PLOS COMPUTATIONAL BIOLOGY

METHODS

DeepPL: A deep-learning-based tool for the prediction of bacteriophage lifecycle

Yujie Zhang[1], Mark Mao[2], Robert Zhang[2], Yen-Te Liao[1], Vivian C. H. Wu[1]*

1 Produce Safety and Microbiology Research Unit, U.S. Department of Agriculture, Agricultural Research Service, Western Regional Research Center, Albany, California, United States of America, **2** Clowit, LLC. Burlingame, California, United States of America

* vivian.wu@usda.gov

Citation: Zhang Y, Mao M, Zhang R, Liao Y-T, Wu VCH (2024) DeepPL: A deep-learning-based tool for the prediction of bacteriophage lifecycle. PLoS Comput Biol 20(10): e1012525. https://doi.org/10.1371/journal.pcbi.1012525

Editor: Yang Lu, University of Waterloo, CANADA

Received: April 10, 2024

Accepted: September 30, 2024

Published: October 17, 2024

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1012525

Data Availability Statement: All data and codes used for this study are available online. The source code of DeepPL is available via https://github.com/Wu-Microbiology/DeepPL. The model can be downloaded at https://figshare.com/articles/software/DeepPL_model/27005053?file=

Abstract

Bacteriophages (phages) are viruses that infect bacteria and can be classified into two different lifecycles. Virulent phages (or lytic phages) have a lytic cycle that can lyse the bacteria host after their infection. Temperate phages (or lysogenic phages) can integrate their phage genomes into bacterial chromosomes and replicate with bacterial hosts via the lysogenic cycle. Identifying phage lifecycles is a crucial step in developing suitable applications for phages. Compared to the complicated traditional biological experiments, several tools have been designed for predicting phage lifecycle using different algorithms, such as random forest (RF), linear support-vector classifier (SVC), and convolutional neural network (CNN). In this study, we developed a natural language processing (NLP)-based tool—DeepPL—for predicting phage lifecycles via nucleotide sequences. The test results showed that our DeepPL had an accuracy of 94.65% with a sensitivity of 92.24% and a specificity of 95.91%. Moreover, DeepPL had 100% accuracy in lifecycle prediction on the phages we isolated and biologically verified previously in the lab. Additionally, a mock phage community metagenomic dataset was used to test the potential usage of DeepPL in viral metagenomic research. DeepPL displayed a 100% accuracy for individual phage complete genomes and high accuracies ranging from 71.14% to 100% on phage contigs produced by various next-generation sequencing technologies. Overall, our study indicates that DeepPL has a reliable performance on phage lifecycle prediction using the most fundamental nucleotide sequences and can be applied to future phage and metagenomic research.

Author summary

Bacteriophages are viruses that infect bacteria and play a critical role in the microbial community within different environments via phage-bacterial evolutionary interactions. The classification of phage lifecycles is of great importance in deploying the potential applications of phages and better understanding complex microbial interactions. However, the traditional biological methods for phage lifecycle identification are complicated and time-consuming. In this study, we proposed a deep learning-based tool—DeepPL—for predicting phage lifecycles using the phage nucleotide genome. Compared with other

49153420. The detailed dataset information, including the NCBI accession number of the phage sequences, verified lifecycle, the usage of training or testing, and the group number used for 5-fold cross-validation, was provided in S1 Data.

Funding: This work was supported by the U.S. Department of Agriculture-Agricultural Research Service Current Research Information System projects (2030-42000-055-000-D to YZ, YL, and VCHW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

bioinformatic tools, DeepPL was developed from the pre-trained transformers model—DNABERT—designed for fundamental nucleotide language combined with representative phage complete genomes. Our in-house biological results were further used to verify the output of DeepPL. Overall, DeepPL performs with high precision for phage lifecycle prediction and could contribute to the genomic data-driven direction of phage research and applications.

Introduction

Bacteriophages (or phages) are viruses that infect bacteria and are widely prevalent in different environments, such as oceans, lakes, and agricultural soil, with an estimated number of 10^{31} virions in the biosphere [1,2]. There is an increasing number of literature and studies investigating the microbial component from different samples, such as the gut microbiome of humans and farm animals, indicating the vital role of the bacteriophage in the microbiome [3,4]. Phages can influence microbial populations by infecting specific bacterial hosts through two different lifecycles—lytic cycle and lysogenic cycle—based on their nature. Virulent phages (or strictly lytic phages) enter the lytic cycle that utilizes machinery from bacterial hosts to replicate and produce new virions before lysing the host cell. Temperate phages (or lysogenic phages) enter the lysogenic cycle by integrating their DNA into the host genome as prophages (also known as lysogenization). The prophages can be further transmitted to daughter cells at each subsequent bacterial cell division [5]. Under certain environmental stresses, prophages can converse to the lytic cycle by carefully being excised or induced from the bacterial genome to proliferate new infectious phage particles [6].

While lytic phages can infect and promptly lyse bacteria, they are promising antimicrobial agents against antibiotic-resistant bacterial strains. Therefore, lytic phage-based applications have gained attention in different areas, such as phage therapy in clinics and phage-based biocontrol in agriculture [7,8]. For example, a U.S. patient was infected by a superbug of *Acinetobacter baumannii*, which was resistant to all antibiotics. After treating with a lytic phage cocktail, the patient overcame this superbug infection and recovered [7]. However, the lysogenization capability enables temperate phages to be used for several applications, including phage display, genetic manipulation, and pathogen detection [9–11]. Phage display has been applied in drug discovery and antibody production to treat human diseases [12].

Identifying the lifecycle of newly isolated phages is a critical step highly associated with subsequent applications. Previously, plaque morphology via plaque assay was used as a traditional biological method to identify phage lifecycle: virulent phages produce clear plaques, and temperate phages produce turbid plaques [13]. Due to the inaccurate results from plaque morphology identification, several methods have been developed for isolating and screening the potential lysogens, such as Digoxigenin (DIG)-labeling in situ hybridization and patch test [14,15]. Though these methods improve accuracy and sensitivity, they are costly and time-consuming. Taking advantage of next-generation sequencing (NGS) development, the whole genome sequencing of isolated phages becomes helpful in determining the phage lifecycle by genomic features. The most common strategy is to compare new phage sequences with the reference phage genomes, which have known lifecycles, and detect the presence of lysogenic genetic modules through basic local alignment search tools (BLAST). In recent years, machine learning (ML) has been utilized to develop tools for classifying phage lifecycles into two categories based on the input file types (nucleotide sequence or amino acid sequences of complete phage genomes). The first ML-based tool—PHACTS—was developed to categorize phage

lifecycles using the amino acid sequences from the annotated phage genome via a random forest algorithm [16]. This tool was published in 2012; however, its accuracy dropped due to the lack of subsequent maintenance. In 2020, Tynecki et al. proposed a novel tool called phageAI, which was able to classify phage lifecycles via the combination of the Word2Vec Skip-gram model and a linear support-vector classifier (SVC) with an average accuracy of 98.90% on the small size of validation sets [17]. Deephage, developed by Wu et al., used a "one-hot" encoding and a convolutional neural network (CNN) to predict phage lifecycles, but it only achieved the best performance of 89% [18].

Bidirectional Encoder Representations from Transformers (BERT) has been widely used in natural language processing (NLP). It is composed of two steps: pre-training and fine-tuning. Notably, the BERT model was pre-trained in many different NLP tasks. Fine-tuning with a relatively small training set can yield great results for new downstream tasks. This distinctive feature of the unified architecture across various downstream tasks demonstrates its potential applications in different areas, including the biological language. Most recently, a newly published tool—PhaTYP—was designed to predict the lifecycles of phages using protein sequences via BERT, showing 98% accuracy [19]. However, protein-based (amino acid) sentences heavily rely on the well-known biological function of the proteins related to the lysogenic cycle, while most newly identified phage sequences do not have known counterparts. Nucleotide sequences (also called DNA sequences), constructed by assembling the four nucleotides (A, T, G, C base), are the most fundamental biological elements and contain all genetic information necessary for encoding functional molecules, including protein [20,21]. Moreover, compared to the BERT, DNABERT has been developed by training BERT using human DNA genomes [22]. Taking advantage of DNABERT with a better understanding of genomic DNA sequences, in this study, we developed a DNABERT-based tool, DeepPL, to identify the phage lifecycle with the input of phage nucleotide sequences (Fig 1). The improvement from amino acid sequences to nucleotide sequences overcame the missing details from translation and imperfect annotation of phage protein but also explored the fundamental instructions of non-coding DNA language related to the phage lifecycle. The pre-trained transformers model designed for DNA language combined with representative phage complete genomes will significantly enhance the accuracy of phage lifecycle classification and further contribute to the data-driven direction of phage-based research and application.

Materials and methods

Data collection

For the training dataset, the datasets from Deephage were collected and further subjected to the following trimming process [18]. First, the downloaded phage sequences were manually curated with reliable lifecycle annotations through the literature review. Second, based on the nature of virulent and temperate phages, the key genetic difference between these two phage types is the lysogenic genes within the temperate phage genomes. Therefore, several representative lysogenic gene markers for temperate phages were selected, including integrase, excisionase, recombinase, regulatory protein cro, antitermination protein Q, cI repressor, cII protein, cIII protein replication protein O, replication protein P, and recombination protein Bet. Third, we extracted these lysogenic genes from the phage genomes prior to manually correcting and removing the incorrect data deriving from their previous gene annotation. Finally, we combined the nucleotide sequences of each selected gene to generate a lysogenic dataset for our current study. After the above processes, we obtained a benchmark dataset with 1,488 lysogenic genes from 557 temperate phage genomes, which were further used for model training. Compared to temperate phages, more virulent phages with biological verification have been

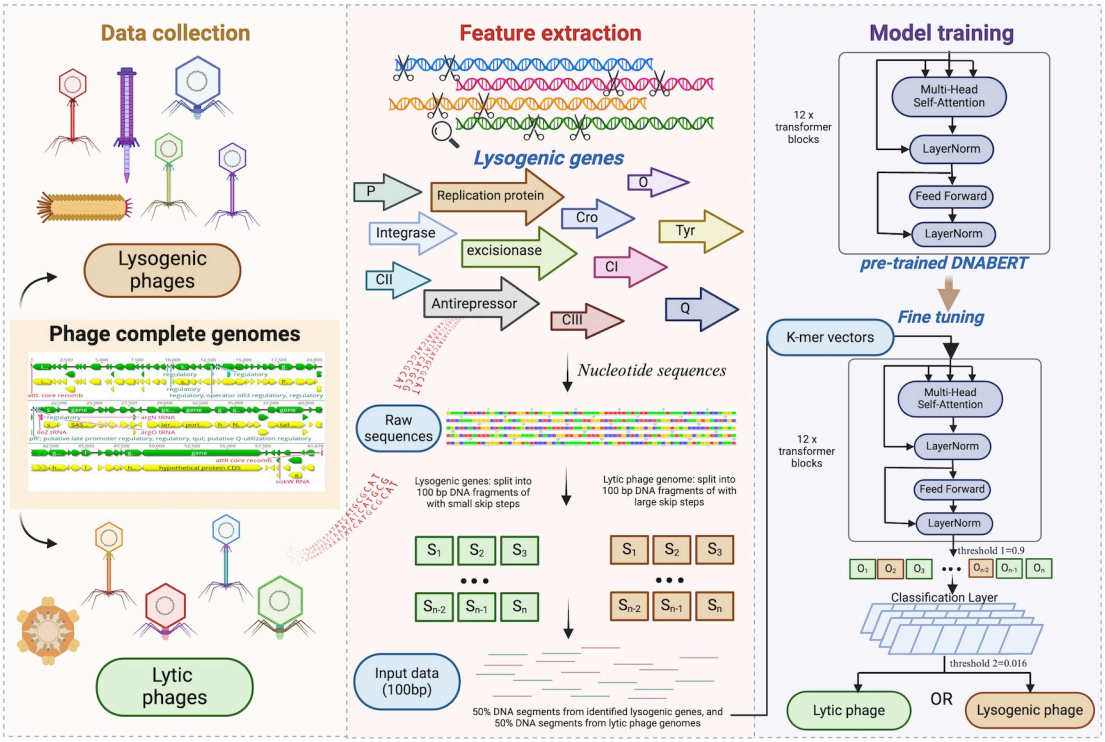

Fig 1. An overview of DeepPL framework for predicting phage lifecycle, including data collection, feature extraction, and model training. Diverse lysogenic and lytic complete phage genomes were collected from the National Center for Biotechnology Information (NCBI) database. Non-ATGC letters within the phage nucleotide sequences were randomly replaced by ATGC letters. The phage lifecycles were manually confirmed by the literature review. Further, the lysogenic genes were identified and extracted from lysogenic phage genomes. The sliding window of 100 bp in length and further conversion of sets of k-mer 6 sequences from phage sequences were used as input for a fine-tuning training process based on the pre-trained DNABERT model. The process generated the binary classification probability (0–1) of each k-mer 6 sequence. Therefore, threshold 1 was used to identify a good match between phage sequences and lysogenic genes. The threshold 1 of the binary classification probability above 0.9 was identified as a good match between 100 bp DNA segments and lysogenic genes. Further, the results from the frame-by-frame classification results were aggregated into one final classification result for phage lifecycle prediction with the threshold 2 of 0.016. The input phage sequence with a threshold 2 below 0.016 was identified as a lytic phage; otherwise, it was predicted as a lysogenic phage. The figure was created using BioRender. Zhang, Y. (2024) BioRender.com/w89b419.

https://doi.org/10.1371/journal.pcbi.1012525.g001

reported in the National Center for Biotechnology Information (NCBI) database. Therefore, a total of 1,262 virulent phage genomes with the strictly lytic cycle were verified and used to balance the training dataset. To test the performance of our model and compare it with other state-of-the-art tools, an additional 374 complete phage genomes with reliable lifecycle labels were collected from the NCBI database as the test dataset, including 245 virulent phages and 129 temperate phages.

All the complete phage genomes were downloaded from the NCBI database. Detailed information on each phage genome was provided in S1 Data, including the phage accession number, verified lifecycle, the usage of training or testing, and the group number used for 5-fold cross-validation.

Model training and testing

Taking advantage of DNABERT with trained knowledge of genome sequences at the nucleotide level, we started with the DNABERT model and used phage nucleotide sequences to do the fine-tuning training process for our specific classification task. Specifically, the phage nucleotide sequences have been pre-checked for the presence of non-ATGC letters for noise

reduction. The sequences of less than 10 non-ATGC letters were replaced with random letters of ATGC. The sequences of more than 10 non-ATGC letters were removed from the training dataset. To classify the phage lifecycle (lysogenic cycle vs. lytic cycle), a balanced training set, consisting of 50% nucleotide sequences from identified lysogenic genes and 50% nucleotide sequences from lytic phage genomes, was obtained by sliding window-based selection and further employed for model training. K-mer 6 was selected in this study based on its best performance, as reported by DNABERT. The sequence sizes between 100 to 512 bp were compared in the fine-tuning process. The comparison of different parameters indicated that the sequence size of 100 bp yielded reasonably balanced results with less computing requirements and was further used in the current study. To that end, the input of training sequences was obtained through a sliding window of 100 bp in length and turned into a set of k-mer 6 sequences. In addition, the lysogenic genes extracted from lysogenic phage genomes are much smaller than the complete genomes of lytic phages. Therefore, different skip steps were selected for lysogenic genes and lytic phage genomes to get a 50%-50% balanced training dataset (547,810 DNA sequences from lytic phage genome and 500,765 DNA sequences from lysogenic genes) with the full coverage of lysogenic genes in this study. Specifically, a small skip step of 1 (i.e., genome location, base pairs 1–100, 2–101, 2–102, etc.) was used for the sliding window in the lysogenic gene to cover limited lysogenic features. Further, a large skip step of 91 (i.e., genome location, base pairs 1–100, 92–191, 183–282, etc.) was selected for the sliding window in the lytic phage genomes to match the number of DNA sequences from lysogenic genes. After evaluation, around 10 epochs of training were conducted during the fine-tuning process, with a larger learning rate in early epochs and a smaller learning rate in later epochs.

The same setting of a sliding window of 100-bp sequence size and the subsequent conversion of k-mer 6 sequence sets was for the model testing. In detail, the entire set of k-mer 6 sequences was run using our model to generate the binary classification probability (0–1) of each k-mer 6 sequence. By optimizing the threshold values, high sensitivity and accuracy were achieved and subsequently used in this study based on the following settings: threshold 1 of 0.9 was set for good matching between each 100 bp segment and lysogenic genes, and threshold 2 of 0.016 was used to aggregate the frame-by-frame classification results into one final classification result for phage lifecycle prediction. The phages with a probability below 0.016 were predicted as lytic phages, whereas phages with a probability equal to or above 0.016 were predicted as lysogenic phages.

Performance evaluation

The test dataset was used to evaluate the performance of DeepPL and compare it with other tools for phage lifecycle prediction. A total of five performance measures, including sensitivity (SN), specificity (SP), accuracy (ACC), F-score, and MCC, were selected to quantify the performance in this study [23]. For evaluation, temperate and virulent phages are referred to as positive and negative samples, respectively.

Results

Performance assessment

The 5-fold cross-validation was conducted to evaluate the performance stability of DeepPL by randomly dividing the training dataset (1,262 virulent phage genomes and 1,488 lysogenic genes from 557 temperate phage genomes) into five equal-sized groups (around 450,000 DNA sequences) (S1 Data). Each group contained approximately 360 phage sequences, consisting of 50% of the lysogenic gene sequences (i.e., around 229,055 sequences in Group 1) extracted from lysogenic phage genomes and 50% of DNA fragments (i.e., around 226,626 sequences in

Table 1. The performance comparison between DeepPL and previously published tools for phage lifecycle prediction.

Tools	Sensitivity (%)	Specificity (%)	Accuracy (%)	F-score	MCC
DeepPL	**92.24**	95.91	94.65	**0.92**	**0.53**
PhaTYP	90.44	97.47	**94.91**	**0.92**	**0.53**
Deephage	78.61	**98.13**	89.83	0.86	0.49
PHACTs	38.94	79.77	48.66	0.53	0.14
PhageAI	83.33	96.08	91.17	0.87	0.50

https://doi.org/10.1371/journal.pcbi.1012525.t001

Group 1) extracted from lytic phage genomes. Specifically, each group (i.e., Group 1) was used as the test dataset, while the rest of the groups (i.e., Groups 2–5) were used as training datasets. The result showed that the sensitivity of the five groups was from 85% to 96%, while the specificity of each group ranged from 89% to 94% (S2 Data). Moreover, the accuracy from 90.6% to 93.9% indicated that DeepPL has an overall reliable and stable performance for phage lifecycle prediction.

Comparison with previously published methods

To estimate the predictive capability of DeepPL, four current available tools designed for phage lifecycle prediction, including PhaTYP, Deephage, PhageAI, and PHACTs, were selected to compare the performance with DeepPL in this study using the test dataset (Table 1). The performance comparison revealed that our DeepPL and PhaTYP yielded better results than the other three tools, suggesting the advantage of NLP applied in the genomic language. DeepPL had a slightly higher sensitivity than PhaTYP, while the specificity of PhaTYP was slightly better than that of DeepPL. The accuracy of DeepPL and PhaTYP showed the best results, around 94%, and outperformed the other tools. Most importantly, compared to the amino-acid-based PhaTYP, DeepPL can achieve a similar performance as PhaTYP using DNA sequences. In addition, the ROC curve comparison on the test dataset was also derived to illustrate the trade-offs between the sensitivity and specificity of each tool (Fig 2). The AUCROC score showed DeepPL had the best performance (0.98), followed by DeePhage (0.97) and PhaTYP (0.85). Together, the results demonstrated that DeepPL could capture the underlying

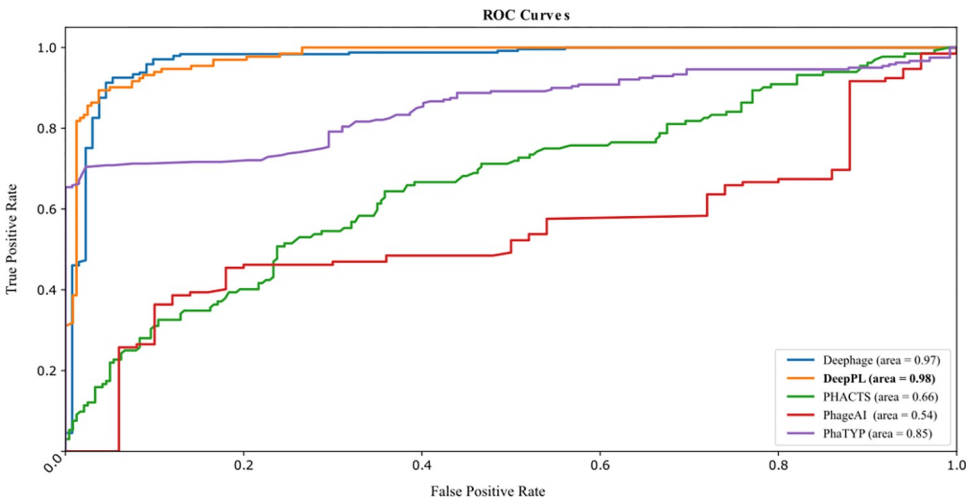

Fig 2. The ROC curve comparison between DeepPL and previously published tools for phage lifecycle prediction on test dataset. The value shown in the legend is AUCROC score.

https://doi.org/10.1371/journal.pcbi.1012525.g002

genomic differences between virulent and temperate phage genomes at the fundamental nucleotide level without trading off the accuracy.

Case study 1-In-house verified phages

In this study, 19 genomes from the phages isolated by our lab were used via DeepPL prediction to validate the prediction results and evaluate the potential application in phage studies (Table 2). Biological and genomic characterizations were used to confirm the lifecycles of 19 phages isolated from different sources. The biological experiments included the spot test, plaque assay, bacterial challenge assay, and lysogen test. The genomic analysis included whole-genome sequencing, lysogenic gene detection, and comparative genomics with the published reference phages and bacterial host. Among these phages, only the phage Sa179lw has a questionable lifecycle. In detail, this phage showed lytic activity against *E. coli* O179 strains via biological tests but was detected with the presence of the lysogenic module by the genomic analysis. The results showed that DeepPL had 100% accuracy in predicting the lifecycles of the 18 in-house verified phages. Moreover, the lysogenic potential of the questionable phage Sa179lw was detected by DeepPL but not by PhaTYP, due to the high sensitivity of DeepPL. Overall, the prediction results demonstrated that our DeepPL can be a promising tool for phage research and phage-based applications.

Case study 2-Metagenomic data

Viral metagenomics has been developed to understand viral diversity and ecology within diverse environments, such as mammalian gastrointestinal (GIT) and agricultural environments. In this case, DeepPL was employed to evaluate its potential application for phage

Table 2. The performance validation of DeepPL on phage lifecycle prediction using in-house verified phages.

Phages	Accession number	Sequence length (bp)	DeepPL prediction	PhaTYP prediction	Confirmed lifecycle	References
Escherichia phage vB_EcoP-Ro103C3lw	MN067430	39,389	Lytic	Lytic	Virulent	[31]
Escherichia phage Ro45lw	MK301532	39,793	Lytic	Lytic	Virulent	[32]
Escherichia phage vB_EcoS-Ro145clw	MG852086	42,031	Lytic	Lytic	Virulent	[33]
Salmonella phage S4lw	OQ660438	42,250	Lytic	Lytic	Virulent	NA
Escherichia phage vB_EcoS-UDF157lw	OQ243221.1	46,604	Lytic	Lytic	Virulent	NA
Escherichia phage Lys8385Vzw	MT225100	50,953	Lysogenic	Lysogenic	Temperate	[34]
Escherichia phage Lys19259Vzw	MT225101	61,072	Lysogenic	Lysogenic	Temperate	[34]
Escherichia phage Lys12581Vzw	NC_049917	62,668	Lysogenic	Lysogenic	Temperate	[35]
Escherichia phage vB_EcoM-Ro157lw	MH051335	72,179	Lytic	Lytic	Virulent	[36]
Escherichia phage vB_EcoM-Ro111lw	MH571750	86,950	Lytic	Lytic	Virulent	[36]
Escherichia phage vB_EcoM-Pr103Blw	MW481326	88,421	Lytic	Lytic	Virulent	[31]
Escherichia phage vB_EcoM-Pr121LW	MH752840	134,575	Lytic	Lytic	Virulent	[37]
Escherichia phage vB_EcoM-Ro121lw	MH160766	149,803	Lytic	Lytic	Virulent	[36]
Escherichia phage vB_EcoM_Sa157lw	MH427377	155,887	Lytic	Lytic	Virulent	[8]
Salmonella phage D5lw	OQ660437	157,399	Lytic	Lytic	Virulent	NA
Escherichia phage vB_EcoM-S1P5QW	OL956808	166,102	Lytic	Lytic	Virulent	[38]
Escherichia phage vB_EcoM-G157lw	OK331996	167,170	Lytic	Lytic	Virulent	NA
Escherichia phage vB_EcoM-Sa45lw	MK977694	167,353	Lytic	Lytic	Virulent	[39]
Escherichia phage vB_EcoS Sa179lw	MH023293	46,833	Lysogenic	Lytic	Questionable*	[40]

NA: the information is not available

* The phage has lysogenic potential and its lifecycle hasn't been confirmed by biological experiments.

https://doi.org/10.1371/journal.pcbi.1012525.t002

lifecycle classification in metagenomic research. However, next-generation metagenomic sequencing is still at an early stage in investigating viral communities; numerous phage genomes have been detected in the metagenomic sequencing dataset, but biological experiments could not isolate and verify the phage particles. Therefore, a mock phage community metagenomic dataset was selected to accurately compare the performance among different tools instead of the real viral metagenomic dataset. Specifically, this mock phage community, composed of 15 sequenced phages with known lifecycles, was constructed by Cook et al., to perform metagenomic sequencing using the most common next-generation sequencing technologies, including Illumina, Pacbio, and Nanopore sequencing [24]. Therefore, we downloaded 15 complete phage genomes and the metagenomic sequences of this mock phage community generated by various sequencers to evaluate whether our DeepPL could identify the phage lifecycle correctly using metagenomic data. In addition, the PhaTYP was also employed for accuracy comparison due to the similar performance shown above.

First, the 15 complete phage genomes were predicted for lifecycles by DeepPL and PhaTYP (Table 3). The results showed that DeepPL and PhaTYP had 100% and 75% accuracy in predicting these phages with known lifecycles, respectively, indicating a better prediction using DeepPL than PhaTYP with complete phage genomes. The phage sequences were generated from different sequencing platforms (Illumina, Pacbio, and Nanopore) alone or in combination and further assembled by various software (Unicycler, Spades, Fyle, wtdgb2). The resulting phage contigs sharing high genomic similarity with the 15 phage genomes via the nucleotide basic local alignment search tool (blastn) were further used to determine the performance of DeepPL and PhaTYP (Fig 3). For the contigs assembled by Illumina short reads, DeepPL and PhaTYP achieved 84% and 44% accuracy, respectively. Furthermore, DeepPL had high accuracies ranging from 71.14% to 100% for the prediction of contigs generated by Nanopore and Pacbio long reads, while PhaTYP showed varied accuracies of 49.66% - 100%. Among the contigs co-assembled by Illumina short reads and Nanopore/Pacbio long reads, DeepPL performed a relatively higher accuracy (83.33% - 100%) than PhaTYP (55.55% - 100%). The quality of metagenomic sequencing (i.e., genome recovery and error rates) was

Table 3. The phages with known lifecycles used to construct a mock phage community for metagenomic sequencing.

Phages	Accession number	Sequence length (bp)	DeepPL Prediction	PhaTYP prediction	Confirmed lifecycle	References
phiX174	NC_001422	5,386	Lytic	Lytic	Virulent	[41]
HP1	NC_001697	32,355	Lysogenic	Lysogenic	Temperate	[42]
KUW1	OQ376857	44,509	Lytic	Lytic	NA	NA
PARMAL1	OQ376858	44,565	Lytic	Lytic	NA	NA
J1	LR027388	50,343	Lysogenic	Lytic	Temperate	[43]
J2	LR027385	50,343	Lysogenic	Lytic	Temperate	[43]
SWAN	LT841304	50,865	Lysogenic	Lytic	Temperate	[43]
CDMH1	NC_024144	54,279	Lysogenic	Lysogenic	Temperate	[44]
VP1	NA	70,044	Lytic	Lytic	Virulent	[45]
SM032	OV032860	79,660	Lytic	Lytic	Virulent	[45]
J3	LR027389	115,471	Lytic	Lytic	Virulent	[43]
DSS3Mal1	NA	149,582	Lytic	Lytic	NA	NA
PHAGE1	LR027390	167,773	Lytic	Lytic	Virulent	[45]
S-RSM4	FM207411	194,454	Lytic	Lytic	Virulent	[46]
vB_Vpa_sm033	OV032902	320,253	Lytic	Lytic	Virulent	[45]

NA: the information is not available

https://doi.org/10.1371/journal.pcbi.1012525.t003

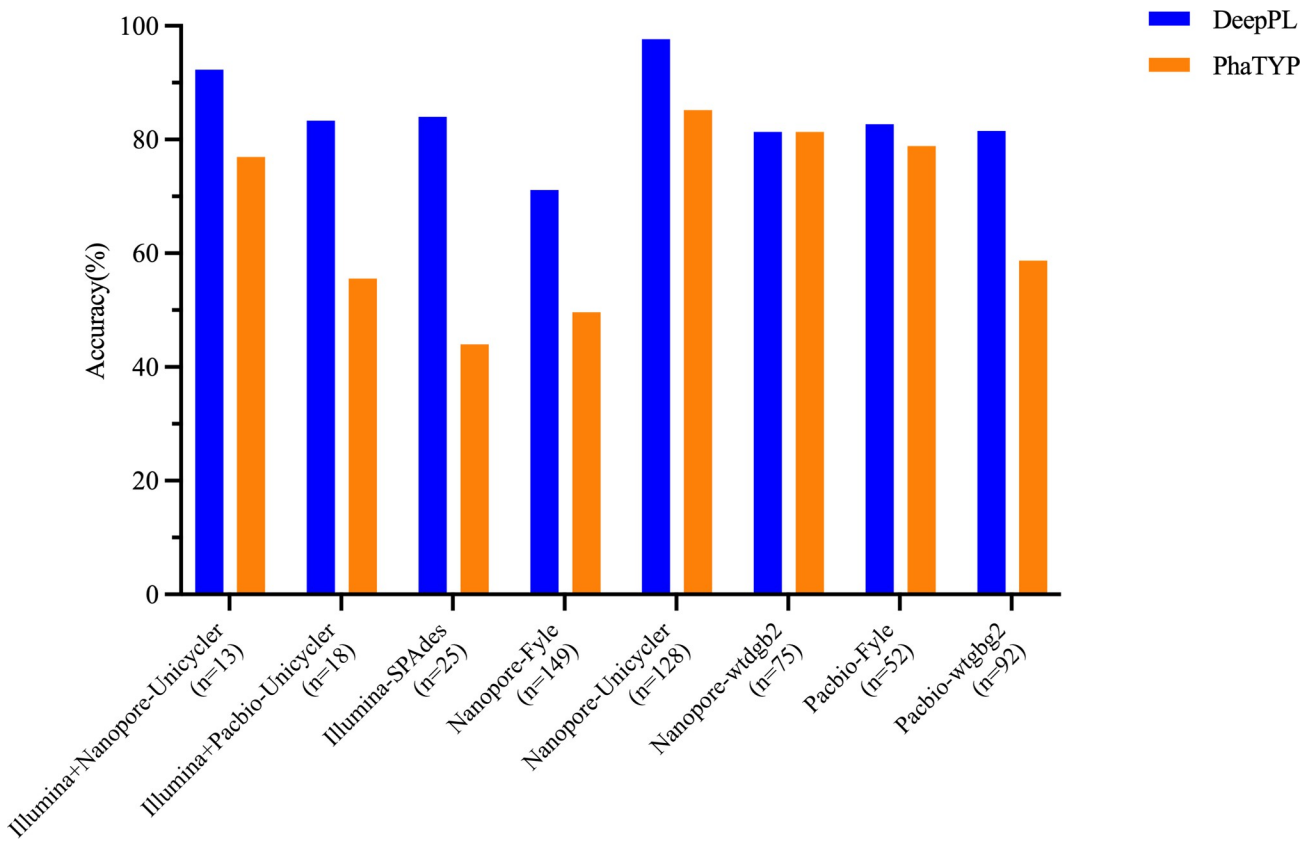

Fig 3. Comparison of DeepPL and PhaTYP performance on phage contigs in a mock phage community generated by different metagenomic sequencing technologies and assemblies.

https://doi.org/10.1371/journal.pcbi.1012525.g003

highly correlated to the sequencer and assembler used in the study and further affected the downstream analysis, such as phage lifecycle prediction. Moreover, the contigs generated by short-read and long-read sequencing approaches were much shorter than complete phage genomes. Even so, DeepPL was able to identify the phage features and predict phage lifecycles with higher accuracy using metagenomic sequences, suggesting its potential as a useful bioinformatic tool in metagenomic research.

Discussion

As the most abundant biological entities in the biosphere, many bacteriophages remain uncharacterized and require further exploration [25]. Diverse phage sequences continue to be discovered through bacteriophage and metagenomic research. But still, a vast majority of available phage sequences create considerable difficulties in navigating through all the data in search of biological meaning, the annotation of phage proteins in particular. Currently, no standard workflow is available for full annotation of a draft phage genome. Due to the imperfect and inconsistent annotation of phage genomes, it is hard to achieve precise characterization based on the phage amino acid sequences. Therefore, nucleotide sequences (DNA sequences) are the most fundamental genomic language that can represent basic biological information accurately. In this study, we utilized the comprehensive language learning ability, DNABERT, to characterize phage lifecycle by feeding the fundamental DNA sequences. It could reduce the noise from protein prediction or rough phage genome annotation from

different pipelines to capture the underlying semantic information of DNA language for phage lifecycle identification.

The phage lifecycle classification requires a complicated validation process through traditional biological experiments, including the spot test, plaque assay, and lysogen test. A precise prediction of the phage lifecycle will improve biological validation tests by saving time and effort. In this study, the precision of DeepPL was also confirmed by the biologically-characterized phages in case study 1. A total of 19 different phages were isolated from agricultural environments by our lab and subjected to verify the phage lifecycles by essential biological experiments. The DeepPL analysis showed the prediction is 100% consistent with our biological results. To our knowledge, this is the first study using actual phages characterized by biological experiments to verify the output of prediction models.

Solid lytic phages can only perform the lytic cycle, whereas lysogenic phages can display either a lytic or lysogenic cycle based on environmental conditions. Therefore, virulent and temperate phages both contained many lysis modules. The primary genomic difference between them relies on the presence of limited lysogenic genes. Therefore, only a few DNA segments within the phage genome can be detected as lysogenic modules. It further explained the reasonable setting of the low threshold 2 that the score of phage classification more than 0.016 was identified as lysogenic phage in this study. In addition, threshold 2 of 0.016 contributes to the high sensitivity of lysogenic gene detection using our model. This capability could benefit phage applications in the clinical and agricultural areas. Specifically, lytic phages used for phage-based therapy or biocontrol have a strict requirement: no virulence, antibiotic resistance, or lysogenic genes can be detected in the phage genome. In nature, some lysogenic phages lose their lysogenic genes during the phage production process and perform the lytic cycle. In some cases, lytic phages could acquire one or a few lysogenic genes by gene exchange or recombination but still show strict lytic infection. Due to the risk of lysogenic potential, these phages will not be considered for phage therapy or biocontrol purposes. For example, one of our lab phages presented in this study, Sa179lw, was isolated from surface water, showing antimicrobial activity against *E. coli* O179 strains. However, this phage was withdrawn from the lytic phage cocktail development for biocontrol application because the lysogenic gene *cro* was present in the genome and also identified by DeepPL. The findings indicate the excellent screening capability of DeepPL on the phenotypic camouflage phages.

Viral metagenomic sequencing has become a popular technique to determine the viral population within different samples, including gut and environmental viromes [26,27]. Bacteriophages, the major component of the virome, drive the diversities and evolution of viral communities and host bacterial populations. Due to the specific interaction between phages and bacteria, there has been a growing interest in lytic and lysogenic phage profiles within the total phageome in recent years. Therefore, the potential application of DeepPL in the metagenomic study was tested in case study 2. DeepPL performed well using short-reads and long-reads-based metagenomic sequences longer than 5,000 bp with accuracy ranging from 82.66% to 100%. Even for sequences less than 5,000 bp, DeepPL has a relatively high accuracy of 71.14% - 97.65%, depending on the sequencer and assembly tools (S3 Data). The results from this study indicated that the accuracy was closely associated with several factors, including the contig lengths, sequencing technologies, and genome assembly tools. Notably, the short contigs require a comprehensive capability to identify genomic features of lysogenic genes for classification. As a result, PhaTYP's prediction accuracy decreased in predicting contigs shorter than 5,000 bp, likely due to limited amino acid sequences by genomic annotation. In contrast, with a relatively high accuracy, the DeepPL model can efficiently identify the lysogenic gene feature based on the nucleotide sequences.

DeepPL has a great performance on phage lifecycle classification compared to currently available tools. Even so, some limitations still need to be improved in future studies. The precision of DeepPL trades off the rapid processing time. The processing time of DeepPL depends on the length of phage genomes and the processor. For example, the CPU (M1 chip) can predict a phage genome with a length of 42,031 bp in 17 mins, while the GPU (NVIDIA V100 GPU, 1 node, memory of 64 GB) can complete the prediction within 2 mins. In addition, a user-friendly web server version of DeepPL will be established for users unfamiliar with coding languages. Furthermore, the diversity and plasticity of phage genomes are highly associated with the model performance. The predictive ability of phylogeny on diverse microbial populations has been shown in several studies [28–30]. Therefore, the phylogenetic-based sampling method, with more newly sequenced phage genomes, should be included in the training dataset to increase the coverage of DeepPL.

Supporting information

S1 Data. Detailed phage information of training and test datasets.
(XLSX)

S2 Data. The performance of DeepPL using 5-fold cross-validation.
(XLSX)

S3 Data. Detailed prediction results of DeepPL and PhatTYP for case-study2.
(XLSX)

Author Contributions

Conceptualization: Vivian C. H. Wu.

Data curation: Yujie Zhang.

Funding acquisition: Vivian C. H. Wu.

Investigation: Yujie Zhang.

Methodology: Yujie Zhang, Mark Mao, Robert Zhang.

Software: Mark Mao, Robert Zhang.

Supervision: Vivian C. H. Wu.

Writing – original draft: Yujie Zhang.

Writing – review & editing: Yujie Zhang, Mark Mao, Robert Zhang, Yen-Te Liao, Vivian C. H. Wu.

References

1. Mushegian AR. Are there 10^{31}virus particles on earth, or more, or fewer? Journal of Bacteriology 2020; 202: 10–1128.
2. Srinivasiah S, Bhavsar J, Thapar K, et al. Phages across the biosphere: contrasts of viruses in soil and aquatic environments. Research in Microbiology 2008; 159:349–357. https://doi.org/10.1016/j.resmic.2008.04.010 PMID: 18565737
3. Callanan J, Stockdale SR, Shkoporov A, et al. RNA Phage Biology in a Metagenomic Era. Viruses 2018, Vol. 10, Page 386 2018; 10:386. https://doi.org/10.3390/v10070386 PMID: 30037084
4. Fitzgerald CB, Shkoporov AN, Upadrasta A, et al. Probing the "Dark Matter" of the Human Gut Phageome: Culture Assisted Metagenomics Enables Rapid Discovery and Host-Linking for Novel Bacteriophages. Frontiers in Cellular and Infection Microbiology 2021; 11: 616918. https://doi.org/10.3389/fcimb.2021.616918 PMID: 33791236

5. Gummalla VS, Zhang Y, Liao Y-T, et al. The Role of Temperate Phages in Bacterial Pathogenicity. Microorganisms 2023; 11:541. https://doi.org/10.3390/microorganisms11030541 PMID: 36985115
6. De Paepe M, Leclerc M, Tinsley CR, et al. Bacteriophages: An underestimated role in human and animal health? Frontiers in Cellular and Infection Microbiology 2014; 4:39. https://doi.org/10.3389/fcimb.2014.00039 PMID: 24734220
7. Schooley RT, Biswas B, Gill JJ, et al. Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter baumannii Infection. Antimicrobial Agents and Chemotherapy 2017; 61:e00954–17. https://doi.org/10.1128/AAC.00954-17 PMID: 28807909
8. Liao Y-T, Zhang Y, Salvador A, et al. Characterization of polyvalent Escherichia phage Sa157lw for the biocontrol potential of *Salmonella* Typhimurium and *Escherichia coli* O157:H7 on contaminated mung bean seeds. Front Microbiol 2022; 13:1053583. https://doi.org/10.3389/fmicb.2022.1053583 PMID: 36439834
9. Azzazy HME, Highsmith WE. Phage display technology: clinical applications and recent innovations. Clinical Biochemistry 2002; 35:425–445. https://doi.org/10.1016/s0009-9120(02)00343-0 PMID: 12413604
10. Schofield D, Sharp NJ, Westwater C. Phage-based platforms for the clinical detection of human bacterial pathogens. Bacteriophage 2012; 2:105–121. https://doi.org/10.4161/bact.19274 PMID: 23050221
11. Groth AC, Calos MP. Phage Integrases: Biology and Applications. Journal of Molecular Biology 2004; 335:667–678. PMID: 14687564
12. Mimmi S, Maisano D, Quinto I, et al. Phage Display: An Overview in Context to Drug Discovery. Trends in Pharmacological Sciences 2019; 40:87–91. PMID: 30606501
13. Levine M. Mutations in the temperate phage P22 and lysogeny in *Salmonella*. Virology 1957; 3:22–41. PMID: 13409758
14. Jofre J, Muniesa M. Bacteriophage Isolation and Characterization: Phages of *Escherichia coli*. Horizontal Gene Transfer: Methods and Protocols 2020; 61–79. https://doi.org/10.1007/978-1-4939-9877-7_4 PMID: 31584154
15. Altamirano FLG, Barr JJ. Screening for lysogen activity in therapeutically relevant bacteriophages. Bio-protocol 2021; 11:e3997. https://doi.org/10.21769/BioProtoc.3997 PMID: 34124298
16. McNair K, Bailey BA, Edwards RA. PHACTS, a computational approach to classifying the lifestyle of phages. Bioinformatics 2012; 28:614–618. https://doi.org/10.1093/bioinformatics/bts014 PMID: 22238260
17. Tynecki P, Guziński A, Kazimierczak J, et al. PhageAI—Bacteriophage Life Cycle Recognition with Machine Learning and Natural Language Processing. 2020; 2020.07.11.198606. https://doi.org/10.1101/2020.07.11.198606
18. S W, Z F, J T, et al. DeePhage: distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. GigaScience 2021; 10:giab056. https://doi.org/10.1093/gigascience/giab056 PMID: 34498685
19. Shang J, Tang X, Sun Y. PhaTYP: predicting the lifestyle for bacteriophages using BERT. Briefings in Bioinformatics 2023; 24:bbac487. https://doi.org/10.1093/bib/bbac487 PMID: 36659812
20. Gauthier J, Vincent AT, Charette SJ, et al. A brief history of bioinformatics. Briefings in Bioinformatics 2019; 20:1981–1996. https://doi.org/10.1093/bib/bby063 PMID: 30084940
21. Calladine CR, Drew H. Understanding DNA: The Molecule and How It Works. Academic press; 1997.
22. Ji Y, Zhou Z, Liu H, et al. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics 2021; 37:2112–2120. https://doi.org/10.1093/bioinformatics/btab083 PMID: 33538820
23. Azadpour M, McKay CM, Smith RL. Estimating confidence intervals for information transfer analysis of confusion matrices. The Journal of the Acoustical Society of America 2014; 135:EL140–EL146. https://doi.org/10.1121/1.4865840 PMID: 24606307
24. Cook R, Brown N, Rihtman B, et al. The long and short of it: Benchmarking viromics using Illumina, Nanopore and PacBio sequencing technologies. Microbial Genomics 2024, 10.2:001198. https://doi.org/10.1099/mgen.0.001198 PMID: 38376377
25. Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol 2020; 18:125–138. https://doi.org/10.1038/s41579-019-0311-5 PMID: 32015529
26. Liang Y, Wang L, Wang Z, et al. Metagenomic Analysis of the Diversity of DNA Viruses in the Surface and Deep Sea of the South China Sea. Frontiers in Microbiology 2019; 10:1951. https://doi.org/10.3389/fmicb.2019.01951 PMID: 31507563

27. Manrique P, Bolduc B, Walk ST, et al. Healthy human gut phageome. Proceedings of the National Academy of Sciences 2016; 113:10400–10405. https://doi.org/10.1073/pnas.1601060113 PMID: 27573828

28. Walkup J, Dang C, Mau RL, et al. The predictive power of phylogeny on growth rates in soil bacterial communities. ISME COMMUN. 2023; 3:1–8. https://doi.org/10.1038/s43705-023-00281-1 PMID: 37454187

29. Zhou B, Zhou H, Zhang X, et al. TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution. Computers in Biology and Medicine 2023; 152:106264. https://doi.org/10.1016/j.compbiomed.2022.106264 PMID: 36535209

30. Hong Q, Chen G, Tang Z-Z. PhyloMed: a phylogeny-based test of mediation effect in microbiome. Genome Biology 2023; 24:72. https://doi.org/10.1186/s13059-023-02902-3 PMID: 37041566

31. Zhang Y, Liao Y-T, Salvador A, et al. Characterization of Two New Shiga Toxin-Producing *Escherichia coli* O103-Infecting Phages Isolated from an Organic Farm. Microorganisms 2021; 9:1527. https://doi.org/10.3390/microorganisms9071527 PMID: 34361962

32. Sun X, Liao Y-T, Zhang Y, et al. A New Kayfunavirus-like Escherichia Phage vB_EcoP-Ro45lw with Antimicrobial Potential of Shiga Toxin-Producing *Escherichia coli* O45 Strain. Microorganisms 2023; 11:77.

33. Liao Y-T, Salvador A, Harden LA, et al. Characterization of a Lytic Bacteriophage as an Antimicrobial Agent for Biocontrol of Shiga Toxin-Producing *Escherichia coli* O145 Strains. Antibiotics 2019; 8:74. https://doi.org/10.3390/antibiotics8020074 PMID: 31195679

34. Zhang Y, Liao Y-T, Salvador A, et al. Genomic Characterization of Two Shiga Toxin–Converting Bacteriophages Induced From Environmental Shiga Toxin–Producing *Escherichia coli*. Frontiers in Microbiology 2021; 12: 587696. https://doi.org/10.3389/fmicb.2021.587696 PMID: 33716997

35. Zhang Y, Liao Y-T, Salvador A, et al. Complete Genome Sequence of a Shiga Toxin-Converting Bacteriophage, Escherichia Phage Lys12581Vzw, Induced from an Outbreak Shiga Toxin-Producing *Escherichia coli* Strain. Microbiology Resource Announcements 2019; 8:e00793–19. https://doi.org/10.1128/MRA.00793-19 PMID: 31488531

36. Liao Y-T, Sun X, Quintela IA, et al. Discovery of Shiga Toxin-Producing *Escherichia coli* (STEC)-Specific Bacteriophages From Non-fecal Composts Using Genomic Characterization. Frontiers in Microbiology 2019; 10:627. https://doi.org/10.3389/fmicb.2019.00627 PMID: 31001216

37. Liao Y-T, Liu F, Wu VCH. Complete Genome Sequence of Escherichia Phage vB_EcoM-Pr121LW, Isolated from Soil in an Organic Farm. Microbiology Resource Announcements 2018; 7:e01236–18. https://doi.org/10.1128/MRA.01236-18 PMID: 30533815

38. Quintela IA, Hwang A, Vasse T, et al. Whole-Genome Analysis of Escherichia Phage vB_EcoM-S1P5QW, Isolated from Manures Collected from Cattle Farms in Maine. Microbiology Resource Announcements 2022; 11:e00041–22. https://doi.org/10.1128/mra.00041-22 PMID: 35254108

39. Liao Y-T, Zhang Y, Salvador A, et al. Characterization of a T4-like Bacteriophage vB_EcoM-Sa45lw as a Potential Biocontrol Agent for Shiga Toxin-Producing *Escherichia coli* O45 Contaminated on Mung Bean Seeds. Microbiology Spectrum 2022; 10:e02220–21. https://doi.org/10.1128/spectrum.02220-21 PMID: 35107386

40. Liao Y-T, Liu F, Sun X, et al. Complete Genome Sequence of *Escherichia coli* Phage vB_EcoS Sa179lw, Isolated from Surface Water in a Produce-Growing Area in Northern California. Genome Announc 2018; 6:e00337–18. https://doi.org/10.1128/genomeA.00337-18 PMID: 29976601

41. Jazwinski SM, Lindberg AA, Kornberg A. The lipopolysaccharide receptor for bacteriophages ∅X174 and S13. Virology 1975; 66:268–282.

42. Waldman AS, Goodman SD, Scocca JJ. Nucleotide sequences and properties of the sites involved in lysogenic insertion of the bacteriophage HP1c1 genome into the Haemophilus influenzae chromosome. J Bacteriol 1987; 169:238–246. https://doi.org/10.1128/jb.169.1.238-246.1987 PMID: 3491821

43. Michniewski S, Redgwell T, Grigonyte A, et al. Riding the wave of genomics to investigate aquatic coliphage diversity and activity. Environmental Microbiology 2019; 21:2112–2128. https://doi.org/10.1111/1462-2920.14590 PMID: 30884081

44. Hargreaves KR, Kropinski AM, Clokie MRJ. What Does the Talking?: Quorum Sensing Signalling Genes Discovered in a Bacteriophage Genome. PLOS ONE 2014; 9:e85131. https://doi.org/10.1371/journal.pone.0085131 PMID: 24475037

45. Michniewski S. Phages infecting marine *Vibrios*: prevalence, diversity and role in the dissemination of antibiotic resistance genes. Thesis, University of Warwick, 2020.

46. Millard AD, Zwirglmaier K, Downey MJ, et al. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: implications for mechanisms of cyanophage evolution. Environmental Microbiology 2009; 11:2370–2387. https://doi.org/10.1111/j.1462-2920.2009.01966.x PMID: 19508343