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

Opeyemi U. Lawal[1] and Lawrence Goodridge[1]**

[1]Canadian Research Institute for Food Safety (CRIFS), Department of Food Science, University of Guelph, Ontario, Canada. N1G 2W1

*Correspondence:

Dr. Opeyemi U. Lawal: lawal@uoguelph.ca

Dr. Lawrence Goodridge: goodridl@uoguelph.ca

Abstract

Phages are ubiquitous viruses that drive bacterial evolution through infection and replication within host bacteria. Phage tailspike proteins (TSPs) are key components of phage tail structures, exhibiting polysaccharide depolymerase activity and host specificity. Despite their potential as novel antimicrobials, few TSPs have been fully characterized due to laborious detection techniques. To address this, we present TSPDB, a curated resource for rapid detection of TSPs in genomics and metagenomics sequence data. We mined public databases, obtaining 17,211 TSP sequences, which were filtered to exclude duplicates and partial sequences, resulting in 8,099 unique TSP sequences. TSPDB contains TSPs from over 400 bacterial genera, with significant diversity among them as revealed by the phylogenetic analysis. The top 13 genera represented were Gram-positive, with *Bacillus*, *Streptococcus*, and *Clostridium* being the most common. Of note, Phage TSPs in Gram-positive bacteria were on average 1 Kbp larger than those in Gram-negative bacteria. TSPDB has been applied in a recent study to screen phage genomes, demonstrating its potential for functional annotation. TSPDB serves as a comprehensive repository and a resource for researchers in phage biology, particularly in phage associated therapy and antimicrobial or biocontrol applications. TSPDB is compatible with bioinformatics tools for *in silico* detection of TSPs in genomics and metagenomic data, and is freely accessible on GitHub and Figshare, providing a valuable resource for the scientific community.

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 towards 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 and colleagues (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 the presence of TSPs in prophages. The development of a database for TSPs would 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).

Here, we present a high-level curated resource called TSPDB for the rapid detection of tailspike proteins in multiomics sequence data.

Data and Methodology

<u>*Data Mining and Quality Check*</u>: 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." Hits were systematically filtered 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% nucleotide similarity and coverage with cd-hit (Li and Godzik, 2006) and Seqkit (Shen et al., 2016), resulting in 9,129 unique TSP sequences (**Figure 1**).

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 or incomplete sequences, were excluded from the dataset. This filtering process resulted in a total of 8,099 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 (**Figure 2A**). Further analysis of TSP genes in the TSPDB reveals a significant difference 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=1616) being the most common, followed by *Streptococcus* (n=1152), *Clostridium* (n=683), *Enterococcus* (n=387), and *Staphylococcus* (n=372). Additionally, TSPs from Gram-negative bacterial genera, *Salmonella* (n=75), *Escherichia* (n=58), *Klebsiella* (n=52), and *Pseudomonas* (n=25) were among the top 38 TSPs in the database (**Figure 2C**).

Diversity of TSPs: To assess the diversity of the 8,099 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 with 1000 bootstrap replicates for node support was constructed using FastTree v2.1.11 (Price et al., 2010). The resulting phylogenetic tree was visualized using the web-based Microreact visualization tool (Argimón et al., 2016) (**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 to screen for the presence of TSPs in a collection of phage genomes using stringent parameters (≥ 90% identity and ≥ 70% coverage). Overall, the TSPDB contains a vast dataset of diverse TSPs found in phages, making it an essential tool for detecting TSPs within large genomic and metagenomic datasets. Integration of this database into phage detection tools will enhance the functional annotation of these genes. The TSPDB described here will undergo regular updates to include new TSP genes as they become available in public databases.

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 datasets associated with this study are hosted in online repositories. Details of the repository/repositories and accession numbers can be found in the links provided in the manuscript.

Funding: This work was supported by the Canada First Research Excellence Fund, and a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to LG.

Author Contributions: OL: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft and review and editing; LG: Conceptualization, Writing – review and editing, Funding acquisition and Resources

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., Feil, E.J., Holden, M.T.G., Yeats, C.A., Grundmann, H., Spratt, G., Aanensen, D.M., 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microbial Genomics 2, 1–11. https://doi.org/10.1099/mgen.0.000093

Ayariga, J.A., Gildea, L., Wu, H., Villafane, R., 2021. The Ɛ34 Phage Tailspike Protein: An In vitro Characterization, Structure Prediction, Potential Interaction with S. newington LPS and Cytotoxicity Assessment to Animal Cell Line. Journal of Clinical Trials 11, 1–18.

Bhandare, S., Lawal, O.U., Colavecchio, A., Cadieux, B., Zahirovich-Jovich, Y., Zhong, Z., Tompkins, E., Amitrano, M., Kukavica-Ibrulj, I., Boyle, B., Wang, S., Levesque, R.C., Delaquis, P., Danyluk, M., Goodridge, L., 2024. Genomic and Phenotypic Analysis of Salmonella enterica Bacteriophages Identifies Two Novel Phage Species. Microorganisms 12, 1–17. https://doi.org/doi.org/10.3390/microorganisms12040695

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

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. https://doi.org/10.1186/1471-2105-10-421

Chatterjee, A., Duerkop, B.A., 2018. Beyond bacteria: Bacteriophage-eukaryotic host interactions reveal emerging paradigms of health and disease. Frontiers in Microbiology 9, 1–8. https://doi.org/10.3389/fmicb.2018.01394

Chen, C., Bales, P., Greenfield, J., Heselpoth, R.D., Nelson, D.C., Herzberg, O., 2014. Crystal Structure of ORF210 from E. coli O157:H1 Phage CBA120 (TSP1), a Putative Tailspike Protein. PLoS ONE 9, e93156. https://doi.org/10.1371/journal.pone.0093156

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

Clokie, M.R.J., Millard, A.D., Letarov, A.V., Heaphy, S., 2011. Phages in nature. Bacteriophage 1, 31–45.

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

Dowah, A.S.A., 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. https://doi.org/10.1007/s12551-017-0382-3

Emond-Rheault, J.-G., Jeukens, J., Freschi, L., Kukavica-Ibrulj, I., Boyle, B., Dupont, M.-J., Colavecchio, A., Barrere, V., Cadieux, B., Arya, G., Bekal, S., Berry, C., Burnett, E., Cavestri, C., Chapin, T.K., Crouse, A., Daigle, F., Danyluk, M.D., Delaquis, P., Dewar, K., Doualla-Bell, F., Fliss, I., Fong, K., Fournier, E., Franz, E., Garduno, R., Gill, A., Gruenheid, S., Harris, L., Huang, C.B., Huang, H., Johnson, R., Joly, Y., Kerhoas, M., Kong, N., Lapointe, G., Larivière, L., Loignon, S., Malo, D., Moineau, S., Mottawea, W., Mukhopadhyay, K., Nadon, C., Nash, J., Ngueng Feze, I., Ogunremi, D., Perets, A., Pilar, A.V., Reimer, A.R., Robertson, J., Rohde, J., Sanderson, K.E., Song, L., Stephan, R., Tamber, S., Thomassin, P., Tremblay, D., Usongo, V., Vincent, C., Wang, S., Weadge, J.T., Wiedmann, M., Wijnands, L., Wilson, E.D., Wittum, T., Yoshida, C., Youfsi, K., Zhu, L., Weimer, B.C., Goodridge, L., Levesque, R.C., 2017. A Syst-OMICS Approach to Ensuring Food Safety and Reducing the Economic Burden of Salmonellosis. Frontiers in Microbiology 8.

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

Gil, J., Paulson, J., Brown, M., Zahn, H., Nguyen, M.M., Eisenberg, M., Erickson, S., 2023. Tailoring the Host Range of Ackermannviridae Bacteriophages through Chimeric Tailspike Proteins. Viruses 15, 286. https://doi.org/10.3390/v15020286

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

Ibrahim, I., Ayariga, J.A., Xu, J., Adebanjo, A., Robertson, B.K., Samuel-Foo, M., Ajayi, O.S., 2023. CBD resistant Salmonella strains are susceptible to epsilon 34 phage tailspike protein. Front. Med. 10, 1075698. https://doi.org/10.3389/fmed.2023.1075698

Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M.B., Pope, B.J., Tomita, T., Zobel, J., Holt, K.E., 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs.

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

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

Knecht, L.E., Veljkovic, M., Fieseler, L., 2020. Diversity and Function of Phage Encoded Depolymerases. Front. Microbiol. 10, 2949. https://doi.org/10.3389/fmicb.2019.02949

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

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

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

Miletic, S., Simpson, D.J., Szymanski, C.M., Deyholos, M.K., Menassa, R., 2016. A Plant-Produced Bacteriophage Tailspike Protein for the Control of Salmonella. Front. Plant Sci. 6. https://doi.org/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., Cornelissen, A., Danis-Wlodarczyk, K., Senchenkova, S.N., Shashkov, A.S., Gula, G., Arabski, M., Wasik, S., Miroshnikov, K.A., Lavigne, R., Leiman, P.G., Knirel, Y.A., Drulis-Kawa, Z., 2017. The O-specific polysaccharide lyase from the phage LKA1 tailspike reduces Pseudomonas virulence. Sci Rep 7, 16302. https://doi.org/10.1038/s41598-017-16411-4

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE 5, 1–10. https://doi.org/10.1371/journal.pone.0009490

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

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

Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS ONE 11, e0163962. https://doi.org/10.1371/journal.pone.0163962

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L.J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., Kandasaamy, S., Lock, A., Luciani, A., Lugaric, M., Luo, J., Lussi, Y., MacDougall, A., Madeira, F., Mahmoudy, M., Mishra, A., Moulang, K., Nightingale, A., Pundir, S., Qi, G., Raj, S., Raposo, P., Rice, D.L., Saidi, R., Santos, R., Speretta, E., Stephenson, J., Totoo, P., Turner, E., Tyagi, N., Vasudev, P., Warner, K., Watkins, X., Zaru, R., Zellner, H., Bridge, A.J., Aimo, L., Argoud-Puy, G., Auchincloss, A.H., Axelsen, K.B., Bansal, P., Baratin, D., Batista Neto, T.M., Blatter, M.-C., Bolleman, J.T., Boutet, E., Breuza, L., Gil, B.C., Casals-Casas, C., Echioukh, K.C., Coudert, E., Cuche, B., de Castro, E., Estreicher, A., Famiglietti, M.L., Feuermann, M., Gasteiger, E., Gaudet, P., Gehant, S., Gerritsen, V., Gos, A., Gruaz, N., Hulo, C., Hyka-Nouspikel, N., Jungo, F., Kerhornou, A., Le Mercier, P., Lieberherr, D., Masson, P., Morgat, A., Muthukrishnan, V., Paesano, S., Pedruzzi, I., Pilbout, S., Pourcel, L., Poux, S., Pozzato, M., Pruess, M., Redaschi, N., Rivoire, C., Sigrist, C.J.A., Sonesson, K., Sundaram, S., Wu, C.H., Arighi, C.N., Arminski, L., Chen, C., Chen, Y., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Zhang, J., 2023. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51, D523–D531. https://doi.org/10.1093/nar/gkac1052

Turner, D., Shkoporov, A.N., Lood, C., Millard, A.D., Dutilh, B.E., Alfenas-Zerbini, P., Van Zyl, L.J., Aziz, R.K., Oksanen, H.M., Poranen, M.M., Kropinski, A.M., Barylski, J., Brister, J.R., Chanisvili, N., Edwards, R.A., Enault, F., Gillis, A., Knezevic, P., Krupovic, M., Kurtböke, I., Kushkina, A., Lavigne, R., Lehman, S., Lobocka, M., Moraru, C., Moreno Switt, A., Morozova, V., Nakavuma, J., Reyes Muñoz, A., Rūmnieks, J., Sarkar, B., Sullivan, M.B., Uchiyama, J., Wittmann, J., Yigang, T., Adriaenssens, E.M., 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. https://doi.org/10.1007/s00705-022-05694-2

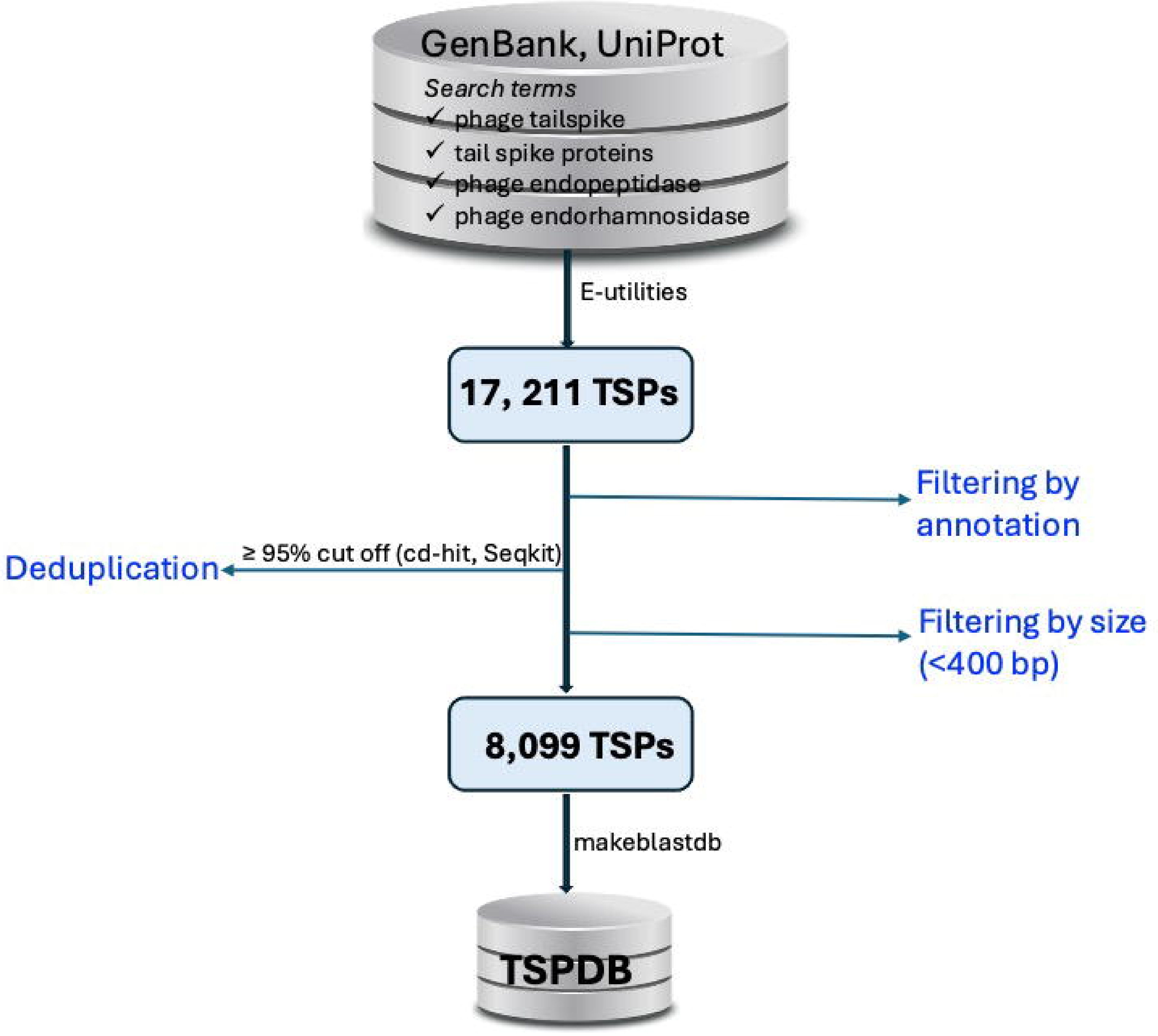
Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., Gillespie, J.J., Gough, R., Hix, D., Kenyon, R., Machi, D., Mao, C., Nordberg, E.K., Olson, R., Overbeek, R., Pusch, G.D., Shukla, M., Schulman, J., Stevens, R.L., Sullivan, D.E., Vonstein, V., Warren, A., Will, R., Wilson, M.J.C., Yoo, H.S., Zhang, C., Zhang, Y., Sobral, B.W., 2014. PATRIC, the bacterial bioinformatics database and analysis resource. Nucl. Acids Res. 42, D581–D591. https://doi.org/10.1093/nar/gkt1099

Yan, J., Mao, J., Xie, J., 2014. Bacteriophage Polysaccharide Depolymerases and Biomedical Applications. BioDrugs 28, 265–274. https://doi.org/10.1007/s40259-013-0081-y

Figure Legend

Figure 1 – Workflow for the construction of the tailspike protein database (TSPDB).

Figure 2 – Analysis of Phage tail spike proteins in the TSPDB. (A). Sequence length distribution of genes encoding phage TSPs contained in the TSPDB. (B). Frequency of top 37 genera of host phages carrying TSPs in the TSPDB. (C). Differential TSPs size between Gram-negative and Gram-positive bacteria in the TSPDB. (D). Phylogenetic diversity of the 8,099 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 colour. An interactive version of this figure is accessible through the following link - https://microreact.org/project/7Kv61nb6aRapgGgHpxsNGL-tspdb-v20.


GenBank, UniProt
Search terms
✓ phage tailspike
✓ tail spike proteins
✓ phage endopeptidase
✓ phage endorhamnosidase
E-utilities
17, 211 TSPs
Filtering by annotation
≥ 95% cut off (cd-hit, Seqkit)
Deduplication
Filtering by size (<400 bp)
8,099 TSPs
makeblastdb
TSPDB

A
Frequency
400
300
200
100
0
0
2000
4000
6000
8000
10000
Sequence length (bp)
B
TSPs from Gram negative bacteria
TSPs from Gram-positive bacteria
80
60
40
20
0
Frequency
Size (bp)
0
2500
5000
7500
10000
C
Frequency
1500
1000
500
0
1616
1152
683
387
372
217
190
148
142
138
137
87
87
75
73
70
63
60
58
54
53
48
48
41
41
40
38
36
33
32
29
28
26
26
26
26
25
Bacillus
Streptococcus
Clostridium
Enterococcus
Staphylococcus
Lactiplantibacillus
Lactococcus
Paenibacillus
Lactobacillus
Listeria
Lacticaseibacillus
Leuconostoc
Ruminococcus
Salmonella
Macrococcus
Ligilactobacillus
Pediococcus
Bifidobacterium
Escherichia
Cytobacillus
Weissella
Klebsiella
Virgibacillus
Caudoviricetes
Blautia
Limosilactobacillus
Eubacterium
Priestia
Oceanobacillus
Paraclostridium
Carnobacterium
Paeniclostridium
Clostridioides
Levilactobacillus
Mammaliicoccus
Dorea
Pseudomonas
Genera
D
0.74
Bacillus
Bifidobacterium
Blautia
Carnobacterium
Caudoviricetes
Clostridioides
Clostridium
Cytobacillus
Dorea
Enterococcus
Escherichia
Eubacterium
Klebsiella
Lacticaseibacillus
Lactiplantibacillus
Lactobacillus
Lactococcus
Leuconostoc
Levilactobacillus
Ligilactobacillus
Limosilactobacillus
Listeria
Macrococcus
Mammalicoccus
Oceanobacillus
Paenibacillus
Paeniclostridium
Paraclostridium
Pediococcus
Priestia
Pseudomonas
Ruminococcus
Salmonella
Staphylococcus
Streptococcus
Uncultured
Virgibacillus
Weissella