Prediction of vOTU lifestyle

The lifestyle of all vOTUs was predicted using PhaTYP (as included in PhaBOX2 v2.1.11) [Shang et al., 2023a], PhaStyle (v0.1) [Juhász et al., 2024] and verified by checking for integrases with DIAMOND (v2.0.8) [Buchfink et al., 2021]. PhaStyle uses a ProkBERT genomic language model to detect similarity of query sequences to phages with known lifestyles. It uses a binary classification model: the negative class in the model is the temperate lifestyle, the positive being virulent. PhaStyle was run using a GPU, greatly accelerating the processing speed compared to CPU-based usage. PhaStyle was run using the "neuralbioinfo/PhaStyle-mini" model available online (huggingface.co/neuralbioinfo/prokbert-mini). PhaTYP also uses a BERT-based model, but differs from PhaStyle in the method of embedding and masking sequences [Shang et al., 2023b, Juhász et al., 2024]. PhaTYP also has virulent lifestyle as positive classification class and temperate as negative class.

vOTUs were also checked for presence of integrase or recombinase, using the cluster of orthologous genes COG0582 (containing 2205 genes) from the COG database and DIAMOND (e-value <1e-20 and minimum query coverage per High-Scoring Pair of 80%), with the genes as predicted by Prodigal as query sequences [Hyatt et al., 2010, Buchfink et al., 2021, Galperin et al., 2025]. The final lifestyle classification was determined by the overlap between PhaTYP and PhaStyle and presence of a DIAMOND hit with a COG0582 sequence for the temperate lifestyle or absence of such a hit for the virulent lifestyle.

2.4. Microdiversity analysis of filtered vOTUs

All vOTUs with mapped reads were used as input for calculating microdiversity using inStrain (v1.9.1) [Olm et al., 2021]. inStrain internally also applies a 95% ANI filter for each aligned reads, reducing spurious mappings. The main output metric of interest was the nucleotide diversity (π) per vOTU, included in the scaffold output files. Additionally, pN/pS ratios in predicted gene sequences were also retrieved from the gene output files. pN/pS ratios indicated the ratio between non-synonymous and synonymous mutations in gene coding sequences and can be an indication of selective pressure on a genetic sequence.

2.5. Association analysis of lifestyle-microdiversity relationship

The association analysis between phage lifestyle and microdiversity was performed using R (v4.4.2) [R Core Team, 2024] and additional libraries from the tidyverse were used for data handling and visualisation [Wickham et al., 2019]. We required there to be at least ten vOTUs in each lifestyle group per dataset analysed. A Wilcoxon rank-sum test was used to calculate the statistical relationship between the two lifestyle groups and their respective microdiversity.

Table 1. Overview of the 17 datasets used in this study. All datasets are of soil-based viruses, most of them focusing specifically on bacteriophages. Studies that performed size fractioning ($< 0.22\mu$m) in their sample treatment show the number of viromes (v) or metagenomes (m). If a study had both viromes and metagenomes available, only the number of viromes is shown. Standard clustering parameters are >= 95% ANI over >= 85% AF. **ANI** = Average Nucleotide Identity, **AF** = Alignment Fraction.

Dataset ID	Source	BioProject ID	Clustering	# of libraries	# of vOTUs
PermaThaw	[Trubl et al., 2025]	PRJNA888099	standard	20 (v)	9 963
MediGrass	[Fudyma et al., 2025]	PRJNA1264504	standard	59 (v)	22 725
NewMexico	[Feeser et al., 2025]	PRJNA1263537	standard	11 (v)	3 233
Conifer	[Geonczy et al., 2025a]	PRJNA1113186	standard	90 (v)	272 017
Wildfire	[Geonczy et al., 2025b]	PRJNA1147726	standard	39 (v)	70 135
Glacial	[Liao et al., 2025]	PRJNA882246	>= 95% ANI over >= 80% AF	24 (v)	59 685
Intertidal	[Ji et al., 2024]	PRJNA957716	standard	96 (m)	20 102
BodegaBay	[Ter Horst et al., 2023]	PRJNA913601	standard	63 (v)	12 689
Watershed	[Coclet et al., 2023]	PRJNA1168581	standard	48 (m)	2 743
WetUp	[Santos-Medellín et al., 2023]	PRJNA859194	standard	152 (v)	12 259
Rhizo	[Braga and Schumacher, 2023]	PRJNA935076	>= 95% ANI over >= 80% AF	12 (v)	13 335
Hopland	[Santos-Medellín et al., 2022]	PRJNA818793	standard	44 (v)	6 088
UKagri	[Muscatt et al., 2022]	PRJEB49313	>= 95% ANI over >= 95% AF	15 (v)	1 059
NCalHabitat	[Durham et al., 2022]	PRJNA831438	standard	30 (v)	3 432
Land-use	[Liao et al., 2022]	PRJNA691683	standard	50 (v)	61 106
Spruce	[Ter Horst et al., 2021]	PRJNA666221	standard	10 (v)	5 006
AgriSoil	[Santos-Medellin et al., 2021]	PRJNA646773	standard	15 (v)	4 065

2.6. Functional annotation of viral genes

vOTUs belonging to the Land-use dataset also underwent functional prediction of genes. For this, genes were first predicted using Prodigal (v2.6.3, anonymous mode) [Hyatt et al., 2010]. After this, eggNOG mapper (v2.1.13, default settings) was used in combination with the eggNOG v5 database to predict the function of genes in vOTUs [Huerta-Cepas et al., 2019, Cantalapiedra et al., 2021]. Annotations were parsed for keywords in the GFF fields for predicted PFAM domains ("em_PFAMS") and description ("em_desc"). Since the focus was on essential genes, only the genes containing the following values in the GFF fields "em_desc" amd "em_PFAMs" were highlighted during visualisation: "terminase", "capsid / head / coat", "portal", "tail", "helicase", "integrase" and "recombinase".

2.7. SRR2strain

The workflow presented here has been largely integrated into a single Snakemake pipeline (v9.3.3), called SRR2strain (`https://github.com/thattimetho/SRR2strain`, Supplemental Figure 1). Snakemake is a python-based workflow management system, that allows the linking of input and output of tools to automate workflows [Mölder et al., 2021]. The steps mentioned here, from the retrieval of the raw reads up until the microdiversity analysis using inStrain, have been incorporated into the pipeline. The acquisition of the respective vOTU databases, functional annotation of Land-use vOTUs and final analysis using R scripts are not incorporated in the pipeline.

3. Results

3.1. Viromes facilitate stringent filtering required for reliable microdiversity estimation

A systematic literature search was performed for studies on bacteriophages and soil environments, taking into account the recency of publication and availability of data. This search resulted in 17 datasets. These 17 datasets formed the starting point of our analysis into the relationship between phage lifestyle and microdiversity. Of the 17 datasets, two contained only metagenomes, while the rest contained only viromes or both viromes and metagenomes (Table 1). The datasets are from diverse soil environments (e.g. coastal, agricultural, urban). These datasets were used as input for the SRR2strain pipeline (Supplemental Figure 1, also see subsection 2.7), which predicts phage lifestyle and estimates phage microdiversity. As part of this pipeline, sequenced reads are mapped back onto the vOTUs to allow for the calculation of microdiversity downstream (using inStrain) [Olm et al., 2021]. During this mapping process, we observed that the metagenomes showed extremely poor mapping efficiency (less than 3% mapped reads, Supplemental Table 1). This was irreconcilable with the thresholds on minimum mapping breadth and coverage necessary for accurate microdiversity estimation. For this reason, the samples with only metagenomes (Watershed and Intertidal) were excluded from the analysis. Only viromes were considered for the remaining datasets and analyses, which led to a significant improvement in mapping efficiency (see Supplemental Table 1).

The 15 remaining datasets were filtered for sufficient mapping coverage and breadth, and vOTU completeness. This led to the removal of Rhizo, Spruce and UKagri, as these datasets did not have (sufficient) vOTUs left for lifestyle microdiversity association analysis. From the other 12 datasets, 117 022 out of 537 397 vOTUs remained after these (mapping) quality control measures were applied. Subsequent taxonomic classification was performed using geNomad and resulted in successful classification of 116 445 vOTUs (99.5% of 117 022 total). Of these, 115 922 (99.6%) were predicted to belong to the *Caudoviricetes* class of viruses. For subsequent analyses and results, only these vOTUs classified as belonging to the *Caudoviricetes* class were considered.

Table 2. Overview of datasets used as input for microdiversity analysis. Taxonomic classification resulted in almost all vOTUs being classified as *Caudoviricetes* viruses. Temperate and virulent phages make up roughly 20% and 80% of each dataset respectively. The last two columns are based on total number of *Caudoviricetes* + filtered vOTUs (fourth column).

Dataset ID	# of vOTUs total	# of vOTUs filtered	# of vOTUs *Caudoviricetes* + filtered	# of vOTUs predicted temperate	# of vOTUs predicted virulent
PermaThaw	9 963	241	237	36	201
MediGrass	22 725	3 405	3 353	694	2 659
NewMexico	3 233	394	394	98	296
Conifer	272 017	21 921	21 639	3 941	17 698
Wildfire	70 135	2 741	2 718	818	1 900
Glacial	59 685	6 444	6 380	1 075	5 305
Intertidal	20 102	-	-	-	-
BodegaBay	12 689	253	249	45	204
Watershed	2 743	-	-	-	-
WetUp	12 259	1 637	1 623	667	956
Rhizo	13 335	-	-	-	-
Hopland	6 088	267	264	71	193
UKagri	1 059	-	-	-	-
NCalHabitat	3 432	112	111	18	93
Land-use	61 106	4 262	4 249	1 053	3 196
Spruce	5 006	-	-	-	-
AgriSoil	4 065	195	195	52	143

3.2. Virulent phages comprise majority in lifestyle distribution

Phage lifestyle, either virulent or temperate, was predicted for the 115 922 vOTUs left (after quality filtering) using a combination of tools (PhaTYP, PhaStyle and DIAMOND). This resulted in 41 412 vOTUs which were assigned the same lifestyle by both PhaTYP and PhaStyle, requiring temperate phages to have a DIAMOND hit with an integrase or recombinase sequence (COG0582). Of these 41 412 vOTUs, 8 568 (21%) were classified as having a temperate lifestyle and 32 844 (79%) as having a virulent lifestyle (Table 2). The ratio between virulent and temperate phages per study was similar between all 12 studies (Figure 1). The lifestyle ratio changed slightly when corrected for relative abundance of vOTUs, most notably in the Land-use dataset (see Supplemental Figure 2). From the 12 datasets, 53 779 vOTUs that passed the initial stringent quality checks were discarded due to disagreement between lifestyle prediction methods (see also Figure 1a).

In order to identify and assess any confounding variables, various other metrics, like mapping breadth, coverage, vOTU length and completeness, have also been tested in relation to the lifestyle assigned to the vOTUs and to vOTU nucleotide diversity. Some of these metrics show a significant statistical association with vOTU lifestyle, but the estimated true difference between the groups is very low (Figure 1b-c, Supplemental Figure 4 and Supplemental Table 3).

3.3. Overview of high quality viral datasets

After filtering on mapping and vOTU quality and taking congruent lifestyle predictions, the final assembled dataset consisted of 12 datasets, containing 41 412 vOTUs (7,7% of all 537 397 vOTUs in these datasets, see also Table 2). The included vOTUs were not distributed equally among the 12 studies: ∼50% originated from the Conifer dataset. On the dataset level, mapping breadth and estimated vOTU completeness showed similar values among the 12 studies, while average vOTU length and average coverage showed more variability between datasets (Figure 1 and Supplemental Table 2). Especially the average vOTU length of vOTUs from the Glacial dataset, originating from a study focusing on so-called huge phages, showed a markedly higher value than the vOTUs from other datasets. Median mapping coverage showed values between 22x (NCalHabitat) and 186x (WetUp).

3.4. Soil-borne bacteriophage microdiversity shows association with lifestyle

Microdiversity (π) present within the vOTUs was calculated identify and analyse any association with phage lifestyle. Generally, microdiversity levels were similar between studies and consistent with the microdiversity reported in the original studies [Liao et al., 2022, 2025]. Overall, the studies showed quite a wide range of microdiversity (Figure 2a), ranging from a median around 0.00074 (BodegaBay) to a median of 0.0027 (Land-use).

We find that eight out of 12 datasets showed a significant difference in the level of microdiversity detected for each lifestyle (Figure 2b and Supplemental Table 4). The direction of the difference in microdiversity of these eight datasets was consistent, all showing a higher level of microdiversity in temperate phages as opposed to virulent phages. The WetUp dataset showed the most distinct microdiversity distributions between the two lifestyles ($p = 3.9e\text{-}39$, estimate = 0.0022, Wilcoxon rank-sum test, two-tailed). The Conifer dataset contained the most vOTUs out of all datasets, which influenced the statistical significance of the difference in microdiversity between the lifestyles ($p = 0$, estimate = 0.0019, Wilcoxon rank-sum test, two-tailed). This association between lifestyle and microdiversity was robust when using different lifestyle prediction methods (see Supplemental Table 4).

Four vOTUs from the Land-use dataset (CC25_NODE_304, G12_NODE_652, Z20_NODE_2880 and Z16_NODE_1649) were chosen for further inspection of read mapping quality and microdiversity in the genome. Of these four vOTUs, two were predicted to be virulent and two to be temperate. Of each lifestyle, both a high and low microdiversity example is given. The predicted genes and corresponding annotations

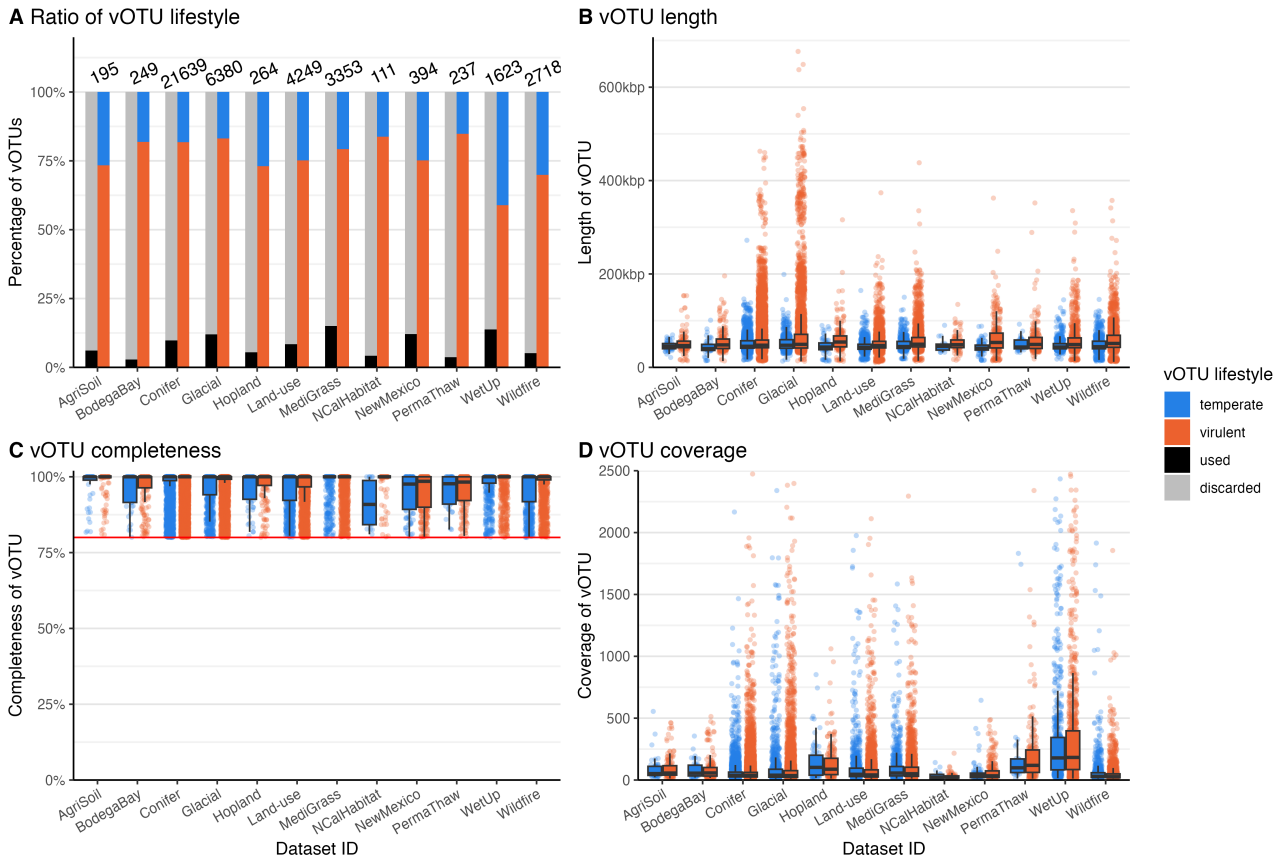

Fig. 1. Overview of vOTU related statistics. All graphs use quality (breadth, completeness, coverage) and lifestyle filtered data. (**a**). Ratios of total and used data, and the ratio between temperate and virulent classified vOTUs. The numbers indicate discarded vOTUs due to quality issues and total vOTUs used for microdiversity analysis. (**b**). Distribution of vOTU lengths among the 12 datasets. Strikingly, only the Glacial and Conifer study show viral genome lengths above 400kbp. (**c**). Predicted vOTU completeness. The used threshold of 80% is indicated with the red line. (**d**). vOTU mapping coverage. The threshold for coverage was put at 10x (data below this threshold not shown), but the plot shows coverages far above this value.

were collected and combined with the coverage of filtered reads (identity >95%) and microdiversity as computed by inStrain (Figure 2c). Overall, there seems to be some conservation of order of proteins. The (pattern of) coverage of these four vOTUs also seems reasonable, containing no clearly depleted or significantly enriched areas of the genome. The pattern (or localisation) of microdiversity seems to vary by vOTU, with Z20_NODE_2880 seemingly showing more microdiversity in the terminase and portal protein predicted proteins compared to the rest of the genome. Meanwhile, Z16_NODE_1649 shows a more even distribution of microdiversity. Lastly, the coverage of the vOTUs does not seem to correlate with microdiversity, which is supported by a correlation analysis on the full dataset (Supplemental Figure 4).

In addition to vOTU-wide microdiversity analysis, the effect of lifestyle on gene-level characteristics was also studied. Besides general statistics, like length and breadth (see Supplemental Figure 3), an estimate of selective pressure on genes was also calculated by inStrain in the form of pN/pS ratios. Generally, ratios between 0 and 1 indicate a purifying selection, while anything above 1 indicates positive selection. Ratios around 1 are a sign of neutral protein evolution. The majority of pN/pS ratios were below 1, indicating purifying selection on the vOTUs in the filtered datasets (Figure 3). Predicted phage lifestyle showed a significant association with pN/pS values in 11 out of 12 datasets (Wilcoxon rank sum test, see also Supplement Table 5): genes from temperate phages showed higher pN/pS values than those from virulent phages.

4. Discussion

Advances in high-throughput, cost-effective sequencing and improved assembly have led to an expanding set of high-quality DNA viromic datasets from diverse environments, including soils. Leveraging these resources, we examined how bacteriophage lifestyle relates to microdiversity. Across datasets, a consistent pattern emerged: temperate phages exhibit higher microdiversity than virulent phages. This trend was robust to alternative lifestyle prediction tools and a range of vOTU quality thresholds. Using stringent filters for both read mapping and lifestyle prediction, we obtained confident association statistics for 12 datasets; eight of the 12 showed significant lifestyle-associated differences, with temperate phages displaying higher microdiversity (Figure 2). This estimated difference in microdiversity between the two lifestyles was ∼36% across those eight datasets, though precise effects differ per dataset. Overall microdiversity levels and lifestyle distribution were in line with those reported by previous soil-focused studies [Liao et al., 2022, Ji et al., 2024, Liao et al., 2025].

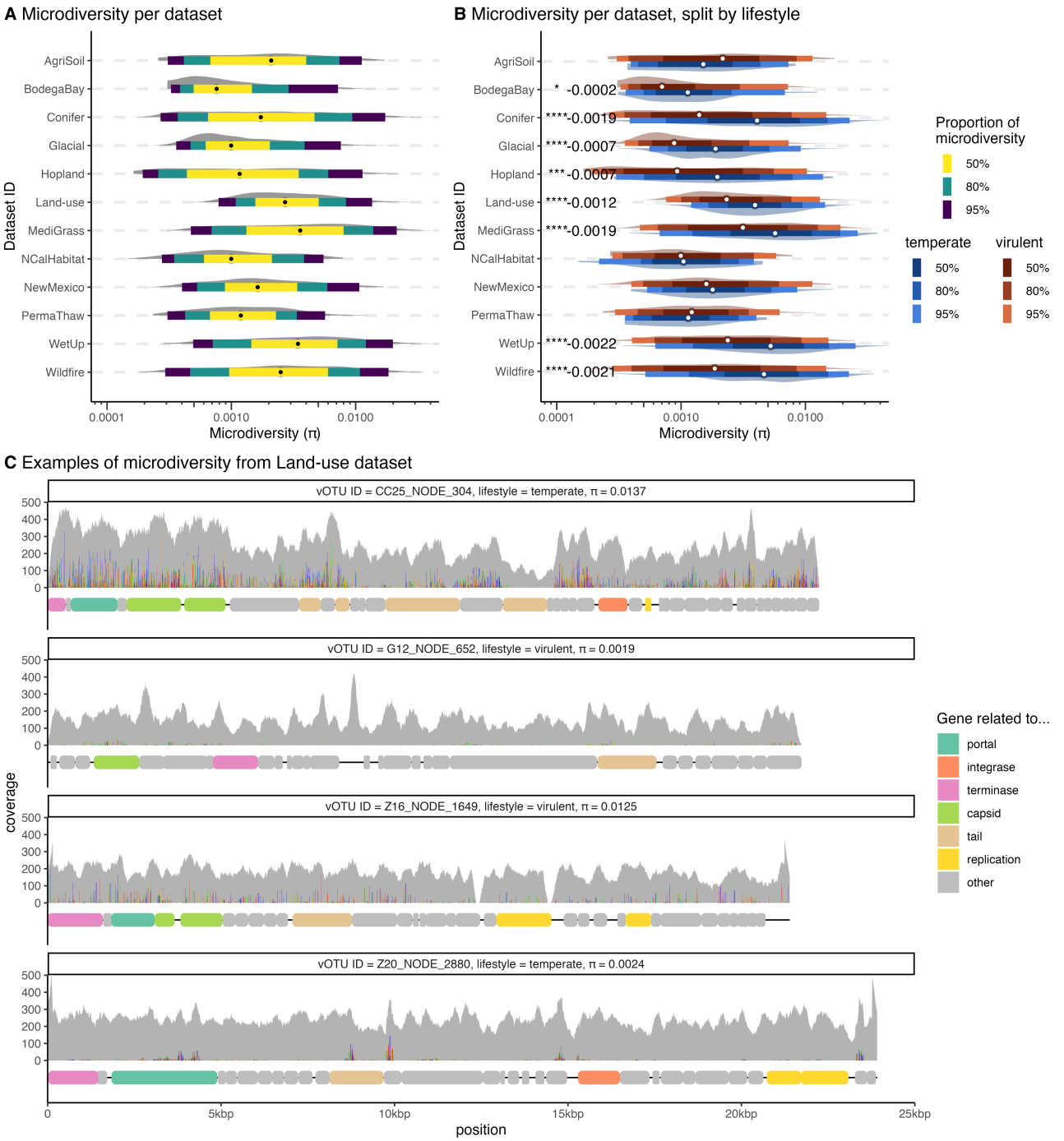

Fig. 2. Higher microdiversity in temperate phages than virulent phages in eight out of 12 tested datasets. Both (**a**) and (**b**) show interval plots with added density distribution and dot indicating the median microdiversity value of the distribution. Value next to significance indication shows estimated true difference between two tested distributions. In panel (**c**) there are four Land-use vOTUs showing different types of microdiversity. (**a**). Microdiversity distribution of vOTUs for the 12 tested datasets. (**b**). Microdiversity distribution split out by predicted lifestyle of vOTUs. The BodegaBay, Conifer, Glacial, Hopland, Land-use, MediGrass, WetUp and Wildfire datasets show the same pattern of differential microdiversity between temperate and virulent phages. This is also visually identifiable in the more right skewed microdiversity distributions of the temperate phages compared to the left skewed distribution of the virulent phages. (**c**). Four vOTUs with coverage, microdiversity and gene annotation. These four vOTUs show the different manifestations of microdiversity in both a virulent and temperate viral genome. The microdiversity and predicted lifestyle is shown in the label of each coverage plot. The temperate phages show the presence of an integrase gene, while the virulent phages do not. The order of proteins seems to be somewhat conserved in vOTUs CC25_NODE_304, Z16_NODE_1649 and Z20_NODE_2880: terminase and portal proteins are situated at start of the vOTU, with integrase (for the temperate phages) or helicase (for the virulent phages) closer to the end of the viral genome. The coverage shown here is from all libraries in the Land-use dataset combined (i.e. the per-dataset approach)

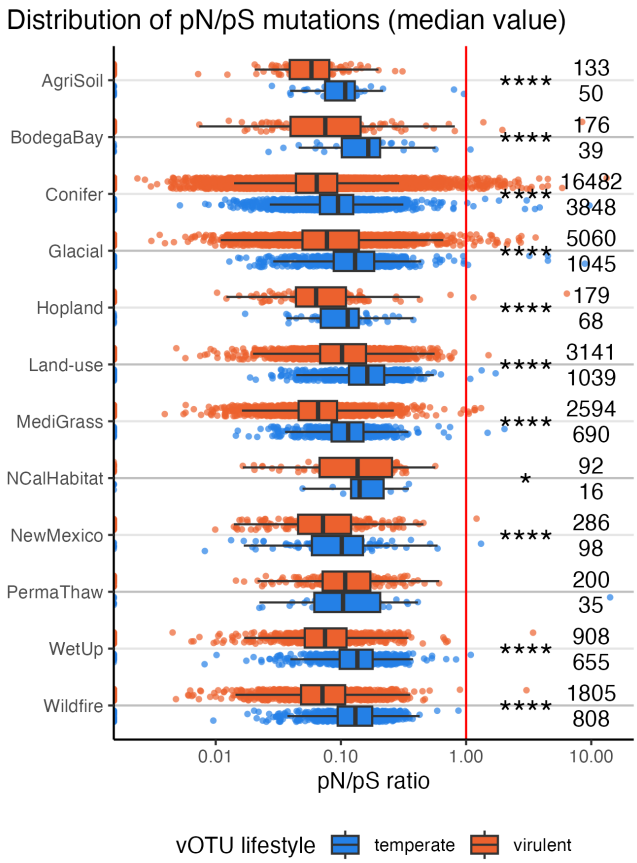

Fig. 3. pN/pS ratios per dataset split by lifestyle. Temperate phages show consistently higher pN/pS ratios than virulent phages, although differences are small. Difference in pN/pS between lifestyle was statistically tested with a Wilcoxon rank sum test, resulting in significant differences in the pN/pS values of 11 of 12 datasets between temperate and virulent lifestyles as indicated by the asterisks (see Supplemental Table 5). The numbers on the right indicate the number of vOTUs that were included in the Wilcoxon rank sum test and plot.

As this project aimed to connect microdiversity with phage lifestyle, accurate microdiversity estimation is key. Microdiversity analysis requires a high read depth, as to filter out sequencing errors and provide certainty on the amount of genetic diversity present. Our initial collection of datasets contained two studies with metagenomes only rather than (size-fractionated) viromes. Mapping the raw (metagenomic) reads against the vOTUs proved to be very inefficient for these metagenomic datasets, often mapping a maximum of 2% of reads and only reaching low coverage of vOTUs (see Supplemental Table 1). When filtering for high mapping coverage and breadth, this led to an extreme reduction in vOTU number, if there were vOTUs with sufficient coverage at all. Indeed, this is a phenomenon previously described, where metagenomes do not contain sufficient viral reads to provide ample coverage for downstream analysis [Santos-Medellin et al., 2021, Kosmopoulos et al., 2024]. Soil samples are especially hard to sequence for viruses using metagenomics, underlining the need for viromes further [Sorensen et al., 2021, Kosmopoulos et al., 2024]. Even when using viromes, the strict quality thresholds we imposed substantially reduced analysable vOTUs. In the BodegaBay dataset, this led to a reduction in vOTUs of 92%, of which: 76% did not pass the quality thresholds (but could be assigned a lifestyle with certainty), and 24% lacked confident lifestyle assignment. For three datasets (Rhizo, Spruce, and UKagri) the stringent filtering resulted in too few vOTUs for association analysis. Although no exact cause is known for these virome studies containing so few vOTUs after filtering, it could be due to the number of libraries included in the datasets. Indeed, these three virome datasets contain the fewest number of libraries, together with the NewMexico dataset (Table 1). Altogether, these results highlight the importance and necessity of (deeply-sequenced) viromes for these fine-grained analyses.

Not only the data used influences the output achieved, so do the various clustering and mapping methods, (tool) settings and filtering steps. For example, the pipeline was benchmarked with and only includes a single mapping tool, Bowtie2. Different mappers with different settings could influence the mapping efficiency and accuracy. We expect only a limited effect of using another mapping tool however, given that the reads from all datasets were mapped with the same tool and the mapping is independent of phage lifestyle (prediction). The subsequent filtering using both coverage and breadth was aimed at reducing spurious mappings. In addition to these thresholds, inStrain was run using default filtering steps, which also includes an ANI cutoff of 95% of mapped reads. Overall, given the many strict thresholds set on vOTU (mapping) quality, we are confident the results presented here are a realistic representation of the microdiversity present in the respective datasets.

Bacteriophage lifestyle is the prime explaining variable in this study, and as such, lifestyle prediction needs to be as accurate and precise as possible. Because of this, we used two sequence-based lifestyle prediction tools, namely PhaTYP and PhaStyle, with DIAMOND search against integrases functioning as positive and negative control [Buchfink et al., 2021, Shang et al., 2023b, Juhász et al., 2024]. In addition to choosing the overlap between these three tools, various combinations of tools have also been tested (Supplemental Table 4). The results of these combinations still show a higher microdiversity in temperate phages compared to virulent phages, showing that the resulting association is robust against variation in choice of (combination of) lifestyle prediction methods. PhaTYP and PhaStyle were chosen for their recent development and extensive benchmarking against established but older tools [Shang et al., 2023b, Juhász et al., 2024]. Despite similar conceptual underpinnings, PhaTYP and PhaStyle yielded substantial variance in lifestyle calls (Supplemental Figure 5). In case of the Land-use dataset, 41.7% of vOTUs had to be discarded due to disagreement between the tools on phage lifestyle. This could also mean the the actual ratio between virulent and temperate phages may differ from our current estimates. This difference between tools is the exact reason we combine the predictions of multiple tools. It also highlights the difficulty of accurate lifestyle prediction. Where feasible, integrating orthogonal evidence such as lysogenic fraction, prophage signatures or long-read phage-host linkage could reduce uncertainty in lifestyle assignment and improve the fidelity of lifestyle-microdiversity associations [McDougall et al., 2018, Liang et al., 2020, Lai et al., 2024].

In order to explain the observed difference in microdiversity between virulent and temperate phages, we need to rewind genetic history to the origin and maintenance of microdiversity in the phage genome. Various population genetic processes, like selective sweeps and mutation rates, are expected to contribute disproportionately in large, rapidly

reproducing viral populations. Yet, the exact link between these processes and the consistent lifestyle-dependent differences we detect, is unknown. Several non-mutually exclusive hypotheses could account for elevated microdiversity in temperate phages. First, temperate phages can alternate between lysogenic and lytic replication, potentially increasing opportunities for accumulation and maintenance of within-species polymorphisms across fluctuating ecological conditions and host states [Maslov and Sneppen, 2015, Van Goethem et al., 2019, Doekes et al., 2021, Aframian et al., 2022]. Second, integration into diverse host genomes may expose temperate phages to host-mediated selection and recombination landscapes, which can elevate microdiversity relative to strictly lytic counterparts [Abedon, 2008, De Paepe et al., 2014]. Third, recurrent induction events may repeatedly reseed the temperate phage population with genetically diverse lineages that have diversified due to genetic drift and mutation during prolonged periods of integration [Nadeem and Wahl, 2017, Wahl and Pattenden, 2017, Van Goethem et al., 2019]. Lastly, and linked to the first hypothesis, virulent phages can and are expected to undergo hard selective sweeps, as they are directly influenced by fluctuating ecological conditions and host states [Burke, 2012]. These hard selective sweeps may remove a part of the microdiversity from the virulent phage population, but not per se from temperate phages which are integrated into a host. These hypotheses should subsequently also be seen and considered in the light of high viral mutation rates. We also briefly looked at localisation of microdiversity within the viral genome, but found no uniform signature or localisation of microdiversity in specific genes or parts of the genome (Figure 2). Limited by the data and methods available, we can only hypothesize on potential explanations for this consistent variation in microdiversity. More research is needed to not only uncover both the exact cause of this difference in microdiversity between lifestyles, but also to analyse if this microdiversity is located in specific parts of the viral genome. Nonetheless, the consistent signal across methods and datasets suggests that lifestyle-linked ecological and evolutionary processes are likely contributors to the lifestyle-microdiversity association detected.

Besides microdiversity, another evolutionary process which can be studied is the ratio between synonymous and non-synonymous mutations (pN/pS or dN/dS ratio) in the coding sequence of genes. The pN/pS ratio is often used as an indication of the type of selection pressure the population is under. The vast majority of the pN/pS values for the 41 412 vOTUs fell between zero and one, indicating that the vOTUs are under purifying selective pressure (Figure 3). Besides this, the pN/pS ratios of 11 out of 12 studies showed a significant difference between the two predicted lifestyles. Additional analyses are necessary to investigate how the distribution of these (mostly synonymous) mutations is spread across the gene repertoire of phages and whether it might be specific to certain (sub)groups of genes, like those involved in host-phage interactions.

Lastly, soil environments are highly spatially heterogeneous, leading to microcosms containing spatially structured bacterial and bacteriophage populations. Prior work has shown that soil viral communities can harbour substantially greater microdiversity than those from marine systems, likely due to this spatial heterogeneity [Gregory et al., 2019, Durham et al., 2022, Liao et al., 2022]. Our results align with previous studies performed in soil environments: viral microdiversity is high in soil environments, and when compared to viral microdiversity in marine environments, microdiversity in soil-associated vOTUs is frequently shown to be higher by an order of magnitude [Gregory et al., 2019]. The unique spatial properties of soil might be essential for the pronounced lifestyle signal we observe, given that spatial heterogeneity can greatly influence both phage and host population dynamics [Haerter and Sneppen, 2012, Wang and Bradburd, 2014, García-García et al., 2019, Durham et al., 2022, Santos-Medellín et al., 2022]. Moreover, while the initial systematic search did not take into account the diversity of sampled biomes, it is worth mentioning that the identified lifestyle-microdiversity association is present across multiple different biomes, like the forests of the Conifer and Wildfire study, and the multiple grassland biomes of the Hopland and MediGrass studies [Santos-Medellín et al., 2022, Geonczy et al., 2025a,b, Fudyma et al., 2025].

In sum, across multiple high-quality soil viromics datasets, temperate phages consistently exhibit higher microdiversity than virulent phages, a pattern robust to alternative prediction tools and filtering thresholds and present throughout biomes. Various evolutionary processes may contribute to this difference in microdiversity between the lifestyles. Further investigation, taking into account the complexity of soil environments, is warranted to identify the ecological and evolutionary mechanisms driving the consistent lifestyle-microdiversity association.

5. Author contributions

TB, HD and AK conceptualization. TB data curation. TB formal analysis. HD and AK funding acquisition. TB investigation. TB and AK methodology. TB, HD and AK project administration. TB software. HD and AK supervision. TB visualization. TB writing - original draft. TB, HD and AK writing - review and editing.

6. Acknowledgements

We would like to thank Marnix Medema and Bas Zwaan for their guidance during the design and final stages of this project. Furthermore, we would like to mention Chen Chen, Dimitris Karapliafis, Laura Patino Medina and Sanne Keizer for their comments on the manuscript.

7. Conflicts of interest

The authors declare no conflicts of interest.

8. Data Availability

A list of IDs and locations for raw read data and vOTU datasets as well as data generated and used for generating visualisations in this study has been uploaded to Zenodo (`https://doi.org/10.5281/zenodo.17988606`). Raw read data of publications used in this study are available on the NCBI SRA database. vOTU datasets of publications used in this study were retrieved from their respective online data repositories.

9. Code Availability

The SRR2strain snakemake pipeline and all other scripts/code generated for this study are available on GitHub (`https://github.com/thattimetho/SRR2strain`). This repository also

contains information on all the tools and libraries used in the processing or visualisation of data.

References

S. T. Abedon, editor. *Bacteriophage ecology: population growth, evolution, and impact of bacterial viruses*. Number 15 in Advances in molecular and cellular microbiology. Cambridge University Press, Cambridge ; New York, 2008. ISBN 978-0-521-85845-8.

N. Aframian, S. Omer Bendori, S. Kabel, P. Guler, A. Stokar-Avihail, E. Manor, K. Msaeed, V. Lipsman, I. Grinberg, A. Mahagna, and A. Eldar. Dormant phages communicate via arbitrium to control exit from lysogeny. *Nature Microbiology*, 7(1):145–153, Jan. 2022. ISSN 2058-5276. doi: 10.1038/s41564-021-01008-5.

S. T. N. Aroney, R. J. P. Newell, J. N. Nissen, A. P. Camargo, G. W. Tyson, and B. J. Woodcroft. CoverM: read alignment statistics for metagenomics. *Bioinformatics (Oxford, England)*, 41(4):btaf147, Mar. 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf147.

L. P. P. Braga and R. I. Schumacher. Awaking the dormant virome in the rhizosphere. *Molecular Ecology*, 32(11):2985–2999, June 2023. ISSN 1365-294X. doi: 10.1111/mec.16893.

H. Brüssow, C. Canchaya, and W.-D. Hardt. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiology and molecular biology reviews: MMBR*, 68(3):560–602, table of contents, Sept. 2004. ISSN 1092-2172. doi: 10.1128/MMBR.68.3.560-602.2004.

B. Buchfink, K. Reuter, and H.-G. Drost. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4):366–368, Apr. 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01101-x. URL `https://www.nature.com/articles/s41592-021-01101-x`.

M. K. Burke. How does adaptation sweep through the genome? Insights from long-term selection experiments. *Proceedings. Biological Sciences*, 279(1749):5029–5038, Dec. 2012. ISSN 1471-2954. doi: 10.1098/rspb.2012.0799.

A. R. Burmeister, R. E. Lenski, and J. R. Meyer. Host coevolution alters the adaptive landscape of a virus. *Proceedings. Biological Sciences*, 283(1839):20161528, Sept. 2016. ISSN 1471-2954. doi: 10.1098/rspb.2016.1528.

A. P. Camargo, S. Roux, F. Schulz, M. Babinski, Y. Xu, B. Hu, P. S. G. Chain, S. Nayfach, and N. C. Kyrpides. Identification of mobile genetic elements with geNomad. *Nature Biotechnology*, 42(8):1303–1312, Aug. 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01953-y.

C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, and J. Huerta-Cepas. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12):5825–5829, Dec. 2021. ISSN 1537-1719. doi: 10.1093/molbev/msab293.

T. Chen, Y. Xiong, J. Zhang, Q. Zhang, J. Wu, N. Xu, and T. Liu. Temporal dynamics, microdiversity, and ecological functions of viral communities during cyanobacterial blooms in Lake Taihu. *NPJ biofilms and microbiomes*, 11(1):178, Aug. 2025. ISSN 2055-5008. doi: 10.1038/s41522-025-00771-1.

C. Coclet, P. O. Sorensen, U. Karaoz, S. Wang, E. L. Brodie, E. A. Eloe-Fadrosh, and S. Roux. Virus diversity and activity is driven by snowmelt and host dynamics in a high-altitude watershed soil ecosystem. *Microbiome*, 11(1):237, Oct. 2023. ISSN 2049-2618. doi: 10.1186/s40168-023-01666-z.

F. H. Coutinho, R. Rosselli, and F. Rodríguez-Valera. Trends of Microdiversity Reveal Depth-Dependent Evolutionary Strategies of Viruses in the Mediterranean. *mSystems*, 4(6):e00554–19, Nov. 2019. ISSN 2379-5077. doi: 10.1128/mSystems.00554-19.

P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2):giab008, Feb. 2021. ISSN 2047-217X. doi: 10.1093/gigascience/giab008.

M. De Paepe, G. Hutinet, O. Son, J. Amarir-Bouhram, S. Schbath, and M.-A. Petit. Temperate phages acquire DNA from defective prophages by relaxed homologous recombination: the role of Rad52-like recombinases. *PLoS genetics*, 10(3):e1004181, Mar. 2014. ISSN 1553-7404. doi: 10.1371/journal.pgen.1004181.

M. B. Dion, F. Oechslin, and S. Moineau. Phage diversity, genomics and phylogeny. *Nature Reviews. Microbiology*, 18(3):125–138, Mar. 2020. ISSN 1740-1534. doi: 10.1038/s41579-019-0311-5.

H. M. Doekes, G. A. Mulder, and R. Hermsen. Repeated outbreaks drive the evolution of bacteriophage communication. *eLife*, 10:e58410, Jan. 2021. ISSN 2050-084X. doi: 10.7554/eLife.58410.

D. M. Durham, E. T. Sieradzki, A. M. Ter Horst, C. Santos-Medellín, C. W. A. Bess, S. E. Geonczy, and J. B. Emerson. Substantial differences in soil viral community composition within and among four Northern California habitats. *ISME communications*, 2(1):100, Oct. 2022. ISSN 2730-6151. doi: 10.1038/s43705-022-00171-y.

K. Feeser, R. Longley, L. V. Gallegos-Graves, M. Albright, and M. Shakya. Recovering new viruses from New Mexico soils. *Microbiology Resource Announcements*, 14(11):e0090825, Nov. 2025. ISSN 2576-098X. doi: 10.1128/mra.00908-25.

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28(23):3150–3152, Dec. 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts565.

J. D. Fudyma, P. Penev, K. Estera-Molina, J. Hoff, S. J. Blazewicz, J. Pett-Ridge, and J. B. Emerson. Spatial and depth structuring predominate over temporal variation in Mediterranean grassland soil viral communities, Dec. 2025. URL `https://www.biorxiv.org/content/10.64898/2025.12.30.696869v1`. ISSN: 2692-8205 Pages: 2025.12.30.696869 Section: New Results.

M. Y. Galperin, R. Vera Alvarez, S. Karamycheva, K. S. Makarova, Y. I. Wolf, D. Landsman, and E. V. Koonin. COG database update 2024. *Nucleic Acids Research*, 53(D1):D356–D363, Jan. 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae983.

N. García-García, J. Tamames, A. M. Linz, C. Pedrós-Alió, and F. Puente-Sánchez. Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *The ISME journal*, 13(12):2969–2983, Dec. 2019. ISSN 1751-7370. doi: 10.1038/s41396-019-0487-8.

S. E. Geonczy, L. S. Hillary, C. Santos-Medellín, J. W. Sorensen, and J. B. Emerson. Patchy burn severity explains heterogeneous soil viral and prokaryotic responses to fire in a mixed conifer forest. *mSystems*, 10(6):e0174924, June 2025a. ISSN 2379-5077. doi: 10.1128/msystems.01749-24.

S. E. Geonczy, A. M. Ter Horst, and J. B. Emerson. Soil viral communities shifted significantly after wildfire in chaparral and woodland habitats. *ISME communications*, 5(1): ycaf073, Jan. 2025b. ISSN 2730-6151. doi: 10.1093/ismeco/ycaf073.

A. C. Gregory, A. A. Zayed, N. Conceição-Neto, B. Temperton, B. Bolduc, A. Alberti, M. Ardyna, K. Arkhipova, M. Carmichael, C. Cruaud, C. Dimier, G. Domínguez-Huerta, J. Ferland, S. Kandels, Y. Liu, C. Marec, S. Pesant, M. Picheral, S. Pisarev, J. Poulain, J.-E. Tremblay, D. Vik, Tara Oceans Coordinators, M. Babin, C. Bowler, A. I. Culley, C. de Vargas, B. E. Dutilh, D. Iudicone, L. Karp-Boss, S. Roux, S. Sunagawa, P. Wincker, and M. B. Sullivan. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, 177(5):1109–1123.e14, May 2019. ISSN 1097-4172. doi: 10.1016/j.cell.2019.03.040.

J. O. Haerter and K. Sneppen. Spatial structure and Lamarckian adaptation explain extreme genetic diversity at CRISPR locus. *mBio*, 3(4):e00126–00112, 2012. ISSN 2150-7511. doi: 10.1128/mBio.00126-12.

Z. Hobbs and S. T. Abedon. Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic'. *FEMS microbiology letters*, 363(7):fnw047, Apr. 2016. ISSN 1574-6968. doi: 10.1093/femsle/fnw047.

J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, S. K. Forslund, H. Cook, D. R. Mende, I. Letunic, T. Rattei, L. J. Jensen, C. von Mering, and P. Bork. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, Jan. 2019. ISSN 1362-4962. doi: 10.1093/nar/gky1085.

D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, Mar. 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119.

J. K. Jansson and R. Wu. Soil viral diversity, ecology and climate change. *Nature Reviews. Microbiology*, 21 (5):296–311, May 2023. ISSN 1740-1534. doi: 10.1038/s41579-022-00811-z.

M. Ji, J. Zhou, Y. Li, K. Ma, W. Song, Y. Li, J. Zhou, and Q. Tu. Biodiversity of mudflat intertidal viromes along the Chinese coasts. *Nature Communications*, 15(1):8611, Oct. 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-52996-x.

J. Juhász, B. Babett, J. Juhász, N. Ligeti-Nagy, S. Pongor, and B. Ligeti. ProkBERT PhaStyle: Accurate Phage Lifestyle Prediction with Pretrained Genomic Language Models, Dec. 2024. URL https://www.biorxiv.org/content/10.1101/2024.12.08.627378v1. Pages: 2024.12.08.627378 Section: New Results.

M. Kimura, Z.-J. Jia, N. Nakayama, and S. Asakawa. Ecology of viruses in soils: Past, present and future perspectives. *Soil Science and Plant Nutrition*, 54(1):1–32, Feb. 2008. ISSN 0038-0768. doi: 10.1111/j.1747-0765.2007.00197.x. URL https://doi.org/10.1111/j.1747-0765.2007.00197.x. _eprint: https://doi.org/10.1111/j.1747-0765.2007.00197.x.

B. Koskella and M. A. Brockhurst. Bacteria-phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS microbiology reviews*, 38 (5):916–931, Sept. 2014. ISSN 1574-6976. doi: 10.1111/1574-6976.12072.

J. C. Kosmopoulos, K. M. Klier, M. V. Langwig, P. Q. Tran, and K. Anantharaman. Viromes vs. mixed community metagenomes: choice of method dictates interpretation of viral community ecology. *Microbiome*, 12(1):195, Oct. 2024. ISSN 2049-2618. doi: 10.1186/s40168-024-01905-x.

S. Lai, H. Wang, P. Bork, W.-H. Chen, and X.-M. Zhao. Long-read sequencing reveals extensive gut phageome structural variations driven by genetic exchange with bacterial hosts. *Science Advances*, 10(33):eadn3316, Aug. 2024. ISSN 2375-2548. doi: 10.1126/sciadv.adn3316.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, Mar. 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923.

A. A. Larkin and A. C. Martiny. Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. *Environmental Microbiology Reports*, 9(2):55–70, Apr. 2017. ISSN 1758-2229. doi: 10.1111/1758-2229.12523.

X. Liang, Y. Zhang, K. E. Wommack, S. W. Wilhelm, J. M. DeBruyn, A. C. Sherfy, J. Zhuang, and M. Radosevich. Lysogenic reproductive strategies of viral communities vary with soil depth and are correlated with bacterial diversity. *Soil Biology and Biochemistry*, 144:107767, May 2020. ISSN 0038-0717. doi: 10.1016/j.soilbio.2020.107767. URL https://www.sciencedirect.com/science/article/pii/S003807172030064X.

H. Liao, H. Li, C.-S. Duan, X.-Y. Zhou, Q.-P. Luo, X.-L. An, Y.-G. Zhu, and J.-Q. Su. Response of soil viral communities to land use changes. *Nature Communications*, 13(1):6027, Oct. 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33771-2.

H. Liao, J. Li, Y.-Z. Wang, H. Li, X.-L. An, T. Wang, R.-Y. Chang, Y.-G. Zhu, and J.-Q. Su. Evolutionary diversification and succession of soil huge phages in glacier foreland. *Microbiome*, 13(1):18, Jan. 2025. ISSN 2049-2618. doi: 10.1186/s40168-024-02017-2.

S. Maslov and K. Sneppen. Well-temperate phage: optimal bet-hedging against local environmental collapses. *Scientific Reports*, 5:10523, June 2015. ISSN 2045-2322. doi: 10.1038/srep10523.

W. M. McDougall, J. M. Perreira, E. C. Reynolds, and A. L. Brass. CRISPR genetic screens to discover host-virus interactions. *Current Opinion in Virology*, 29:87–100, Apr. 2018. ISSN 1879-6265. doi: 10.1016/j.coviro.2018.03.007.

S. Moineau, S. Pandian, and T. R. Klaenhammer. Evolution of a Lytic Bacteriophage via DNA Acquisition from the Lactococcus lactis Chromosome. *Applied and Environmental Microbiology*, 60(6):1832–1841, June 1994. ISSN 0099-2240. doi: 10.1128/aem.60.6.1832-1841.1994.

G. Muscatt, S. Hilton, S. Raguideau, G. Teakle, I. D. E. A. Lidbury, E. M. H. Wellington, C. Quince, A. Millard, G. D. Bending, and E. Jameson. Crop management shapes the diversity and activity of DNA and RNA viruses in the rhizosphere. *Microbiome*, 10(1):181, Oct. 2022. ISSN 2049-2618. doi: 10.1186/s40168-022-01371-3.

F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, and J. Köster. Sustainable data analysis with Snakemake. *F1000Research*, Apr. 2021. doi: 10.12688/f1000research.29032.2. URL https://f1000research.com/articles/10-33.

A. Nadeem and L. M. Wahl. Prophage as a genetic reservoir: Promoting diversity and driving innovation in the host community. *Evolution; International Journal of Organic Evolution*, 71(8):2080–2089, Aug. 2017. ISSN 1558-5646. doi: 10.1111/evo.13287.

S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrpides. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, 39(5):578–585, May 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-00774-7.

M. R. Olm, A. Crits-Christoph, K. Bouma-Gregson, B. A. Firek, M. J. Morowitz, and J. F. Banfield. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology*, 39(6):727–736, June 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-00797-0.

K. M. Peck and A. S. Lauring. Complexities of Viral Mutation Rates. *Journal of Virology*, 92(14):e01031–17, July 2018. ISSN 1098-5514. doi: 10.1128/JVI.01031-17.

G. J. Piedade, M. E. Schön, C. Lood, M. V. Fofanov, E. M. Wesdorp, T. E. G. Biggs, L. Wu, H. Bolhuis, M. G. Fischer, N. Yutin, B. E. Dutilh, and C. P. D. Brussaard. Seasonal dynamics and diversity of Antarctic marine viruses reveal a novel viral seascape. *Nature Communications*, 15(1):9192, Oct. 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53317-y.

D. Piel, M. Bruto, Y. Labreuche, F. Blanquart, D. Goudenège, R. Barcia-Cruz, S. Chenivesse, S. Le Panse, A. James, J. Dubert, B. Petton, E. Lieberman, K. M. Wegner, F. A. Hussain, K. M. Kauffman, M. F. Polz, D. Bikard, S. Gandon, E. P. C. Rocha, and F. Le Roux. Phage-host coevolution in natural populations. *Nature Microbiology*, 7(7):1075–1086, July 2022. ISSN 2058-5276. doi: 10.1038/s41564-022-01157-1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

S. Roux and C. Coclet. Viromics approaches for the study of viral diversity and ecology in microbiomes. *Nature Reviews. Genetics*, July 2025. ISSN 1471-0064. doi: 10.1038/s41576-025-00871-w.

S. Roux and J. B. Emerson. Diversity in the soil virosphere: to infinity and beyond? *Trends in Microbiology*, 30(11):1025–1035, Nov. 2022. ISSN 1878-4380. doi: 10.1016/j.tim.2022.05.003.

S. Roux, E. M. Adriaenssens, B. E. Dutilh, E. V. Koonin, A. M. Kropinski, M. Krupovic, J. H. Kuhn, R. Lavigne, J. R. Brister, A. Varsani, C. Amid, R. K. Aziz, S. R. Bordenstein, P. Bork, M. Breitbart, G. R. Cochrane, R. A. Daly, C. Desnues, M. B. Duhaime, J. B. Emerson, F. Enault, J. A. Fuhrman, P. Hingamp, P. Hugenholtz, B. L. Hurwitz, N. N. Ivanova, J. M. Labonté, K.-B. Lee, R. R. Malmstrom, M. Martinez-Garcia, I. K. Mizrachi, H. Ogata, D. Páez-Espino, M.-A. Petit, C. Putonti, T. Rattei, A. Reyes, F. Rodriguez-Valera, K. Rosario, L. Schriml, F. Schulz, G. F. Steward, M. B. Sullivan, S. Sunagawa, C. A. Suttle, B. Temperton, S. G. Tringe, R. V. Thurber, N. S. Webster, K. L. Whiteson, S. W. Wilhelm, K. E. Wommack, T. Woyke, K. C. Wrighton, P. Yilmaz, T. Yoshida, M. J. Young, N. Yutin, L. Z. Allen, N. C. Kyrpides, and E. A. Eloe-Fadrosh. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology*, 37(1):29–37, Jan. 2019. ISSN 1546-1696. doi: 10.1038/nbt.4306.

C. Santos-Medellin, L. A. Zinke, A. M. Ter Horst, D. L. Gelardi, S. J. Parikh, and J. B. Emerson. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. *The ISME journal*, 15(7):1956–1970, July 2021. ISSN 1751-7370. doi: 10.1038/s41396-021-00897-y.

C. Santos-Medellín, K. Estera-Molina, M. Yuan, J. Pett-Ridge, M. K. Firestone, and J. B. Emerson. Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. *Proceedings of the National Academy of Sciences of the United States of America*, 119(45):e2209132119, Nov. 2022. ISSN 1091-6490. doi: 10.1073/pnas.2209132119.

C. Santos-Medellín, S. J. Blazewicz, J. Pett-Ridge, M. K. Firestone, and J. B. Emerson. Viral but not bacterial community successional patterns reflect extreme turnover shortly after rewetting dry soils. *Nature Ecology & Evolution*, 7(11):1809–1822, Nov. 2023. ISSN 2397-334X. doi: 10.1038/s41559-023-02207-5.

E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B. W. Trawick, K. D. Pruitt, and S. T. Sherry. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 50(D1):D20–D26, Jan. 2022. ISSN 1362-4962. doi: 10.1093/nar/gkab1112.

J. Shang, C. Peng, H. Liao, X. Tang, and Y. Sun. PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data. *Bioinformatics Advances*, 3(1):vbad101, 2023a. ISSN 2635-0041. doi: 10.1093/bioadv/vbad101.

J. Shang, X. Tang, and Y. Sun. PhaTYP: predicting the lifestyle for bacteriophages using BERT. *Briefings in Bioinformatics*, 24(1):bbac487, Jan. 2023b. ISSN 1477-4054. doi: 10.1093/bib/bbac487. URL `https://doi.org/10.1093/bib/bbac487`.

J. W. Sorensen, L. A. Zinke, A. M. Ter Horst, C. Santos-Medellín, A. Schroeder, and J. B. Emerson. DNase Treatment Improves Viral Enrichment in Agricultural Soil Viromes. *mSystems*, 6(5):e0061421, Oct. 2021. ISSN 2379-5077. doi: 10.1128/mSystems.00614-21.

A. M. Ter Horst, C. Santos-Medellín, J. W. Sorensen, L. A. Zinke, R. M. Wilson, E. R. Johnston, G. Trubl, J. Pett-Ridge, S. J. Blazewicz, P. J. Hanson, J. P. Chanton, C. W. Schadt, J. E. Kostka, and J. B. Emerson. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome*, 9(1):233, Nov. 2021. ISSN 2049-2618. doi: 10.1186/s40168-021-01156-0.

A. M. Ter Horst, J. D. Fudyma, J. L. Sones, and J. B. Emerson. Dispersal, habitat filtering, and eco-evolutionary dynamics as drivers of local and global wetland viral biogeography. *The ISME journal*, 17(11):2079–2089, Nov. 2023. ISSN 1751-7370. doi: 10.1038/s41396-023-01516-8.

G. Trubl, S. Roux, M. A. Borton, A. Varsani, Y.-F. Li, C. L. Sun, H. B. Jang, B. J. Woodcroft, G. W. Tyson, K. C. Wrighton, S. R. Saleska, E. A. Eloe-Fadrosh, M. B. Sullivan, and V. I. Rich. Population ecology and biogeochemical implications of ssDNA and dsDNA viruses along a permafrost thaw gradient. *Nature Communications*, 17(1):375, Dec. 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-67057-0.

M. W. Van Goethem, T. L. Swenson, G. Trubl, S. Roux, and T. R. Northen. Characteristics of Wetting-Induced Bacteriophage Blooms in Biological Soil Crust. *mBio*, 10(6):e02287–19, Dec. 2019. ISSN 2150-7511. doi: 10.1128/mBio.02287-19.

L. M. Wahl and T. Pattenden. Prophage Provide a Safe Haven for Adaptive Exploration in Temperate Viruses. *Genetics*, 206(1):407–416, May 2017. ISSN 1943-2631. doi: 10.1534/genetics.116.197541.

I. J. Wang and G. S. Bradburd. Isolation by environment. *Molecular Ecology*, 23(23):5649–5662, Dec. 2014. ISSN 1365-294X. doi: 10.1111/mec.12938.

H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.