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, thereby experiencing distinct evolutionary pressures, has an unknown influence on the microdiversity present in the viral genome. This difference in microdiversity could impact the coevolution of phages with their hosts or the resilience of phages to new environments or hosts. To identify a possible association between microdiversity and lifestyle, we used four existing metagenomics datasets focusing on bacteriophages from soil environments as input for a newly constructed pipeline calculating microdiversity and predicting phage lifestyles. These four datasets together consist of 11 073 viral genomes and span across environments. We found that temperate phages consistently exhibit significantly higher microdiversity than their virulent counterparts in three of four datasets, a pattern that held 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 hold the title of most ubiquitous biological entities on earth, exerting immense influence over prokaryotic and eukaryotic life. Focusing on prokaryotic life, we find bacteriophages shunting and shuttling 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 in 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 meant 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. Microdiversity is the genetic diversity *within* a species, as opposed to *between* species with macrodiversity.

In microbial models, from both the bacterial and phage side, microdiversity is involved in maintaining the possibility for long-term host-virus coexistence [Avrani et al., 2011]. 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 bacterial models [Dion et al., 2020].

This 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]. First of all, the lytic lifestyle can be performed by both virulent and temperate phages: the phage infects the cell, starts replication of the 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, or excise itself out the host genome and return to the lytic lifestyle. A well-studied class of double-stranded DNA viruses that can have either a virulent or temperate lifestyle is *Caudoviricetes* [Roux and Coclet,

© The Author 2025.

2025]. 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., 1994, Brüssow et al., 2004]. How the lifestyle of bacteriophages affects the genetic microdiversity is unknown however.

Here we studied the association between phage lifestyle and microdiversity, using the integration of lysogenic lifestyles as starting point for differential microdiversity between virulent and temperate phages. Four soil-borne bacteriophage datasets were chosen, as soil-environments tend to contain a more even ratio between virulent and temperate phages [Liang et al., 2020]. Besides their focus on soil-borne bacteriophages, these four datasets were chosen due to their recent publication and FAIR data (i.e. data was Findable, Accessible, Interoperable and Reusable). The four datasets together contained 137 545 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 11 073 high-quality *Caudoviricetes* vOTUs were assigned a lytic or lysogenic lifestyle and microdiversity values. We show that three of the four datasets indicate temperate phages as having a higher microdiversity than virulent phages.

2. Methods

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 spring 2025) and data-availability of raw reads and vOTUs. Using these criteria, six 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: Intertidal, BodegaBay, Watershed and Land-use followed the predominant standard for clustering of contigs into vOTUs: at least 95% average nucleotide identity over 85% alignment fraction [Gregory et al., 2019]. Liao et al. [2025] used a different alignment fraction threshold of 80%. Santos-Medellin et al. [2021] used both a different average nucleotide identity threshold of 90% and a different alignment fraction threshold of 70%. Datasets were also re-clustered 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 prevent any artefacts caused by different vOTU clustering definitions between datasets [Fu et al., 2012].

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]. Two approaches were proposed for the mapping of reads against the vOTUs: a per-library approach where reads per individual library were mapped, and a per-dataset approach where all libraries were combined for the mapping. As the per-dataset approach retained the most vOTUs by pooling coverage, we will focus on the per-dataset approach here. Raw reads were mapped against the vOTUs 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 with mapped reads were filtered on a minimum coverage of 10x and breadth of coverage of at least 80% after mapping individual libraries (per-library approach) or after mapping all libraries at once (per-dataset approach). 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 of PhaStyle 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 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 COG0582 hit for temperate lifestyle or absence for 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]. 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.

Table 1. Overview of the six 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) in addition or instead of metagenomes (m). **ANI** = Average Nucleotide Identity, **AF** = Alignment Fraction.

Dataset ID	Source	BioProject ID	Clustering	# of libraries	# of vOTUs
Glacial	[Liao et al., 2025]	PRJNA882246	>= 95% ANI over >= 80% AF	24 (v)	59685
Intertidal	[Ji et al., 2024]	PRJNA957716	>= 95% ANI over >= 85% AF	96 (m)	20102
BodegaBay	[Ter Horst et al., 2023]	PRJNA913601	>= 95% ANI over >= 85% AF	63 (v), 63 (m)	12689
Watershed	[Coclet et al., 2023]	PRJNA1168581	>= 95% ANI over >= 85% AF	48 (m)	2743
Land-use	[Liao et al., 2022]	PRJNA691683	>= 95% ANI over >= 85% AF	50 (v)	61106
AgriSoil	[Santos-Medellin et al., 2021]	PRJNA646773	>= 90% ANI over >= 70% AF	15 (v), 16 (m)	4065

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 library or dataset analysed for the per-library and per-dataset approach respectively. A Wilcoxon rank-sum test was employed to calculate the statistical relationship between the two lifestyle groups and their respective microdiversity. This analysis was performed both on individual libraries and each dataset. P-values were corrected per dataset for multiple testing using the Benjamini-Hochberg procedure when performing the per-library approach. No correction was applied for the per-dataset approach.

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

Six datasets had been selected for their focus on bacteriophages and soil environments, recency of publication and availability of data. These six datasets formed the starting point of our analysis into the relationship between phage lifestyle and microdiversity. Of the six datasets, two contained only metagenomes, two contained both metagenomes and viromes and two contained only viromes (Table 1). The datasets are all from soil-focused studies, although the exact environment type differs (e.g. coastal, agricultural, urban). These datasets were used as input for the SRR2strain pipeline (Supplemental Figure 1, also see subsection 2.7). As part of this pipeline, sequenced reads are mapped back onto the vOTUs. This is used as input for the analysis of microdiversity downstream. During this mapping process, two of the datasets, Intertidal and Watershed (which both contained metagenomes only), showed extremely poor mapping efficiency (less than 2% mapped reads, Supplemental Table 2). This proved irreconcilable with the thresholds put in place on minimum mapping breadth and coverage. For this reason, 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 2).

3.2. Overview of high quality viral datasets

The final collection of datasets consisted of four datasets, containing a total of 152 viromes, which contain 137 545 vOTUs. After performing filtering on breadth, coverage and completeness of vOTUs, 30 762 vOTUs (22.3%) remain. These 30 762 vOTUs were not distributed equally among the four studies however, as 49% and 45% originate from the Glacial and Land-use datasets, respectively. On the dataset level, mapping breadth and estimated vOTU completeness showed similar values among the four studies, while average vOTU length and average coverage showed more variability between datasets (Figure 1 and Supplemental Table 1). Average vOTU length was similar among BodegaBay, Land-use and AgriSoil datasets, but showed a markedly higher value for the Glacial dataset. Average mapping coverage showed values between 112x (BodegaBay) and 141x (Glacial).

Re-clustering the initial 137 545 vOTUs using a single set of thresholds (subsection 2.1) resulted in 137 195 vOTU clusters, of which 136 860 were singletons (99.8%). Of the 335 non-singleton clusters, 321 contained two vOTUs, 13 contained three vOTUs and 1 contained four vOTUs. No clustering between datasets was observed and therefore only the unclustered datasets were considered for subsequent analyses. Overall, we are confident the stringent thresholds we have

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. Last two columns are based on total number of *Caudoviricetes* vOTUs.

Dataset ID	# of vOTUs	# of *Caudoviricetes* vOTUs	# of temperate vOTUs	# of virulent vOTUs
Glacial	6444	6380	1075	5305
BodegaBay	253	249	45	204
Land-use	4262	4249	1053	3196
AgriSoil	195	195	52	143

chosen have resulted in a collection of vOTUs of high quality. This high quality viral dataset was subsequently used in additional analyses of lifestyle and microdiversity.

Taxonomic classification was performed using geNomad and resulted in successful classification of 30 655 vOTUs (99.6% of 30 762 total). Of these 30 655 vOTUs, 30 545 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.

3.3. Virulent phages comprise majority in lifestyle distribution

The lifestyle, either virulent or temperate, was predicted for all vOTUs (i.e. before quality filters are applied) in the four selected datasets using a combination of tools (PhaTYP, PhaStyle and DIAMOND). This resulted in 120 358 vOTUs with a predicted lifestyle from either PhaTYP and PhaStyle, not including vOTUs that are filtered out or below default thresholds. The DIAMOND BLASTn search with known integrase and recombinase sequences (COG0582) showed significant hits for 11 143 vOTUs. Taking the overlap of the two tools and using the BLAST results as verification of temperate phages results in 34 760 *Caudoviricetes* vOTUs with a predicted lifestyle. Applying quality thresholds on mapping breadth, coverage and estimated vOTU completeness resulted in 11 073 vOTUs, of which 2225 (20%) were classified as having a temperate lifestyle and 8848 (80%) as having a virulent lifestyle (Table 2). The ratio between virulent and temperate phages per study was similar between all four 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 5). From our four datasets, 14 136 vOTUs that passed the 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 4).

3.4. Soil-borne bacteriophage microdiversity shows association with lifestyle

The microdiversity (π) was calculated for all vOTUs in the dataset in an effort to identify and analyse any association with phage lifestyle. The microdiversity analysis was performed using inStrain and the "scaffold_info" output file used as primary output data [Olm et al., 2021]. inStrain was run using all 137 545 vOTUs and filtered for the 11 073 vOTUs that passed the quality thresholds, lifestyle prediction and were classified as *Caudoviricetes*.

Generally, microdiversity levels were similar between studies and consistent with the microdiversity as reported in the original studies respectively [Liao et al., 2022, 2025]. BodegaBay showed the lowest level of microdiversity, with a median around 0.00074, while Land-use had the highest median microdiversity, around 0.0027. Overall, the studies showed quite a wide range of microdiversity (Figure 2a).

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. The predicted genes and corresponding annotations were collected and combined with the coverage of filtered reads (identity >95%) and microdiversity as computed by inStrain (Figure 2c). 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. Overall, there seems to be somewhat of a 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 our correlation analysis (Supplemental Figure 4).

On the dataset level (using the per-dataset approach), three out of four datasets showed a significant difference in the level of microdiversity detected for each lifestyle (Figure 2b and Supplemental Table 3), with the exception of the AgriSoil dataset ($p = 0.17$, Wilcoxon rank-sum test, two-tailed). The direction of the difference in microdiversity of these three datasets was consistent, all showing a higher level of microdiversity in temperate phages as opposed to virulent phages. The Land-use showed the most distinct microdiversity distributions between the two lifestyles ($p = 1.4e\text{-}58$, estimate = 0.0012, Wilcoxon rank-sum test, two-tailed). The Glacial dataset contained the most vOTUs, which influenced the statistical significance of the difference in microdiversity between the lifestyles, even though it resulted in a lower estimated true difference in microdiversity ($p = 4.7e\text{-}114$, estimate = 0.0007, Wilcoxon rank-sum test, two-tailed). This association between lifestyle and microdiversity was robust when using different lifestyle prediction methods (see Supplemental Table 3).

In addition to vOTU wide microdiversity analysis, 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

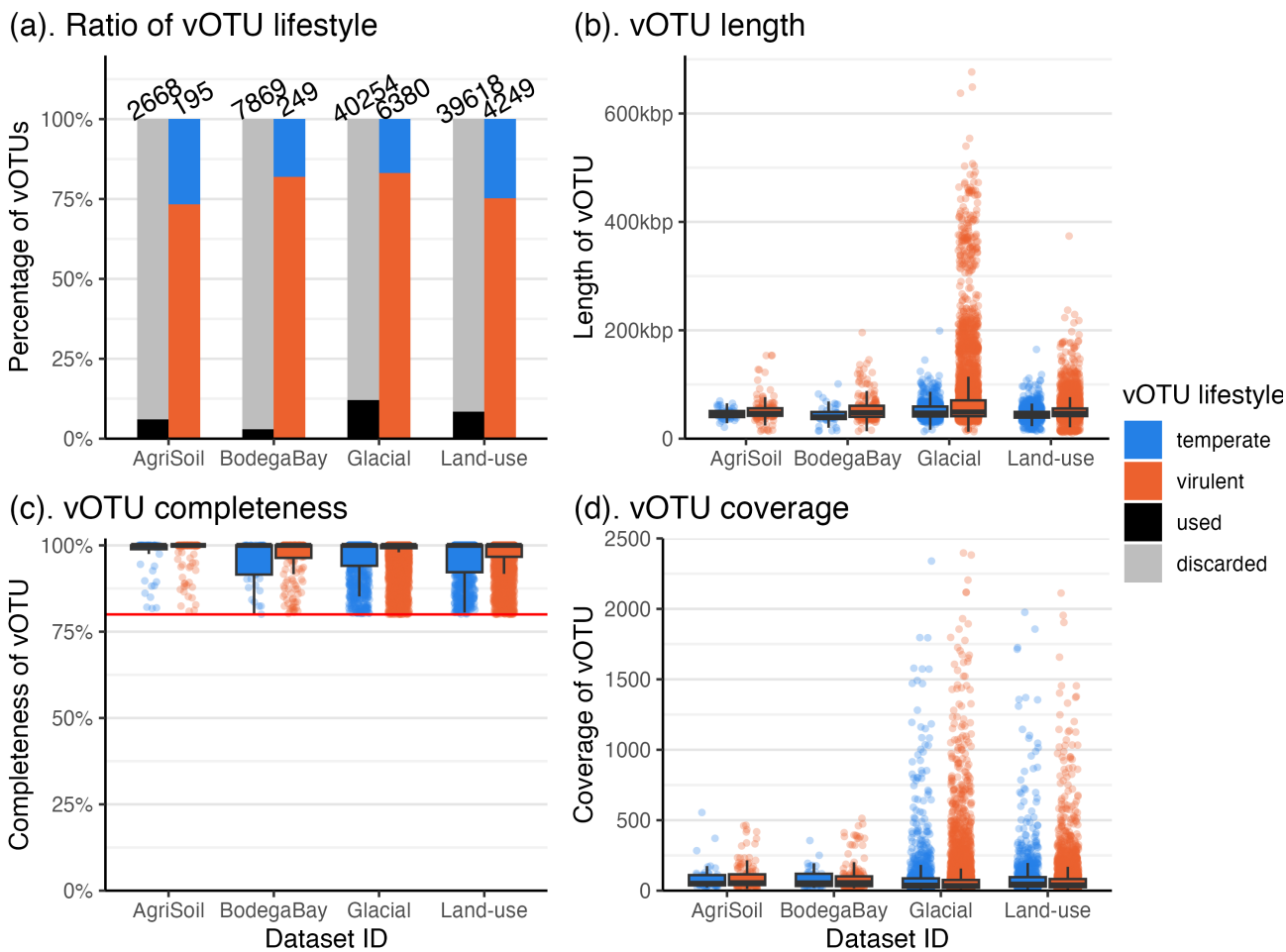

Fig. 1. Overview of vOTU related statistics. All graphs use quality (breadth, completeness, coverage) and lifestyle filtered data. (**a**). Ratios of total and used data, and the ratio between temperate and virulent classified vOTUs. The numbers indicate discarded vOTUs due to quality issues and total vOTUs used for microdiversity analysis. (**b**). Distribution of vOTU lengths among the four datasets. Strikingly, only the Glacial study, which itself focuses on huge phages, shows viral genome lengths above 200kbp. (**c**). Predicted vOTU completeness. The used threshold of 80% is indicated with the red line. The AgriSoil and BodegaBay studies contain more vOTUs skewed towards the 100% completeness, while the Glacial and Land-use studies contain more vOTUs throughout the 80-100% range. (**d**). vOTU mapping coverage. The threshold for coverage was put at 10x, but the plot shows coverages far above this value.

1 indicates positive selection. Ratios around 1 are considered to be signs 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). Small but significant differences were identified between temperate and virulent lifestyles (Wilcoxon rank sum, see also Figure 3).

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 (see Supplemental Table 3). Using stringent filters for both read mapping and lifestyle prediction, we obtained confident association statistics for four datasets; three of the four showed significant lifestyle-associated differences, with temperate phages displaying higher microdiversity (Figure 2). This estimated difference in microdiversity between the two lifestyles was approximately ∼40% across those three 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].

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 a number of studies represented only by metagenomes 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 2). When filtering for high

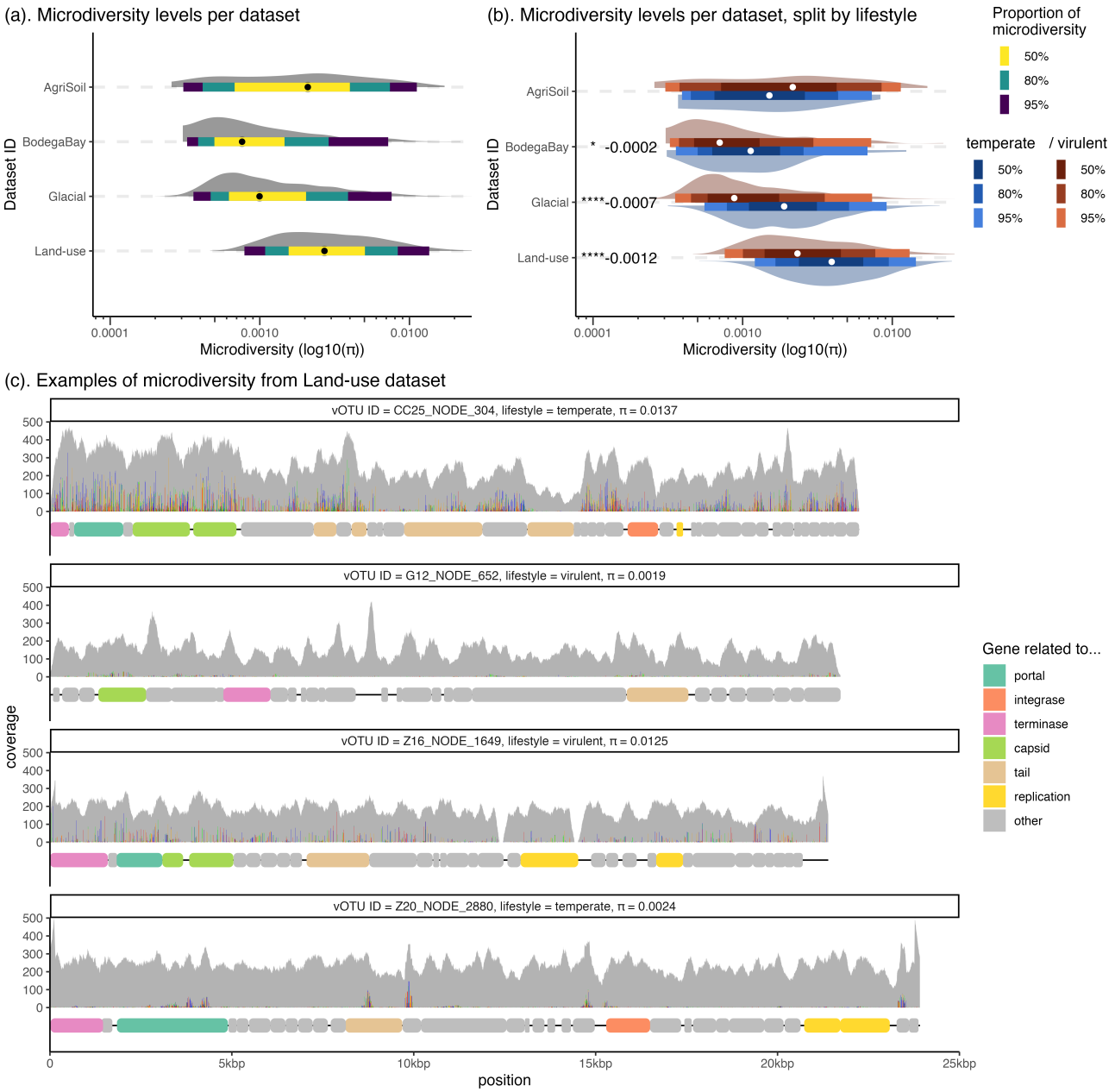

Fig. 2. Higher microdiversity in temperate phages than virulent phages in three out of four tested datasets. Both (**a**) and (**b**) show interval plots with added density distribution and dot indicating the median microdiversity value of the distribution. Value next to significance indication shows 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 four tested datasets. Glacial, BodegaBay and Land-use show similar distributions, all being left skewed. AgriSoil shows a more uniform distribution, showing only a slight peak at 0.003. (**b**). Microdiversity distribution split out by predicted lifestyle of vOTUs. The Glacial, BodegaBay and Land-use dataset show the same pattern of differential microdiversity between temperate and virulent phages. This is also visually identifiable in the more bell-shaped and 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)

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,

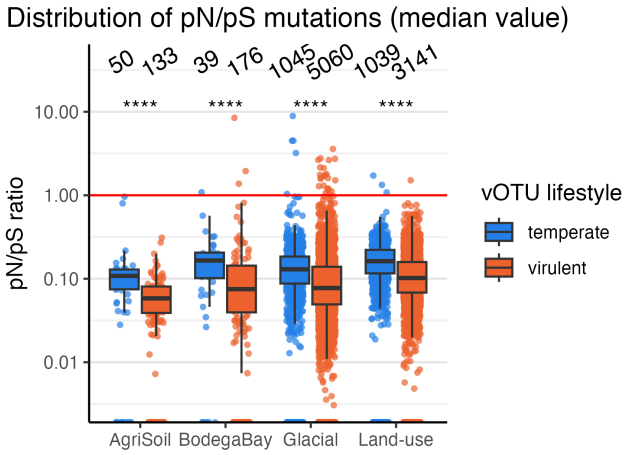

Fig. 3. pN/pS ratios per dataset split by lifestyle. Temperate phages show consistently higher pN/pS ratios than virulent phages, although differences are small. This was also statistically tested with a Wilcoxon rank sum test, resulting in significant differences in the pN/pS values of all four datasets between temperate and virulent lifestyles (glacial: p-value = 0, land-use: p-value = 0, AgriSoil: p-value = 3.05e-14, BodegaBay: p-value = 1.91e-10).

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. Altogether, these results still highlight the importance and necessity of 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. Take the mapping process as example; the pipeline was benchmarked with and only includes a single 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%. Another related topic is the definition of vOTU sequences as set by the respective studies (Table 1). Especially of interest is the aberrant vOTU definition used by Santos-Medellin et al. [2021] in their AgriSoil dataset, with low average nucleotide identity compared to other studies. The AgriSoil dataset is also the only of the four virome datasets not to show the same microdiversity-lifestyle association. Additional analyses should be performed using an updated definition of vOTU clustering. These analyses will aid in determining whether it is the vOTU definition influencing the microdiversity-lifestyle association or a true signal present in the dataset. 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.

As this study uses a meta-analysis, it is limited in the specialized experimental variables it can capture. 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 functioning as positive and negative control [Buchfink et al., 2021, Shang et al., 2023b, Juhász et al., 2024]. Despite similar conceptual underpinnings, PhaTYP and PhaStyle yielded substantial variance in lifestyle calls (Supplemental Figure 2). 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. These discrepancies underscore the need for application of multiple tools for these kinds of predictions, but perhaps also hesitancy with exclusively sequence based tools. 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 genetics related 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]. Thirdly, 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 11 073 vOTUs fell between zero and one, indicating that the vOTUs are under purifying selective pressure (Figure 3). Besides this, the pN/pS ratios of all four studies showed a significant difference between the two predicted lifestyles. Unfortunately, it is difficult to link these pN/pS ratios to any specific (a)biotic factor across these four datasets, since not every dataset has high-precision data available on host states or abiotic factors.

In addition to these neutral process, the effect of a highly spatial heterogeneous environment should also be considered. Prior work has shown that soil viral communities can harbour substantially greater microdiversity than those from marine systems, likely reflecting the extreme spatial heterogeneity characteristic of soils [Gregory et al., 2019, Peng et al., 2023, Zhong et al., 2023, Piedade et al., 2024]. Our results align with previous studies performed in soil environments: when compared with microdiversity statistics from marine-focused publications, 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 provides a plausible context for the pronounced lifestyle signal we observe, given that spatial heterogeneity can interact with the potential microdiversity influencing processes previously described [Haerter and Sneppen, 2012, Wang and Bradburd, 2014, García-García et al., 2019, Santos-Medellín et al., 2022].

Given the pronounced heterogeneity of soils, we also considered the role of anthropogenic disturbance across the four datasets analysed. Among them, the AgriSoil dataset is the most likely to be influence by direct human activity: the AgriSoil study explicitly manipulated agricultural fields through the addition of carbon-rich amendments, which may help explain its divergent lifestyle-microdiversity relationship relative to the other datasets (Figure 2) [Santos-Medellin et al., 2021]. The Land-use dataset spans forest, agricultural and urban soils, so is likely to be marginally influenced by anthropogenic disturbance [Liao et al., 2022]. The BodegaBay dataset derives from a protected coastal reserve used for ecosystem research [Ter Horst et al., 2023]. The Glacial dataset is the least likely to be influenced by any anthropogenic activity, given its remote sampling location [Liao et al., 2025]. Disturbance regimes can restructure host availability, alter induction rates, and shift selection-drift balances, any of which could influence lifestyle-associated microdiversity patterns [Hughes et al., 2008, García-García et al., 2019, Santos-Medellin et al., 2021, Zhong et al., 2023, Geonczy et al., 2025]. More studies are necessary to fully untangle different drivers of microdiversity. When this is not possible using environmental samples, concepts like mathematical or agent-based models could be employed as they allow for more fine-grained control of experimental variables [Doekes et al., 2021].

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. Various neutral processes were discussed as possible contributors 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. 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.17988607`). 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.

S. Avrani, O. Wurtzel, I. Sharon, R. Sorek, and D. Lindell. Genomic island variability facilitates Prochlorococcus-virus

coexistence. *Nature*, 474(7353):604–608, June 2011. ISSN 1476-4687. doi: 10.1038/nature10172.

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. Publisher: Nature Publishing Group.

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.

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.

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, 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. 2025. 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.

A. R. Hughes, B. D. Inouye, M. T. J. Johnson, N. Underwood, and M. Vellend. Ecological consequences of genetic diversity. *Ecology Letters*, 11(6):609–623, June 2008. ISSN 1461-0248. doi: 10.1111/j.1461-0248.2008.01179.x.

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. Publisher: Taylor & Francis _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.

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.

Y. Peng, Z. Lu, D. Pan, L.-D. Shi, Z. Zhao, Q. Liu, C. Zhang, K. Jia, J. Li, C. R. J. Hubert, and X. Dong. Viruses in deep-sea cold seep sediments harbor diverse survival mechanisms and remain genetically conserved within species. *The ISME journal*, 17(10):1774–1784, Oct. 2023. ISSN 1751-7370. doi: 10.1038/s41396-023-01491-0.

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.

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.

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.

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, 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.

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.

Z.-P. Zhong, D. Vik, J. Z. Rapp, O. Zablocki, H. Maughan, B. Temperton, J. W. Deming, and M. B. Sullivan. Lower viral evolutionary pressure under stable versus fluctuating conditions in subzero Arctic brines. *Microbiome*, 11 (1):174, Aug. 2023. ISSN 2049-2618. doi: 10.1186/s40168-023-01619-6.