frontiers | Frontiers in Big Data

TYPE Data Report
PUBLISHED 27 November 2024
DOI 10.3389/fdata.2024.1437580

OPEN ACCESS

EDITED BY
Hariharan Shanmugasundaram,
Vardhaman College of Engineering, India

REVIEWED BY
Yongsheng Hao,
Nanjing University of Information Science and Technology, China
Viyyapu Lokeshwari Vinya,
Vardhaman College of Engineering, India

*CORRESPONDENCE
Opeyemi U. Lawal
✉ lawal@uoguelph.ca
Lawrence Goodridge
✉ goodridl@uoguelph.ca

RECEIVED 24 May 2024
ACCEPTED 12 November 2024
PUBLISHED 27 November 2024

CITATION
Lawal OU and Goodridge L (2024) TSPDB: a curated resource of tailspike proteins with potential applications in phage research. *Front. Big Data* 7:1437580.
doi: 10.3389/fdata.2024.1437580

TSPDB: a curated resource of tailspike proteins with potential applications in phage research

Opeyemi U. Lawal* and Lawrence Goodridge*

Canadian Research Institute for Food Safety (CRIFS), Department of Food Science, University of Guelph, Guelph, ON, Canada

KEYWORDS

phage, tailspike proteins, genomics, big data, data mining

Background

Bacteriophages (phages) are viruses that infect and replicate within host bacteria and archaea (Chatterjee and Duerkop, 2018; Dion et al., 2020). Phages are the most abundant entities in the biosphere (Dion et al., 2020) and are distributed across different biomes populated by bacterial and archaeal hosts, including the gastrointestinal tract of humans and animals, and oceanic beds (Chevallereau et al., 2022; Clokie et al., 2011). They play a vital role in the rapid evolution and adaptation of their hosts in various environments (Dion et al., 2020).

Phages exhibit high genomic, morphological, and structural diversity, composed of DNA or RNA that can be single-stranded or double-stranded and packaged into a capsid (Dion et al., 2020; Fokine and Rossmann, 2014). The structural form of the capsid was a major feature used in the taxonomic classification of phages until the advent of whole-genome sequencing, which has now become the gold standard for this classification (Dion et al., 2020; Fokine and Rossmann, 2014; Turner et al., 2023). Phages are broadly classified as tailed or non-tailed, with double-stranded DNA tailed phages constituting about 96% of all known phages (Dion et al., 2020). Phages possess a diverse array of tail structures essential for host recognition, attachment, and penetration, making them important targets in phage therapy research (Fokine and Rossmann, 2014; Gil et al., 2023). Phage infection of its host begins with the recognition of a receptor on the bacterial cell surface for attachment (Dowah and Clokie, 2018; Latka et al., 2017). To penetrate the host cell, phages must overcome various complex barriers on the bacterial cell wall, such as the outer membrane of Gram-negative bacteria and the lipoteichoic acids of Gram-positive bacteria (Chen et al., 2014; Latka et al., 2017). Phages encode virion-associated carbohydrate-degrading enzymes called depolymerases, which are distinct from the endolysins produced by phages during the lysis stage (Knecht et al., 2020; Yan et al., 2014). These depolymerases, encoded by tailspike protein (TSP) genes, recognize, bind, and degrade cell-surface associated polysaccharides, unmasking phage receptors and making them accessible for bacterial infection (Gil et al., 2023; Greenfield et al., 2019; Latka et al., 2017).

Tailspike proteins are integral components of phage tail structures, and their activities as polysaccharide depolymerases are related to host specificity and infectivity (Greenfield et al., 2019). A hallmark of TSPs is their host specificity, high thermostability, resistance to protease treatment, and stability in the presence of high concentrations of urea and sodium dodecyl sulfate (Chen et al., 2014). Phage TSPs possess carbohydrate depolymerase activity and recognize capsule, and lipopolysaccharides (LPS) where they cleave components of the LPS to position the phage toward a secondary membrane receptor during infection (Knecht et al., 2020). TSPs have been observed to decrease bacterial viability, leading to antimicrobial applications. For example, Ayariga et al. (2021) demonstrated that the ε34 phage tailspike protein has enzymatic property as a LPS hydrolase and synergizes with Vero Cell culture supernatant in killing *Salmonella* Newington. The ε34 TSP also showed bactericidal efficacy against different *Salmonella* serovars in various matrices (Ibrahim et al., 2023). Miletic and colleagues (Miletic et al., 2016) expressed the receptor binding domain of the Phage P22 Gp9 tailspike protein in plant tissue (*Nicotiana benthamiana*), and demonstrated that, upon oral administration of lyophilized leaves expressing Gp9 TSP to newly hatched chickens, *Salmonella* concentrations were reduced on average by approximately 0.75 log relative to controls. Others have shown that TSPs can be used to control the growth of plant pathogens. For example, expression of the *Erwinia* spp. phage TSP DpoEa1h in transgenic apple and pear plants significantly reduced fire blight (*Erwinia amylovora*) susceptibility (Malnoy et al., 2005; Roach and Donovan, 2015), likely due to removal of the main virulence factor amylovoran and exposing the *E. amylovora* cells to host plant defenses (Kim et al., 2004). Finally, phage LKA1 TSP exhibits disruptive activity against biofilms while also reducing virulence in *Pseudomonas* in an infection model (Olszak et al., 2017). Collectively, these studies demonstrate the utility of TSPs as novel antimicrobials to control the growth of food and plant-borne pathogens in foods.

Despite the known antimicrobial applications of TSPs, only a few have been fully characterized to date. This could be partly due to the laborious nature of detection techniques, which include plaque assays followed by examination under a transmission electron microscope (TEM) to identify "bulb-like" baseplate structures at the base of phage tails indicative of TSPs (Bhandare et al., 2024; Knecht et al., 2020). The decreasing costs of sequencing and the availability of improved bioinformatics tools have facilitated the construction of large-scale genome and metagenome datasets (Emond-Rheault et al., 2017; Wattam et al., 2014). High-throughput *in silico* detection of TSP-encoding genes in genomic data would not only provide further details regarding the diversity of TSPs in virulent phages but could also be used to identify TSPs in prophages. In this report, we present a high-level curated resource called TSP database (TSPDB) for the rapid detection of tailspike proteins in multiomics sequence data. This TSPDB will be an indispensable resource for researchers in phage biology, drug discovery, and antimicrobial resistance domains to further contribute to the understanding of the structure and function of these proteins to harness their potential for diverse applications, such as the development of phage therapy for bacterial infections or phage-based biocontrol of foodborne pathogens, and drug discovery (Brives and Pourraz, 2020; Roach and Donovan, 2015).

Data and methodology

Data mining and quality check

The DDBJ/ENA/GenBank and UniProt databases (Sayers et al., 2022; The UniProt Consortium et al., 2023) were queried for TSPs using search terms commonly associated with tailspike proteins, such as "phage tailspike," "tail spike proteins," "phage endopeptidase," and "phage endorhamnosidase." (Figure 1). Hits were systematically filtered based on annotation criterion to exclude duplicate results. Nucleotide sequences of TSPs were retrieved from public databases using accession numbers obtained from the database query via NCBI Entrez Programming Utilities (E-utilities) (National Center for Biotechnology Information, 2023).

Dataset curation

From this exercise, 17,211 sequences were obtained from the queried public databases. Duplicated sequences were removed using thresholds of ≥95% sequence coverage and nucleotide similarity with cd-hit (Li and Godzik, 2006) and Seqkit (Shen et al., 2016), resulting in 9,129 unique TSP sequences. To assess the sequence length distribution and perform quality checks on unique TSP sequences, Gaussian distribution analysis was conducted. Sequences shorter than 400 bp, which could represent partial region or incomplete sequences that may lack critical functional domains required for accurate annotation and functional prediction, were excluded from the dataset. By excluding these shorter sequences, we reduce the possibility of including fragments that could introduce noise or inaccuracies into the database. This threshold helps ensure that the TSPDB contains more reliable and complete sequences for functional analysis and annotation. This filtering process resulted in a total of 8,105 unique TSP sequences (Figure 1). TSP sequences with a length of ≤10,000 bp were retained to include those originating from Gram-positive bacteria such as *Clostridium* and *Streptococcus*, among others. Overall size range of TSPs retrieved from the public databases is 405 to 9,990 bp (Figure 2A). Further analysis of TSP genes in the TSPDB reveals a significant difference ($p < 0.001$) in the sizes of TSPs between Gram-negative and Gram-positive bacteria. Specifically, the average size of TSPs for Gram-negative bacteria is 2,070 bp, while the average size for Gram-positive bacteria is substantially larger, at 3,255 bp (Figure 2B). The TSPDB contains TSPs from more than 400 bacterial genera. Among these, the top 13 genera represented were Gram-positive bacteria, with TSPs from *Bacillus* ($n = 1{,}616$) being the most common, followed by *Streptococcus* ($n = 1{,}152$), *Clostridium* ($n = 683$), *Enterococcus* ($n = 387$), and *Staphylococcus* ($n = 372$). Additionally, TSPs from Gram-negative bacterial genera, *Salmonella* ($n = 80$), *Escherichia* ($n = 58$), *Klebsiella* ($n = 52$), and *Pseudomonas* ($n = 25$) were among the top 38 TSPs in the database (Figure 2C). To assess the normality of the distribution

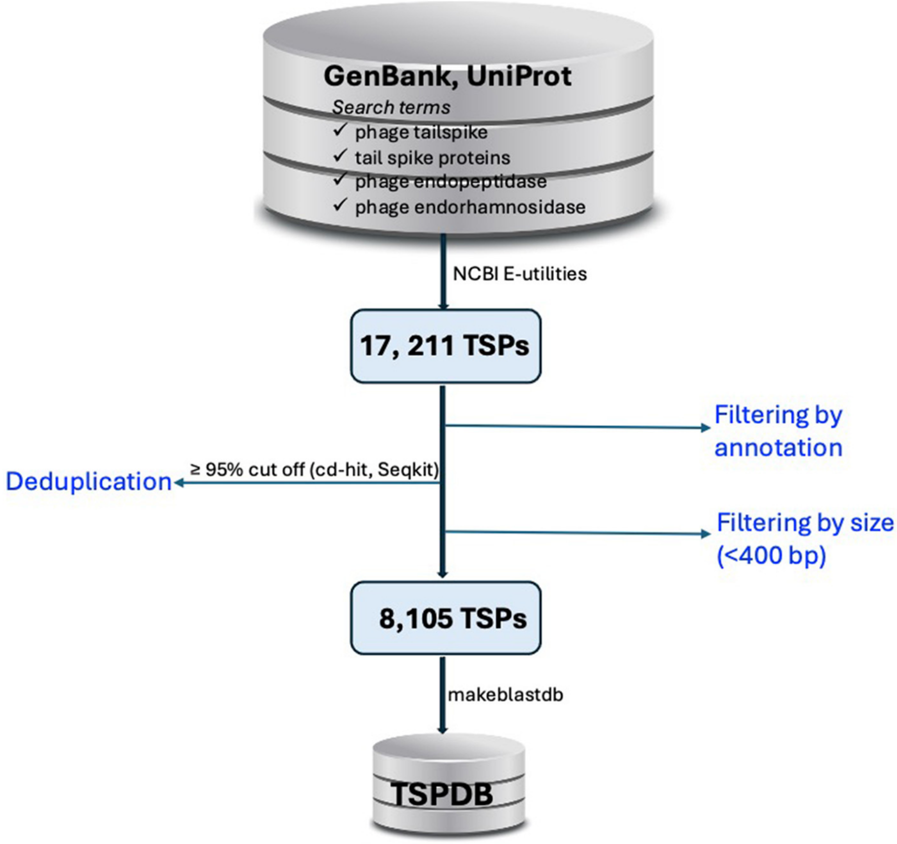

FIGURE 1
Workflow for the construction of the tailspike protein database (TSPDB). TSP sequences were retrieved from GenBank and UniProt using specific search terms, yielding 17,211 sequences. After filtering by annotation and excluding sequences <400 bp and >10,000 bp, deduplication at ≥95% similarity reduced the dataset to 8,105 unique sequences, which were then compiled into the TSPDB for efficient access.

of TSP frequencies across bacterial genera, we performed a Shapiro-Wilk test. This test yielded a statistic of 0.487 and a p-value < 0.0001, confirming a significant departure from normality. This result supports the observation of a skewed distribution, where Gram positive bacteria host genera (e.g., *Bacillus* and *Streptococcus*) exhibit notably high TSP counts compared to others.

Diversity of TSPs

To assess the diversity of the 8,105 unduplicated TSP sequences and their suitability for database creation, we employed a phylogeny-based approach. The TSP sequences were aligned using MAFFT v7.453 (Katoh, 2002), and a maximum likelihood tree was constructed with FastTree v2.1.11 (Price et al., 2010) using the generalized time reversible mode and 1,000 bootstrap replicates for node support. The resulting phylogenetic tree was visualized using the web-based Microreact visualization tool (Argimón et al., 2016) (Figure 2D). The phylogeny revealed the high diversity of TSPs in the TSPDB, further supporting the uniqueness of individual TSPs. TSPs from the same species often belonged to different clusters. For example, TSPs from *Bacillus* and *Listeria* were distributed across multiple clusters in the phylogeny. While the majority of TSPs from *Salmonella* belonged to the same cluster, there were also a few instances of TSPs from this host genus in separate clusters (Figure 2D).

TSPDB construction

The deduplicated TSP nucleotide sequences were utilized to construct the TSP database using makeblastdb (Camacho et al., 2009). This database is compatible for use with ABRicate (https://github.com/tseemann/abricate) and other bioinformatics tools equipped with embedded BLAST algorithms, such as BLAST suites and SRST2 (Inouye et al., 2014), among others.

TSPDB application

The TSPDB was recently utilized in a study by Bhandare et al. (2024), where the database was implemented within an ABRicate container. The database index files suitable for use with blast was generated using makeblast_db option in

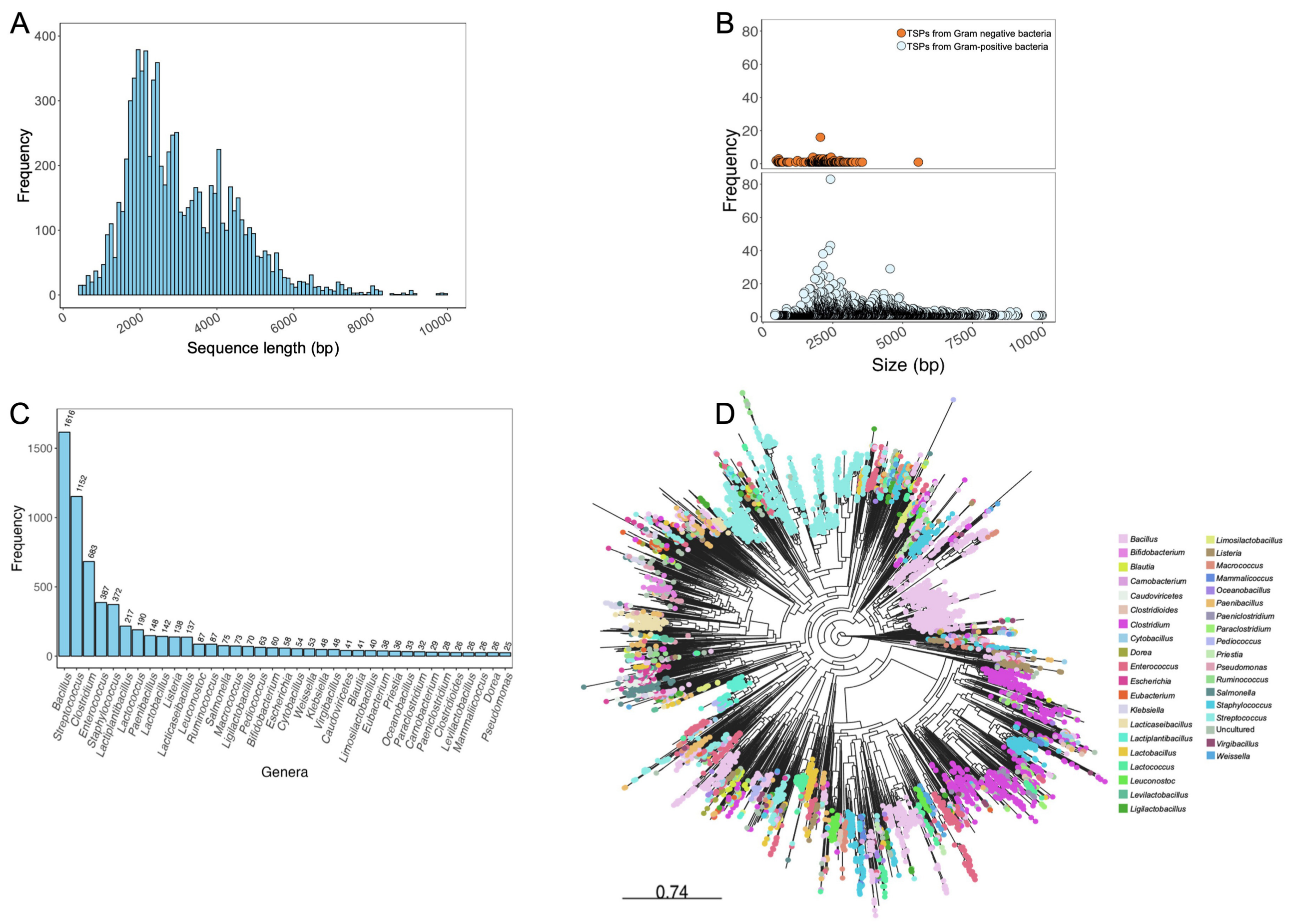

FIGURE 2
Analysis of phage tail spike proteins in the TSPDB. **(A)** Sequence length distribution of genes encoding phage TSPs contained in the TSPDB. TSP size in the database ranged from 405–9,990 bp. **(B)** Differential frequency distribution of TSP gene sizes in Gram-negative (orange circles) and Gram-positive (blue circles) bacteria in the TSPDB. **(C)** Frequency across the top 37 genera of host phages carrying TSPs in the TSPDB. **(D)** Phylogenetic diversity of the 8,105 TSPs in the TSPDB. Each node represents a unique TSP contained in the TSPDB, with nodes of similar color belonging to the same genera. The top 37 genera are displayed in color. An interactive version of this figure is accessible through the following link - https://microreact.org/project/7Kv61nb6aRapgGgHpxsNGL-tspdb-v20. To assess the normality of the distribution of TSP frequencies across bacterial genera, we performed a Shapiro-Wilk test. In this analysis, the Shapiro-Wilk test yielded a statistic of 0.487 and a *p*-value < 0.0001, confirming a significant departure from normality. This result supports the observation of a skewed distribution, where a small number of genera (e.g., *Bacillus* and *Streptococcus*) exhibit notably high TSP counts compared to others.

ABRicate. The step-by-step guide on how to incorporate TSPDB into ABRicate for rapid screening of large genomic dataset is detailed on the ABRicate Github page (https://github.com/tseemann/abricate). The presence of TSPs in a collection of phage genomes were determined using stringent parameters (≥90% identity and ≥70% coverage). TSPDB provides valuable applications across various fields, particularly in phage therapy, biocontrol, and functional genomics and would contribute to advancing the application of TSPs in biocontrol strategies in agriculture and food safety. Overall, the TSPDB contains a vast dataset of diverse TSPs found in phages, and the integration of this database into phage detection tools will enhance the functional annotation of these genes in large genomic and metagenomic datasets. Lastly, the TSPDB described here will undergo regular updates and expansion to include new TSPs as they become available in public databases ensuring that the database remains comprehensive.

Limitations

It is acknowledged that mis-annotation of some TSPs as hypothetical proteins or tail fibers in public databases may have resulted in the omission of certain TSP genes in this study. However, the TSPDB will be continually updated to incorporate additional TSP genes.

Dataset description

The TSPDB is freely accessible on GitHub at the following link: https://github.com/yemilawal/Tailspike-proteins or by searching for the title "TSPDB: A curated resource of tailspike proteins with potential applications in phage research" on GitHub. Additionally, accession numbers of genes encoding phage tailspike proteins in TSPDB are available on the GitHub page. A backup version is also available for download on Figshare at https://doi.org/10.6084/m9.figshare.25526323.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

OL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing. LG: Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Canada First Research Excellence Fund, the Canada Research Chair Program, and a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to LG.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., et al. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microbial Genomics* 2, 1–11. doi: 10.1099/mgen.0.000093

Ayariga, J. A., Gildea, L., Wu, H., and Villafane, R. (2021). The ε34 phage tailspike protein: an in vitro characterization, structure prediction, potential interaction with S. newington lps and cytotoxicity assessment to animal cell line. *J. Clin. Trials* 11, 1–18. doi: 10.1101/2021.09.20.461090

Bhandare, S., Lawal, O. U., Colavecchio, A., Cadieux, B., Zahirovich-Jovich, Y., Zhong, Z., et al. (2024). Genomic and phenotypic analysis of salmonella enterica bacteriophages identifies two novel phage species. *Microorganisms* 12, 1–17. doi: 10.3390/microorganisms12040695

Brives, C., and Pourraz, J. (2020). Phage therapy as a potential solution in the fight against AMR: obstacles and possible futures. *Palgrave Commun.* 6:100. doi: 10.1057/s41599-020-0478-4

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinform.* 10:421. doi: 10.1186/1471-2105-10-421

Chatterjee, A., and Duerkop, B. A. (2018). Beyond bacteria: bacteriophage-eukaryotic host interactions reveal emerging paradigms of health and disease. *Front. Microbiol.* 9:1394. doi: 10.3389/fmicb.2018.01394

Chen, C., Bales, P., Greenfield, J., Heselpoth, R. D., Nelson, D. C., and Herzberg, O. (2014). Crystal structure of ORF210 from *E. coli* O157:H1 phage CBA120 (TSP1), a putative tailspike protein. *PLoS ONE* 9:e93156. doi: 10.1371/journal.pone.0093156

Chevallereau, A., Pons, B. J., van Houte, S., and Westra, E. R. (2022). Interactions between bacterial and phage communities in natural environments. *Nat. Rev. Microbiol.* 20, 49–62. doi: 10.1038/s41579-021-00602-y

Clokie, M. R. J., Millard, A. D., Letarov, A. V., and Heaphy, S. (2011). Phages in nature. *Bacteriophage* 1, 31–45. doi: 10.4161/bact.1.1.14942

Dion, M. B., Oechslin, F., and Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.* 18, 125–138. doi: 10.1038/s41579-019-0311-5

Dowah, A. S. A., and Clokie, M. R. J. (2018). Review of the nature, diversity and structure of bacteriophage receptor binding proteins that target Gram-positive bacteria. *Biophys. Rev.* 10, 535–542. doi: 10.1007/s12551-017-0382-3

Emond-Rheault, J.-G., Jeukens, J., Freschi, L., Kukavica-Ibrulj, I., Boyle, B., Dupont, M.-J., et al. (2017). A Syst-OMICS approach to ensuring food safety and reducing the economic burden of salmonellosis. *Front. Microbiol.* 8:996. doi: 10.3389/fmicb.2017.00996

Fokine, A., and Rossmann, M. G. (2014). Molecular architecture of tailed double-stranded DNA phages. *Bacteriophage* 4:e28281. doi: 10.4161/bact.28281

Gil, J., Paulson, J., Brown, M., Zahn, H., Nguyen, M. M., Eisenberg, M., et al. (2023). Tailoring the host range of ackermannviridae bacteriophages through chimeric tailspike proteins. *Viruses* 15:286. doi: 10.3390/v15020286

Greenfield, J., Shang, X., Luo, H., Zhou, Y., Heselpoth, R. D., Nelson, D. C., et al. (2019). Structure and tailspike glycosidase machinery of ORF212 from *E. coli* O157:H7 phage CBA120 (TSP3). *Sci. Rep.* 9:7349. doi: 10.1038/s41598-019-43748-9

Ibrahim, I., Ayariga, J. A., Xu, J., Adebanjo, A., Robertson, B. K., Samuel-Foo, M., et al. (2023). CBD resistant Salmonella strains are susceptible to epsilon 34 phage tailspike protein. *Front. Med.* 10:1075698. doi: 10.3389/fmed.2023.1075698

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M. B., Pope, B. J., Tomita, T., et al. (2014). SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* 6, 1–16. doi: 10.1186/s13073-014-0090-6

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30, 3059–3066. doi: 10.1093/nar/gkf436

Kim, W.-S., Salm, H., and Geider, K. (2004). Expression of bacteriophage φEa1h lysozyme in Escherichia coli and its activity in growth inhibition of Erwinia amylovora. *Microbiology* 150, 2707–2714. doi: 10.1099/mic.0.27224-0

Knecht, L. E., Veljkovic, M., and Fieseler, L. (2020). Diversity and function of phage encoded depolymerases. *Front. Microbiol.* 10:2949. doi: 10.3389/fmicb.2019.02949

Latka, A., Maciejewska, B., Majkowska-Skrobek, G., Briers, Y., and Drulis-Kawa, Z. (2017). Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. *Appl. Microbiol. Biotechnol.* 101, 3103–3119. doi: 10.1007/s00253-017-8224-6

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Malnoy, M., Faize, M., Venisse, J.-S., Geider, K., and Chevreau, E. (2005). Expression of viral EPS-depolymerase reduces fire blight susceptibility in transgenic pear. *Plant Cell Rep.* 23, 632–638. doi: 10.1007/s00299-004-0855-2

Miletic, S., Simpson, D. J., Szymanski, C. M., Deyholos, M. K., and Menassa, R. (2016). A plant-produced bacteriophage tailspike protein for the control of salmonella. *Front. Plant Sci.* 6:1221. doi: 10.3389/fpls.2015.01221

National Center for Biotechnology Information (2023). *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information, Bethesda.

Olszak, T., Shneider, M. M., Latka, A., Maciejewska, B., Browning, C., Sycheva, L. V., et al. (2017). The O-specific polysaccharide lyase from the phage LKA1 tailspike reduces Pseudomonas virulence. *Sci. Rep.* 7:16302. doi: 10.1038/s41598-017-16411-4

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Roach, D. R., and Donovan, D. M. (2015). Antimicrobial bacteriophage-derived proteins and therapeutic applications. *Bacteriophage* 5:e1062590. doi: 10.1080/21597081.2015.1062590

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50, D20–D26. doi: 10.1093/nar/gkab1112

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE* 11:e0163962. doi: 10.1371/journal.pone.0163962

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., et al. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 51, D523–D531. doi: 10.1093/nar/gkac1052

Turner, D., Shkoporov, A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfenas-Zerbini, P., et al. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.* 168:74. doi: 10.1007/s00705-022-05694-2

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucl. Acids Res.* 42, D581–D591. doi: 10.1093/nar/gkt1099

Yan, J., Mao, J., and Xie, J. (2014). Bacteriophage polysaccharide depolymerases and biomedical applications. *BioDrugs* 28, 265–274. doi: 10.1007/s40259-013-0081-y