OXFORD (GIGA)n SCIENCE

GigaScience, 2024, **13**, 1–14
DOI: 10.1093/gigascience/giae017
Research

Large-scale genomic survey with deep learning-based method reveals strain-level phage specificity determinants

Yiyan Yang [1], Keith Dufault-Thompson [1], Wei Yan [1], Tian Cai[2], Lei Xie [2,3], and Xiaofang Jiang [1,*]

[1]National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[2]Ph.D. Program in Computer Science, The Graduate Center, The City University of New York, New York, NY 10016, USA
[3]Department of Computer Science, Hunter College, The City University of New York, New York, NY 10065, USA
*Correspondence address. Xiaofang Jiang, Building 38A, Room 6N607, 8600 Rockville Pike, Bethesda, MD 20894, USA. E-mail: xiaofang.jiang@nih.gov

Abstract

Background: Phage therapy, reemerging as a promising approach to counter antimicrobial-resistant infections, relies on a comprehensive understanding of the specificity of individual phages. Yet the significant diversity within phage populations presents a considerable challenge. Currently, there is a notable lack of tools designed for large-scale characterization of phage receptor-binding proteins, which are crucial in determining the phage host range.

Results: In this study, we present SpikeHunter, a deep learning method based on the ESM-2 protein language model. With SpikeHunter, we identified 231,965 diverse phage-encoded tailspike proteins, a crucial determinant of phage specificity that targets bacterial polysaccharide receptors, across 787,566 bacterial genomes from 5 virulent, antibiotic-resistant pathogens. Notably, 86.60% (143,200) of these proteins exhibited strong associations with specific bacterial polysaccharides. We discovered that phages with identical tailspike proteins can infect different bacterial species with similar polysaccharide receptors, underscoring the pivotal role of tailspike proteins in determining host range. The specificity is mainly attributed to the protein's C-terminal domain, which strictly correlates with host specificity during domain swapping in tailspike proteins. Importantly, our dataset-driven predictions of phage–host specificity closely match the phage–host pairs observed in real-world phage therapy cases we studied.

Conclusions: Our research provides a rich resource, including both the method and a database derived from a large-scale genomics survey. This substantially enhances understanding of phage specificity determinants at the strain level and offers a valuable framework for guiding phage selection in therapeutic applications.

Keywords: phage–host specificity, phage receptor-binding protein, bacterial polysaccharide, serotype, phage therapy

Introduction

Phage therapy is gaining renewed interest as a solution to antimicrobial resistance, which is reflected by the increased number of case reports describing the use of phage treatments [1–4]. Understanding phage–host interactions and the determinants of phage host specificity at the strain and species level is crucial for improving phage therapy and applying phage proteins in biotechnology [5–7]. Extensive work has been done to understand what bacteria specific phages can infect and which proteins are involved in this process [8–10]. Phage receptor-binding proteins, in particular, are gaining popularity due to their direct interaction with host receptors and their potential to streamline the selection of effective phages for therapy, a key bottleneck in its broader application [11–13].

Tailspike proteins are a type of phage receptor-binding protein that specifically recognizes and breaks down bacterial cell surface polysaccharides, such as the capsular polysaccharides (K-antigens), the O-specific polysaccharides of the lipopolysaccharide (O-antigens), and the outer core (OC) of the lipooligosaccharide (OC-antigen) to initiate infection [14–17]. They have also shown promise as antimicrobials that can be used to sensitize resistant strains [18] and degrade biofilms [19]. Given their central role in phage–host interactions, many studies have attempted to understand the associations of tailspike proteins with bacterial serotypes, the types of polysaccharides on the bacterial cell surface. Recent studies have shown that the host range of *Klebsiella* phages is generally restricted and have shed light on the role of phage tailspike proteins in defining this specificity [9]. Similarly, other studies related to Ackermannviridae and *Escherichia* viruses have demonstrated that tailspike proteins are tightly associated with host serotype and that recombination of tailspike protein domains may be a key driver in tailspike protein evolution [9, 10, 20].

Experimental phage host–range determination is laborious and time-consuming. While these studies have provided valuable examples of the association of phage receptor-binding proteins with specific host receptors, their scale and scope, often focusing on single bacterial species and fewer than a hundred phages, limits their generalizability and predictive power. Computational approaches, which predict host ranges utilizing genomic information [12], offer a significant advantage. Methods employing large-scale genomic data grounded in phage host–range mechanisms can provide enhanced sensitivity and predictive power at finer taxonomic resolutions, thereby supporting future phage therapy initiatives.

In this study, we conducted a large-scale, multispecies genomic analysis to better understand the role of tailspike proteins in phage host specificity at the strain level. Utilizing a deep learning–based method, SpikeHunter, we identified tailspike proteins

Received: September 28, 2023. **Revised:** January 23, 2024. **Accepted:** March 24, 2024

specific to serotypes from 5 prevalent human pathogens and created a comprehensive phage–host association database. Our findings indicate that host range is primarily governed by the specificity of the tailspike protein, enabling phages with identical tailspike proteins to infect diverse bacterial species sharing the same serotypes. This specificity is mainly attributed to the C-terminal domain, as host specificity was observed to strictly follow this domain during extensive domain swapping in tailspike proteins. Furthermore, our dataset-driven phage–host specificity predictions align well with established phage–host pairs employed in real-world phage therapy cases. By expanding the knowledge of the molecular basis of phage host specificity, our research enhances both the applications and the engineering of phages to target new strains [21] or circumvent bacterial resistance [22], thereby advancing phage therapy. The analysis performed in this study is provided at [23]. The expansive dataset of tailspike proteins and the SpikeHunter model are available at TailspikeDB [24] and via GitHub [25], respectively, which can guide future phage applications and predictions on phage host range.

Methods

Training and validation data

A total of 3,659 bacteriophage genomes were downloaded from the INPHARED database (v1.7) [26]. This collection of proteins was split into tailspike and non-tailspike phage protein datasets. A possible tailspike protein dataset was generated based on keyword searches and comparisons to other annotated viral proteins in the NCBI nr database using BlastP, PDB using SCOP (RRID:SCR_007039) [27], and the viral ortholog databases, PHROG [28], pVOG [29], ViPhOG [30], eggNOG viral ortholog groups (RRID:SCR_002456) [31], and VOGDB [32], using HMMER (RRID:SCR_005305). Proteins that were annotated as tailspike, tail fiber, or receptor binding were included in the candidate tailspike dataset along with proteins whose top hit was annotated as tailspike proteins in other databases. The set of candidate tailspike proteins were then clustered at 70% identity using CD-HIT (RRID:SCR_007105) [33,34] and their structures were predicted using AlphaFold v2.3.2 [35]. The structures were then manually curated to identify a final set of 1,912 tailspike proteins based on the presence of the distinctive beta-helix receptor-binding domain. The remaining 200,732 non-tailspike proteins (excluding those classified as part of the "unknown" category in INPHARED) were included in the non-tailspike dataset.

Independent testing data

The independent dataset was compiled using a recently annotated dataset of 100,081 proteins from 96 phages that infect 403 strains of the *Escherichia* genus [36]. Within this dataset, 81 curated tailspike proteins were designated as positive samples, and the rest of the proteins were categorized as negative samples.

Model architecture

The SpikeHunter (RRID:SCR_024831) was developed using the PyTorch framework (RRID:SCR_018536) [37]. First, the phage sequences are first tokenized and transformed into numerical vectors using the *batch_converter* function in the ESM python package [38]. The sequences are then embedded as 1,280 length representations using a pretrained transformer protein language model ESM-2 (esm2_t33_650M_UR50D) [39]. The sequence representations are fed into a 4-fully-connected layer network with 1280, 568, 128, and 2 nodes, respectively. The output from the last layer is converted into a probability representing each sequence being a tailspike protein or not with a softmax activation function. The SpikeHunter model and code is available on GitHub [25].

We further conducted ablation studies on SpikeHunter by altering its modules to analyze the effect of the different components. The modifications included (i) removing the 568-neuron hidden layer from the fully connected layers; (ii) removing the 128-neuron hidden layer from the fully connected layers; (iii) replacing the pretrained ESM-2 encoder, used for input sequence embedding, with the SeqVec encoder [40]; and (iv) integrating dropout layers with a 0.2 dropout ratio into each linear layer of the fully connected layers. After these architectural changes, the models were retrained. Their performance on the validation dataset was then evaluated using metrics such as accuracy, precision, recall, specificity, F1-score, and Matthew's correlation coefficient (MCC). These modified models are available on GitHub at [23].

Training and validation of the deep learning model

To train and validate the SpikeHunter, the manually curated set of phage proteins, consisting of both tailspike proteins and non-tailspike proteins, was first clustered into 20,274 clusters at 30% identity using CD-HIT [34]. Each cluster contained only tailspike or non-tailspike proteins, with no mixed clusters being observed in the dataset. The sequences were then divided into training, validation, and testing datasets in a ratio of 3:1:1 using the *StratifiedGroupKFold* function in the *Scikit-learn* python package [41], resulting in a training set of 122,506 proteins (comprising 1,023 positive samples and 121,483 negative samples belonging to 12,170 clusters), a validation set of 40,838 proteins (comprising 343 positive samples and 40,495 negative samples belonging to 4,054 clusters), and a testing set of 39,300 proteins (comprising 546 positive samples and 38,754 negative samples belonging to 4,050 clusters). The model training was performed with the cross-entropy loss function and the Pytorch implementation of the Adam optimizer, with the parameters of the ESM-2 model frozen. The training was halted when the model's performance on the validation dataset did not improve for 3 consecutive epochs. The model with the lowest validation loss was then used for testing and prediction.

Identification of tailspike proteins in bacterial genomes

A total of 787,566 genomes of 5 common pathogens, *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas aeruginosa*, *Salmonella enterica*, and *Acinetobacter baumannii*, were obtained from the NCBI Pathogen Detection database [42], downloaded on 2 April 2023 [43]. Prophage regions were predicted in the pathogen genomes using VIBRANT (version 1.2.0) with default parameters to identify phages within the bacterial genomes. The protein sequences of the phages with lengths greater than 200 amino acids were then extracted, and SpikeHunter was used to classify them as either tailspike or non-tailspike proteins. Only proteins with greater than 50% probability of being a tailspike protein were positive hits for tailspike proteins. All identified tailspike protein IDs with their corresponding clusters at various protein identities are provided at [23].

Bacterial genome serotyping

Serotype prediction for the 5 analyzed microbial species was performed using a variety of species-specific tools. O-antigen

serotypes were predicted for *E. coli* using ECTyper [44], SerotypeFinder [45], and fastKaptive [46]. Serotypes were grouped into serogroups according to Iguchi et al. [47]. K-antigen prediction for *E. coli* genomes was done using fastKaptive [46]. *K. pneumoniae* K- and O-antigen serotypes were inferred using Kaptive (RRID: SCR_024046) [48] and fastKaptive [46]. *A. baumannii* K-antigen and OC serotypes were predicted using Kaptive [48]. *P. aeruginosa* O-antigen serotypes were predicted using PAst [49]. O-antigen serotypes were predicted for *S. enterica* using SeqSero2 [50] and fastKaptive [46]. All tools were run using default settings. Results from Kaptive with match confidence scores of "None" and results from fastKaptive with best match coverages lower than 90 were excluded from the downstream analysis. To facilitate the comparison of serotypes between species, shared serotypes were merged and consistently named for the set of *K. pneumoniae* and *E. coli* K-antigens and the set of *E. coli* and *S. enterica* O-antigens. Common serotypes between *K. pneumoniae* and *E. coli* were identified by predicting serotypes in the *K. pneumoniae* genome using the fastKaptive tool. The predicted *K. pneumoniae* serotypes from Kaptive and fastKaptive were then used to map the shared serotypes between the 2 species. The same procedure was followed for identifying common serotypes between *S. enterica* and *E. coli*, but with the *S. enterica* genomes being annotated using SeqSero2 and fastKaptive to provide the mapping to the *E. coli* genomes.

Associating tailspike proteins with serotypes

Tailspike protein sequences were hierarchically clustered at 30%, 40%, 50%, 60%, 70%, 80%, 85%, 90%, and 95% identities using cd-hit v.4.8.1 [34]. The tailspike protein clusters at different levels were associated with the vOTUs (viral Operational Taxonomic Units) they were found in and the serotypes of the genomes that the vOTUs were in. In this study, a vOTU is defined as a cluster that contains phage genomes with an Average Nucleotide Identity (ANI) ≥ 95% and an alignment coverage of shorter sequence ≥ 85%. When encountering multiple instances of the same 95% tailspike protein cluster, vOTU, and serotype within the dataset, only 1 instance was retained for subsequent analysis to eliminate redundancy. An overall network of tailspike protein clusters at 60% identity and serotypes was then generated with the links between them being determined by the vOTUs. Associations between the tailspike protein clusters at 60% identity and serotypes were classified into 3 categories: "highly confident," "confident," and "uncertain." "Highly confident" associations were tailspike to serotype pairs that were supported by at least 90% of the vOTUs containing that 60% tailspike protein cluster for tailspike proteins with more than 4 vOTU connections or that were supported by all vOTUs for tailspike protein clusters with less than or equal to 4 vOTUs. "Confident" associations were tailspike to serotype pairs that were supported by 10% to 90% of the vOTUs with the tailspike protein cluster, and the remaining associations were classified as "uncertain." More information regarding the assignment of serotypes to a tailspike protein cluster can be found within the code hosted on GitHub [23]. A database featuring tailspike proteins and their associations with serotypes is accessed at TailSpikeDB [24].

Identification of domain swapping between tailspike proteins

An all-by-all BlastP [51] search was performed to compare all nonredundant tailspike protein sequences, and the hits were filtered using an e-value cutoff of 1e-8, identity threshold of 60%, and a coverage range of 10% to 90% of the sequence. Candidate domain swapping was found by identifying proteins that aligned to at least 2 other proteins that cumulatively covered at least 90% of the original query sequence. Breakpoints for the N- and C-terminal domains were determined by computing the average alignment start and end positions from all of the alignments to each query tailspike protein. The domain sequences were then extracted and clustered using psi-cd-hit v.4.8.1 at 30% and 95% identity thresholds [52]. Domain swapping was only considered plausible if they were validated by multiple proteins with the same domain at 95% identity while having different versions of the other domain at 30%. Circlize v0.4.15 was used to visualize the N/C-terminal domain swapping for each species [53].

The collection of phage host range determination trials and phage therapy preclinical and clinical cases

The phage host range determination trial**s** were collected from PhReD [54, 55] and VHRdb [56, 57] databases. The preclinical trials were collected from a list of trials in Gómez-Ochoa et al. [58] while the clinical cases were collected from case studies listed at [59] and cases reported in Green et al. [60]. The phage genomes and bacterial genomes or serotype information were extracted from literature or searched in GenBank. The missing bacterial serotypes were inferred by the serotype tools previously mentioned.

Results

Detecting sequence-divergent tailspike proteins in phages using deep learning

To investigate the association between tailspike proteins and bacterial surface polysaccharide antigens, we first developed a deep learning method named SpikeHunter to identify tailspike proteins (Fig. 1A). SpikeHunter was built on the ESM-2 large protein language model [39] to embed a protein sequence into a representative vector and predict the probability of that protein being a tailspike protein using a fully connected 3-layer neural network (Fig. 1B). A reference set of 1,912 tailspike protein sequences and 200,732 non-tailspike protein sequences was curated from the INPHARED database [26]. The labeled protein sequences were then partitioned into training, validation, and testing datasets at a ratio of 3:1:1. This partition was conducted to preserve an equivalent proportion of positive samples across each set and to ensure that no sequence within a particular set displayed over 30% identity with a sequence in another set (Table 1).

Early stopping was adopted to stop training once the model performance was no longer improved on the validation dataset for 3 consecutive epochs to avoid overfitting (Supplementary Fig. S1). Evaluation of the model on the testing dataset demonstrated that SpikeHunter was accurate and sensitive, achieving an F1-score of 0.99991, precision of 0.99995, recall of 0.99987, specificity of 0.99634, an MCC of 0.99352, and the area under the precision–recall curve (PRAUC) of 0.99360 (Fig. 1C). Evaluated on an independent testing set consisting of 100,081 phage proteins with 81 as positive samples [36], SpikeHunter achieved an F1-score of 0.99985, precision of 1.0, recall of 0.99970, specificity of 1.0, an MCC of 0.98183, and the PRAUC of 0.99985.

We further investigated the effect of data imbalance on model performance. To address this, we oversampled positive samples in the training and validation datasets, creating a balanced dataset with an equal number of positive and negative samples (N = 121,483 for each class for training and N = 40,495 for each class for validation). The comparison between models trained on the original dataset and the balanced dataset revealed minor changes

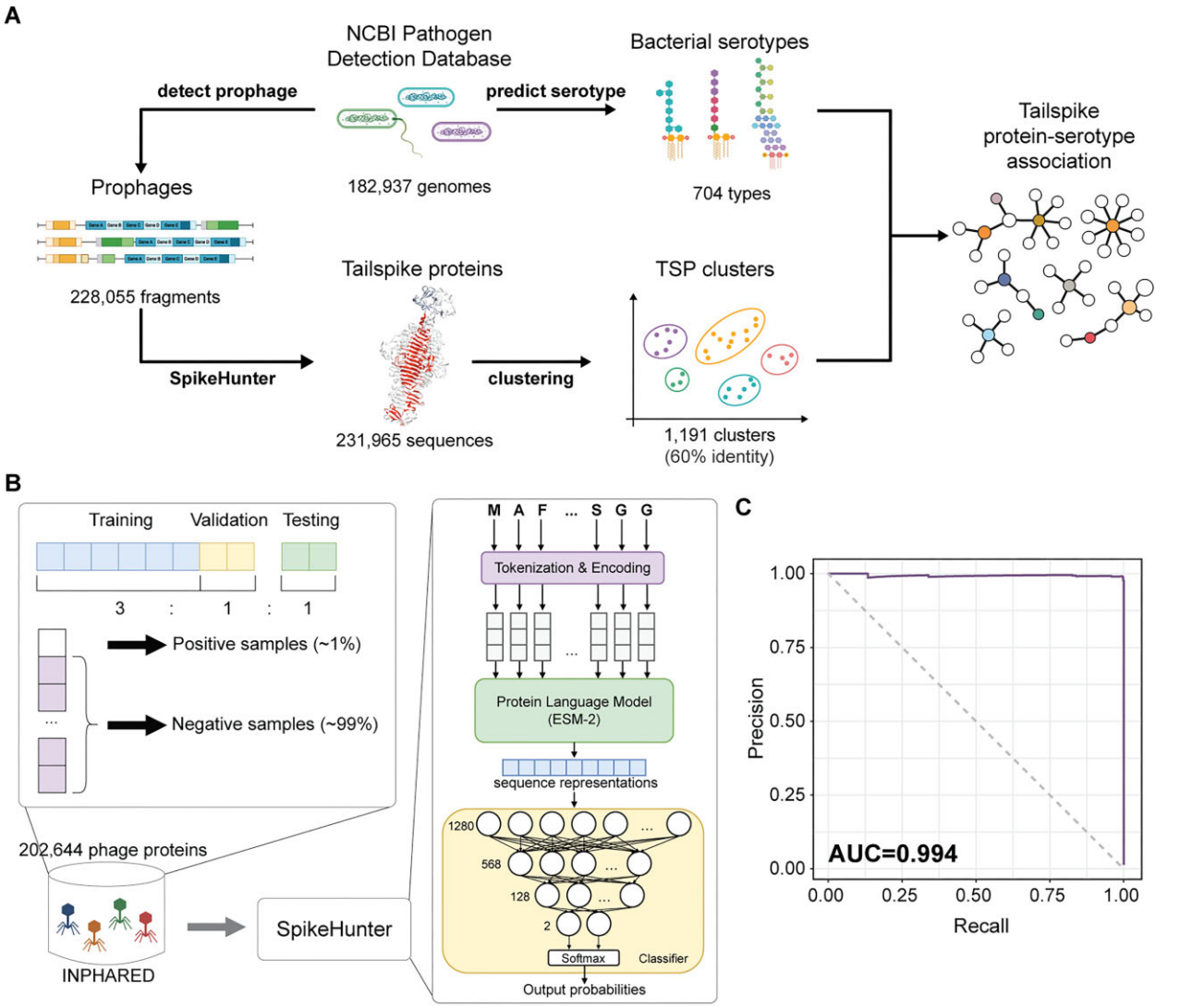

Figure 1: Development of SpikeHunter. (A) Diagram showing the project workflow for the identification of tailspike protein and serotype associations. (B) Organization of the training data and the model architecture for SpikeHunter. (C) The precision–recall curve with an area under the curve for SpikeHunter evaluated on a testing dataset consisting of 39,300 phage proteins.

Table 1: Overview of training, validation, and testing dataset splits

Dataset type	Protein cluster count	Tailspike protein count, *n* (%)	Non-tailspike protein count, *n* (%)	Total protein count	Percentage
Training	12,170	1,023 (0.835%)	121,483 (99.165%)	122,506	60.5%
Validation	4,054	343 (0.840%)	40,495 (99.160%)	40,838	20.2%
Testing	4,050	546 (1.389%)	38,754 (98.611%)	39,300	19.4%

in performance, with a slight increase in F1-score by 1.182% and a small decrease in specificity by 0.025% on the validation data. These results suggest that data imbalance does not significantly impact the model's performance (Supplementary Table S1). Additionally, we evaluated model performance based on ablation studies. Adding dropout layers to, or removing the 568-neuron hidden layer from, the fully connected neural network resulted in unchanged performance for SpikeHunter. However, removing the 128-neuron layer led to decreases in both F1-score (by 0.145%) and specificity (by 0.002%). This indicates the greater importance of the 128-neuron layer compared to the 568-neuron layer in the model, suggesting that the features extracted from this layer are more useful in assisting the model with classification. We also substituted the pretrained ESM-2 model in SpikeHunter with another pretrained model, SeqVec [40], which focuses less on structure, for embedding the input sequences. This substitution resulted in more pronounced declines in performance metrics, with a 6.59% reduction in F1-score and a 0.09% decrease in specificity. This finding indicates that the ESM-2 encoder is a key contributor to SpikeHunter's performance, suggesting that the structural features of the tailspike protein are the key to reliable tailspike protein identification (Supplementary Table S1).

In summary, based on the testing results, we concluded that SpikeHunter is an effective tool for tailspike protein identification and utilized it in subsequent analyses.

Identification of 231,965 prophage-encoded tailspike proteins from the genomes of 5 common pathogens

A total of 787,566 genomes from *E. coli*, *P. aeruginosa*, *Klebsiella*, *Acinetobacter*, and *Salmonella* in the NCBI Pathogen Detection database were analyzed to predict prophages, prophage-encoded tailspike proteins, and bacterial serotypes. Prophages were predicted in 99.4% (783,033) of the genomes, resulting in 8,434,760 prophage genomes containing 177,337,485 protein-encoding genes across the entire dataset (Fig. 2A). The SpikeHunter was used to identify tailspike proteins within the prophages, resulting in the identification of 231,965 tailspike proteins in 10,676,658 proteins from 228,055 prophages, representing 17,932 vOTUs. Despite the relatively small fraction of prophage genomes with a predicted tailspike protein (2.7%), 25.7% of the bacterial genomes analyzed contained at least 1 prophage genome with a predicted tailspike protein. Typically, only 1 tailspike protein was observed in each prophage genome, but 1.67% of prophages (3,814 out of 228,055) were found to encode 2 or more tailspike proteins (Fig. 2A). In the most extreme case, a single prophage from a *K. pneumoniae* (GCA_003037395.1) genome contained 14 predicted and was similar to the ϕKp24 jumbo phage, which is known for its highly branched tail structure and expanded host range [61] (Supplementary Fig. S2). Overall, 208,900 prophages with tailspike proteins were identified in 182,937 bacterial genomes that also had a serotype prediction, providing good coverage across the species and facilitating the subsequent multispecies analysis of tailspike protein–serotype associations.

While most vOTUs (15,022 of 16,535, 90.85%) only contained tailspike proteins from 1 group (clustered at 60% identity), multiple examples (1,513 of 16,535, 9.15%) of vOTUs composed of prophages with tailspike proteins belonging to different groups were found. In 1 example, a vOTU associated with *E. coli* was composed of prophages with 24 distinct tailspike proteins associated with distinct predicted serotypes, including a mix of K- and O-antigens (Supplementary Fig. S3). Similarly, a *Klebsiella*-associated vOTU was identified where the 5 prophages each had a distinct tailspike protein and were associated with different K-antigen serotypes (Fig. 2C). The differences in tailspike protein and serotype associations for these otherwise similar prophages provide further evidence of the importance of the tailspike protein in driving host specificity. Similar tailspike proteins were found in distinct vOTUs from the same genomes (Fig. 2B). Despite the difference in genomic content and organization between these vOTUs, their shared bacterial host further corroborates the link between the tailspike proteins and serotypes.

The distribution of serotypes in relation to tailspike proteins was then investigated. Bacterial serotypes were predicted for each of the bacterial genomes, resulting in 182,937 bacterial genomes that had a predicted serotype and at least 1 prophage with an identified tailspike protein (Fig. 2D). The tailspike proteins were subjected to hierarchical clustering analysis based on 30%, 60%, and 95% identity thresholds, and their relationship with the corresponding host serotypes of their vOTUs was examined. While the specific outcomes varied across different clusters, the predominant serotype associated with each cluster at the 30% identity level generally exhibited consistency with its descendant clusters at the 60% identity level. However, certain exceptions were observed. For instance, the cluster labeled cl30_19 in *Klebsiella*, which emerged at the 30% identity level, displayed high diversity within its descendant cluster at the 60% identity level (Supplementary Fig. S4). Consequently, an identity threshold of 60% was proposed for all tailspike proteins to strike a reasonable balance between cluster purity and reducing redundancy in the results (Supplementary Fig. S5).

Prophage-encoded tailspike proteins are reliable indicators of bacterial polysaccharide receptors

In the overall dataset, there were 1,180 unique tailspike protein to serotype associations identified, of which 715 (60.59%) were found to be strong associations (Fig. 3, Supplementary Fig. S6). A majority of these strong associations were found from the *E. coli* (Fig. 3A) and *Klebsiella* (Fig. 3B) genomes, but multiple strongly associated tailspike–serotype pairs were still predicted for *Salmonella* (Fig. 3C), *Acinetobacter* (Supplementary Fig. S7), *and P. aeruginosa* (Supplementary Fig. S7), providing coverage across all 5 taxa (Supplementary Table S2). These strongly associated tailspike–serotype pairs consisted of 681 unique tailspike clusters and 322 unique serotypes (127 K-antigens and 195 O/OC-antigens). Most of the strongly serotype-associated tailspike proteins were connected to only 1 serotype (99.56%) (Fig. 3), suggesting that these particular phage receptor-binding proteins typically target a narrow range of hosts.

The association between prophage tailspike proteins and bacterial serotypes cannot be attributed to phylogenetic relatedness alone. The loci encoding bacterial surface polysaccharides, which determine serotypes, are known to undergo frequent horizontal gene transfer and exhibit a polyphyletic distribution of serotypes [46]. For instance, the serotype O81 in *E. coli* is distributed among multiple clades, and the tailspike protein cluster cl60_897 is found to strictly follow the clades that encompass the O81 serotype. Another example pertains to the K57 serotype in *Klebsiella*, where the associated tailspike cluster cl60_351 is present in 3 of the 4 main K57 clades (Supplementary Fig. S8). These findings suggest that the distribution of tailspike proteins aligns with serotypes rather than the phylogeny of the bacterial hosts.

Tailspike protein clusters that were strongly associated with multiple serotypes may be indicators of associations with other polysaccharides or shared polysaccharide structures between the serotypes. A few instances were observed in *E. coli* and *Klebsiella* where 1 tailspike protein was associated with 2 differently annotated serotypes. While polysaccharide structural information for many of the serotypes is not available, examples were found for some of these components where the polysaccharide structures were similar. One tailspike protein was found to be associated with the *E. coli* O123 and O186 serotypes and the *S. enterica* O58 serotype, which have all been found to have a shared glycan backbone structure composed of 4-(N-acetylalanyl)amido-4,6-dideoxyglucose, N-acetyl-D-glucosamine, 2,5-dideoxy-2-(3-hydroxybutyramido)-glucose, and N-acetyl-D-glucosamine [62] (Fig. 3A). Similarly, a tailspike protein was strongly associated with the K2 and K13 serotypes in *Klebsiella*, which both share a backbone structure of 2 D-glucose units and a D-mannose unit [63] (Fig. 3B), and strains with these serotypes have been observed to be infected by the same phages [64]. These examples provide further evidence that tailspike proteins are associated with specific polysaccharide structures and can provide valuable information about their associated bacterial host

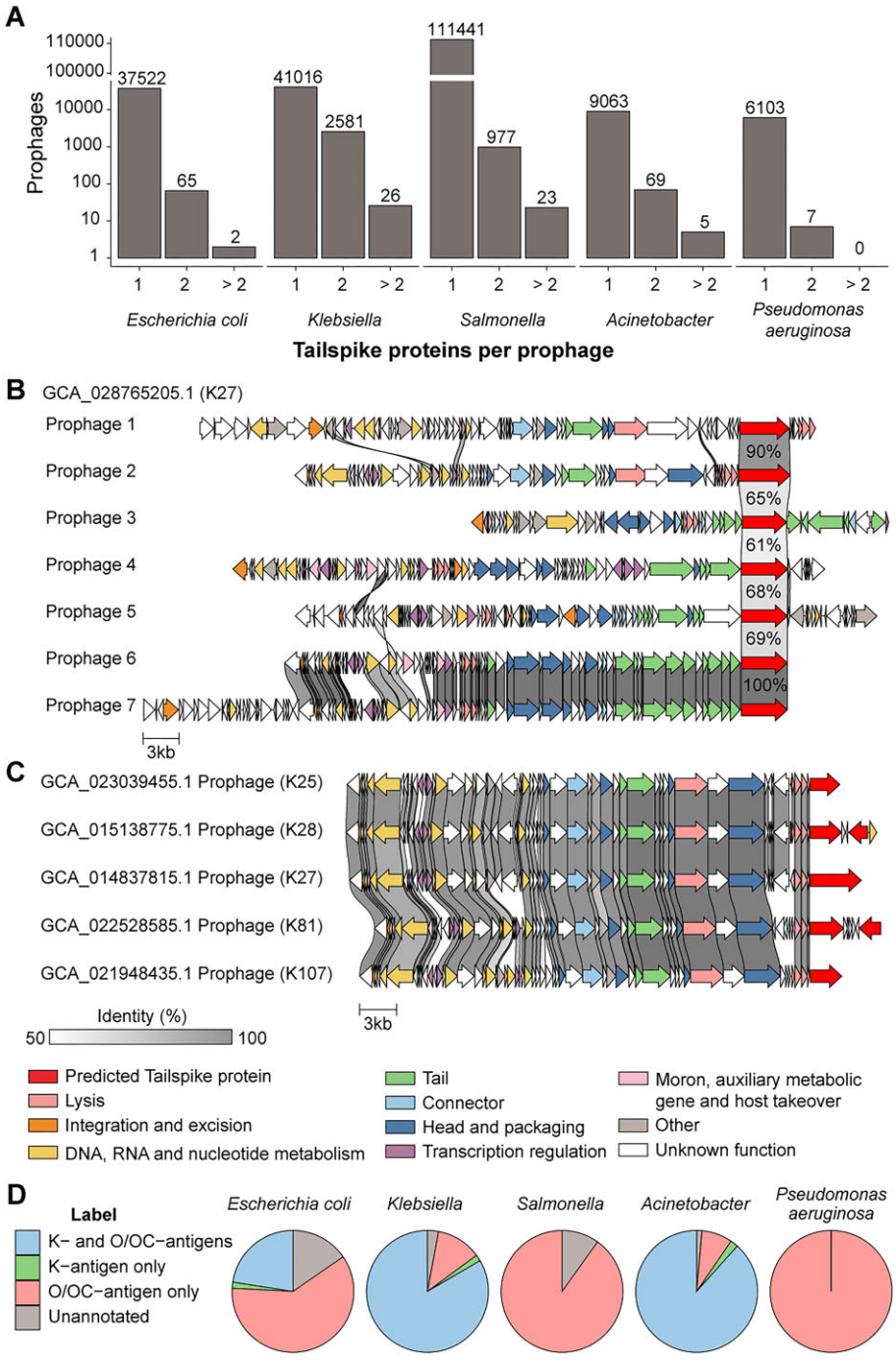

Figure 2: Identification of tailspike proteins and serotypes. (A) Number of prophages detected in the genomes of each pathogen. (B) Prophages detected in a *Klebsiella* genome (GCA_028765205.1) with the same tailspike protein clustered at 60% identity. (C) Similar prophages with different tailspike proteins detected in multiple *Klebsiella* genomes. For panels B and C, the amino acid similarity between genes is shown by the shaded region between the genes and the tailspike proteins are colored in red. (D) Predicted serotypes for genomes of each of the pathogens. The relatively small number of genomes with K-antigen predictions for the *E. coli* can be partially attributed to the lack of group 4 K-antigen prediction, which would be functionally redundant with the strain O-antigen typing.

serotypes. Larger components containing links between multiple tailspike proteins and serotypes were observed for each of the 5 taxa (Fig. 3). While some of these associations may be biologically relevant, these links are likely the result of either unpredicted serotypes that are common to the genomes and associated with the tailspike proteins or due to other factors like the association of the tailspike proteins with polysaccharides that are not K- or O/OC-antigens.

Highly similar tailspike proteins are found in different host species with closely related surface polysaccharides

Highly similar tailspike proteins (above 95% identity) were found in prophages from genomes of different host species. Nearly all the tailspike protein clusters (at 95% identity) were composed of tailspike proteins found in prophages from just 1 species (*i.e.*, only *E. coli*, only *S. enterica*, etc.). However, a small fraction

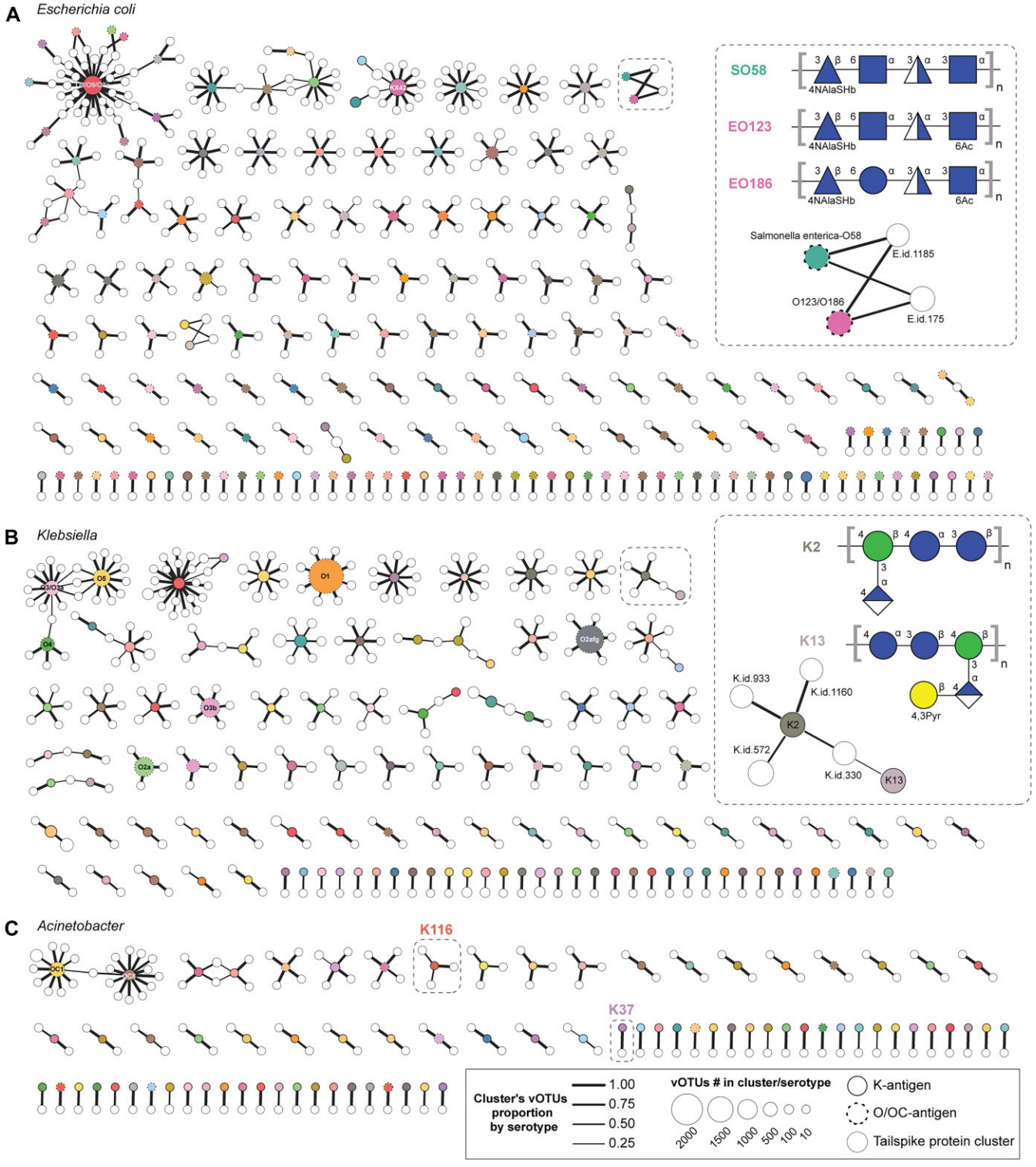

Figure 3: Association of tailspike protein clusters with bacterial serotypes. Networks showing the strongly associated serotypes (colored circles) with tailspike protein clusters (white circles). The size of the circle indicates the number of serotypes or tailspike proteins. Serotype circles with solid outlines represent K-antigens and circles with dashed outlines represent O/OC-antigens. Width of the lines indicates the fraction of vOTUs containing that tailspike protein that support the association. Panels show the tailspike–serotype associations for (A) *E. coli*, (B) *Klebsiella*, and (C) *Acinetobacter*. Insets in panels A and B show the surface polysaccharide structures associated with example mixed serotype clusters.

was made up of a mixture of tailspike proteins from prophages found in *E. coli* and *S. enterica* (26/41) or *E. coli* and *K. pneumoniae* (15/41) genomes (Supplementary Table S3). These mixed-species tailspike clusters were relatively rare but were all supported by at least 2 genomes from each species, reducing the chance that these were due to contamination or errors in the genomes.

The identification of similar tailspike proteins in *E. coli* and *Klebsiella* genomes with the same K-antigen suggests that the relationship between serotype and tailspike protein serves as a pivotal determinant of the host range. Instances of horizontal gene transfer of the K-antigen–specific loci between *E. coli* and *Klebsiella* strains have been documented numerous times [65], providing a possible route for phages to become associated with new host species. When examining the tailspike protein clusters that contained tailspike from prophages from different species, it was found that, despite their shared tailspike proteins, the prophages were distinct from each other (Fig. 4). When examined in the context of their host serotypes, it was found that these different phages from different species were associated with the same or similar serotypes. The *E. coli* genomes containing these tailspike proteins were predicted to have serotypes that were the same as the *Klebsiella* strains in the cluster in the instance of the K47 (Fig. 4A) and K63 (Fig. 4B) serotypes or possessed a similar glycan backbone in the instance of the K13 and K2 serotypes (Fig. 4B and Fig. 3B, Supplementary Table S3).

Examples of similar tailspike proteins were also observed in *E. coli* and *Salmonella* genomes, suggesting that they shared similar O-antigens. *E. coli* and *Salmonella* strains have also been found to possess similar O-antigen serotypes due to their evolutionary relatedness and horizontal gene transfer [66, 67]. One tailspike protein cluster was found in prophages associated with *E. coli* genomes predicted to have the O23 serotype and *Salmonella* genomes predicted to have the O51 serotype (Fig. 4D), which have been found to have highly similar glycan backbone structures with slightly different side groups (Supplementary Fig. S9A) [68]. Additionally, the *E. coli* serotypes O118 and O151 are similar to the *Salmonella* serotype O47, differing by only the linkages between N-acetyl-beta-D-glucosamine and ribitol sugar residues [69], suggesting that the tailspike proteins observed in the phages associated with these bacteria could interact with both polysaccharide types (Fig. 4E, Supplementary Fig. S9B). Another set of phages was found to be associated with the *Salmonella* O45 serotype and *E. coli* OX13 serotype (Fig. 4F). While the structure for the OX13 surface O-polysaccharide is not known, the corresponding gene loci of the *E. coli* and *Salmonella* genomes were found to be highly similar in gene orders and gene identities, with over 50% identity observed in 13 of 14 gene pairs, suggesting a likely relationship between the polysaccharide structures (Supplementary Fig. S9C).

These cross-species correlations between tailspike proteins and serotypes offer additional proof of the strong links between tailspike proteins and bacterial surface polysaccharides. This dynamic also illustrates the ongoing phage–host arms race: bacteria can evolve their surface receptors through horizontal gene transfer, while phages have the flexibility to modify their receptor-binding proteins to target these newly adapted bacterial hosts. Additionally, the association of tailspike proteins with different serotypes that have structurally similar polysaccharides provides evidence that a limited degree of cross-reactivity between tailspike proteins and polysaccharides may contribute to the increased host range of phages.

Domain swapping in the tailspike proteins

Extensive domain swapping between tailspike protein was observed in all 5 species. Tailspike proteins are commonly classified into an N-terminal head domain and a C-terminal domain (Fig. 5A), with the C-terminal domain encompassing a beta-helix domain that is responsible for both receptor binding and depolymerase activity [20, 70]. Previous studies have reported naturally occurring domain swapping in phages based on the presence of highly similar domains [20, 71]. Leveraging the large dataset of tailspike protein, a comprehensive assessment of domain swapping was performed. To convincingly identify these potential swaps, one of the domains is required to remain within a highly similar cluster (at 95% identity), while the other domain is placed in a distinct cluster, even if its identity threshold is significantly lower (at 30% identity). Evidence of N- and C-terminal domain swaps was observed in all 5 species, with a majority of the observed swaps being in tailspike proteins associated with *E. coli* or *Klebsiella* genomes. This observation can be attributed to the higher abundance of tailspike proteins identified in these 2 species, emphasizing the advantages offered by large datasets.

For the tailspike proteins that displayed putative domain swaps, an additional investigation was carried out to explore the relationship between their associated serotypes and the clustering of their N-terminal and C-terminal domains (Fig. 5B–D, Supplementary Fig. S10). Within the 95% identity cluster, it was observed that all tailspike proteins sharing the same C-terminal cluster were associated with the same serotype. However, this consistent serotype association was not observed across all N-terminal clusters. At the 30% identity level, the C-terminal domain clusters continued to exhibit a notably stronger association with serotypes compared to the N-terminal domain clusters. For example, at the 30% identity level, the C-terminal domain cluster, C_cl30_83, containing 274 vOTUs in *Klebsiella* exclusively corresponded to 2 serotypes, whereas a N-terminal domain cluster, N_cl30_15, with a similar amount of vOTUs (215) in *Klebsiella* at the same identity level originated from a tailspike protein associated with 20 serotypes. Analysis of the serotype cluster entropy associated with each N-terminal and C-terminal domain corroborated these trends, revealing that C-terminal domains were characterized by significantly lower levels of serotype diversity compared to N-terminal domains in all the species (Fig. 5E).

Applying phage–host specificity predictions in phage therapy cases

To demonstrate the predictive power of our data in aiding real-world phage host specificity determination in phage therapy, we compiled a dataset consisting of 173 cases for the 5 pathogens. These cases were extracted from scientific literature and encompassed phage–host infection experiments, as well as preclinical and clinical phage therapy trials for which both phage and bacterial genomic data, or host serotype information, were available (Supplementary Table S4). Within this dataset, tailspike proteins were identified in the phage genomes from 92 cases. Remarkably, 64 of these cases (accounting for 69.57%) exhibited associations consistent with our prediction. When examining just the preclinical and clinical phage therapy studies, 22 of 66 phages contained tailspike proteins. In 16 of these cases (accounting for 72.73%), the bacterial host serotypes were in alignment with our findings. Collectively, these data suggest that the specific association between tailspike proteins and host serotypes can be used to reliably guide the prediction of phage–host specificity in phage therapy applications.

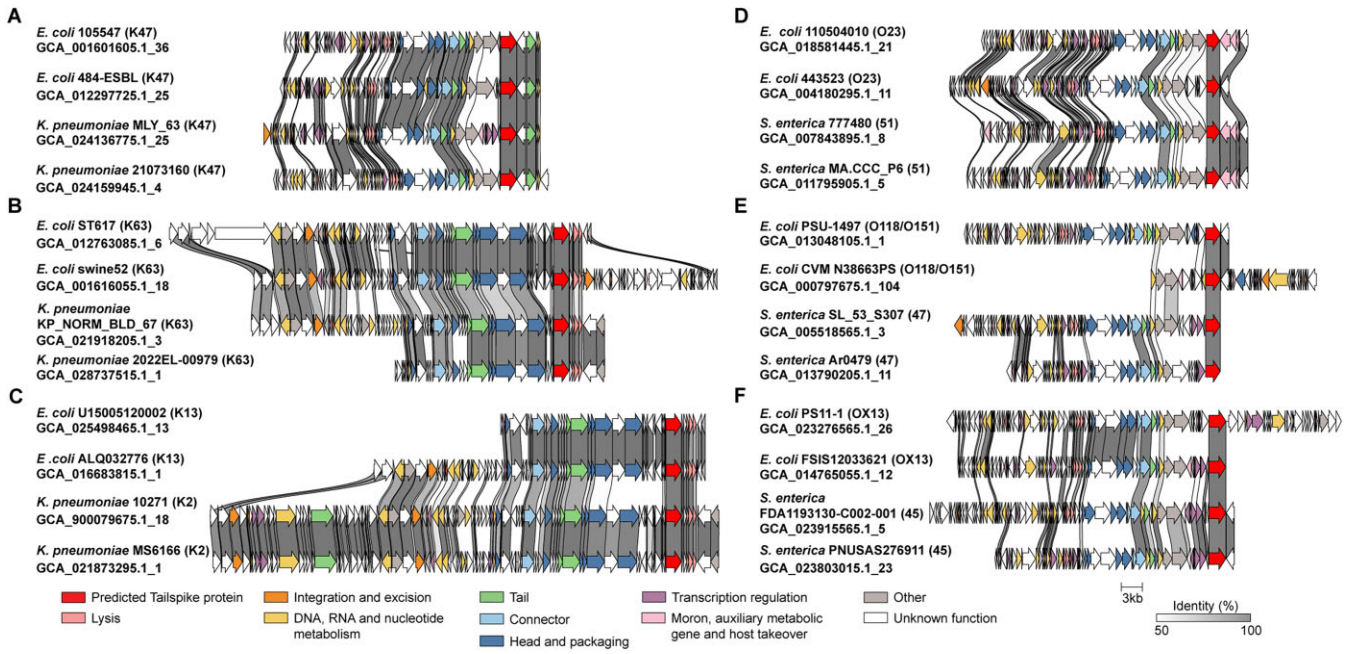

Figure 4: Highly similar tailspike proteins in phages targeting different host species. Prophage genomes derived from different host species with highly similar tailspike proteins are shown for (A–C) *E. coli* and *K. pneumoniae* strains and (E, F) *E. coli* and *S. enterica* strains. Genes are colored based on their annotations using Pharokka with tailspike proteins shown in red, and the amino acid identities of the genes are shown as the shaded regions between genes.

A successful example of phage therapy being used can be seen in the phage AbTP3phi1, which was effective against a multidrug-resistant *A. baumannii* infection with the K116 serotype [72, 73]. When compared to the set of tailspike proteins identified in this study, the tailspike protein from the AbTP3phi1 phage (OL770263) was found to be homologous to 2 clusters of tailspike proteins associated with the *Acinetobacter* K37 (38.0% amino acid identity) and K116 (37.4% amino acid identity) serotypes (Fig. 3D), 2 of the rarer serotypes being predicted for only 2 (0.022%) and 26 (0.282%) of the 9,217 *Acinetobacter* genomes analyzed, respectively. The K116 serotype has been suggested to be a hybrid of genes from the *Acinetobacter* K37 and K14 serotype-specific loci, and the polysaccharide structures were found to have similar structures [74]. This result demonstrates the predictive capacity of tailspike protein–serotype associations, which stems from the comprehensive and large-scale analysis of the dataset. The knowledge about tailspike proteins can subsequently be employed to evaluate the viability of specifically targeting a phage toward a particular serotype, offering a resource to optimize the search process in phage therapy and rational phage engineering.

Another interesting example lies in the phage–host range experiment for bacteriophage CBA120, which was reported to possess 4 tailspike proteins [16]. CBA120 infects *S. enterica* serovar Minnesota (also known as the *Salmonella* O21-antigen) due to the function of tailspike protein (TSP) 1. Meanwhile, TSP2, TSP3, and TSP4 hydrolyze the O157, O77, and O78 *E. coli* O-antigens, respectively. Notably, all 4 of these tailspike proteins were accurately identified by SpikeHunter, and the top predicted serotypes for each protein matched the experimental findings. The percentages of the *E. coli* O157-, O77-, and O78-antigens and the *Salmonella* O21-antigen in our database are relatively low, 0.725%, 0.767%, 2.124%, and 0.403%, respectively, indicating that the accurate identification of these serotypes is unlikely to be due to random chance. This underscores the significance of our study in streamlining the screening process for effective phages in therapeutic applications.

Discussion

Through this large-scale, multispecies genomic analysis, we have demonstrated the strong association of tailspike proteins and bacterial surface polysaccharide receptors at the strain level. By applying protein language models, we have developed a sensitive approach for the detection of tailspike proteins, something that has been challenging due to the sequence diversity seen in these proteins. Our use of prophage-encoded tailspike proteins provides a way to examine confirmed tailspike–serotype associations across thousands of genomes without the need for isolating phages and bacterial strains. These factors have resulted in a rich dataset that captures phage and serotype diversity and can serve as a resource for future phage research.

Phage tailspike proteins are crucial in determining phage–host range and in degrading pathogen polysaccharides, as they serve dual functions as both phage receptor-binding proteins and phage-encoded depolymerases. Despite their significance, there has been a lack of computational tools specifically tailored for the effective and precise identification of tailspike proteins. Existing computational methods and pipelines, such as PhageRBPdetect [71], PhageHostLearn [75], and BacteriophageHostPrediction [12], are designed to recognize phage receptor-binding proteins. Similarly, tools like DePP [76] and PhageDPO [77] aim to identify phage-encoded depolymerases in phages. However, these tools fail to differentiate tailspike proteins from the other phage-encoded depolymerases and receptor-binding proteins that they detect. Traditionally, the detection of tailspike proteins has not posed a significant challenge, primarily because they possess this unique structural feature of the single-stranded beta-helix domain. However, this identification process has largely been manual, involving analyses of gene annotations, genomic context, and predicted structures. This labor-intensive approach has restricted its application to small-scale studies and to specific bacterial species [9, 10, 36, 78]. Consequently, SpikeHunter is a pioneering tool, being

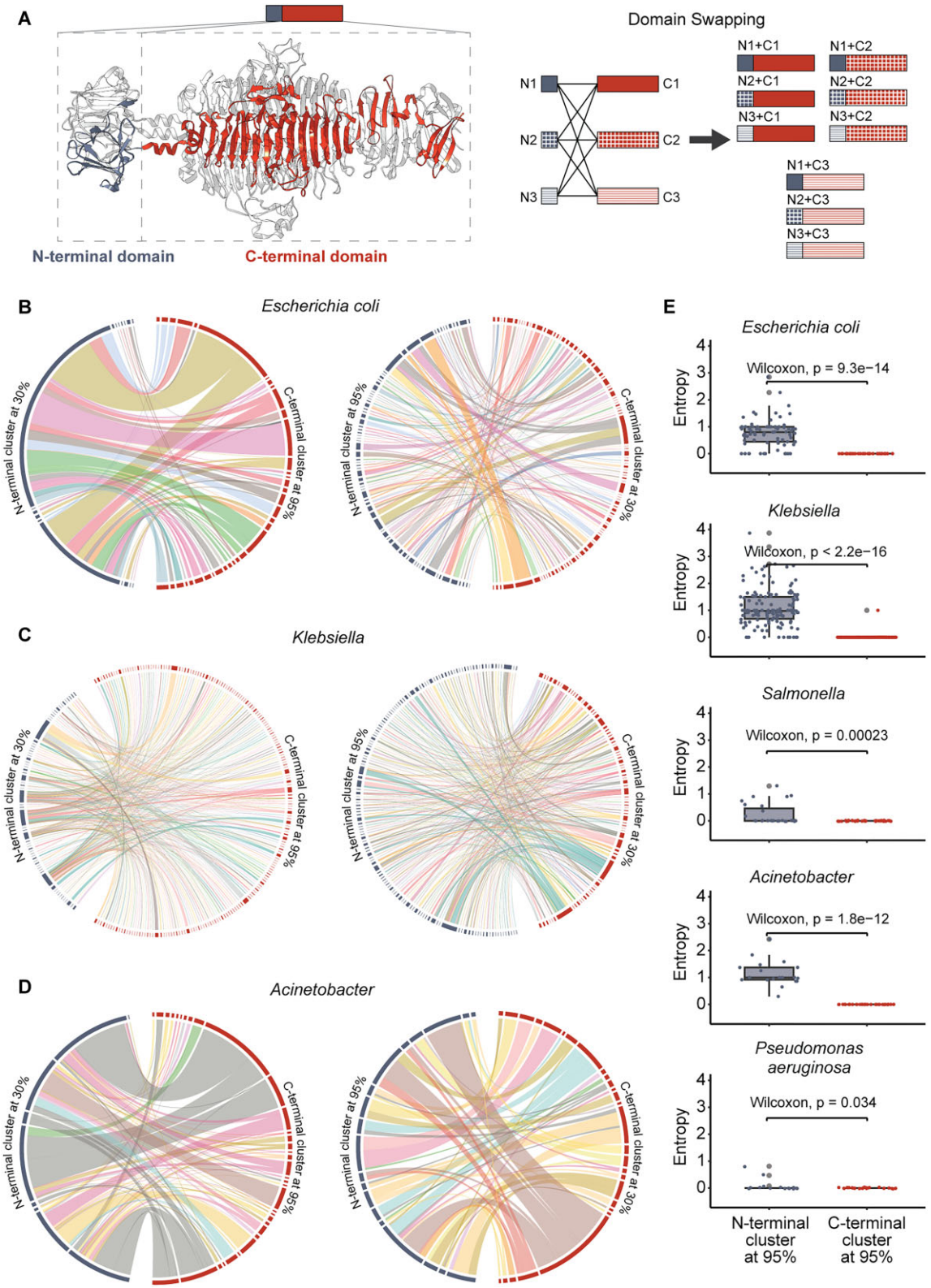

Figure 5: Domain swapping in tailspike proteins. (A) Example domain partitioning of tailspike protein (PDB: 2XC1) and diagram illustrating domain swapping between hypothetical tailspike proteins. Putative domain swapping in tailspike proteins from (B) *E. coli*, (C) *Klebsiella*, and (D) *Acinetobacter* are shown as connections between the N-terminal and C-terminal domains clustered at different amino acid identities. Bands connecting N- and C-terminal domains are colored based on the serotype of the associated bacterial genomes. (E) Boxplots comparing the entropy of serotypes observed in bacterial genomes with N- and C-terminal domains clustered at 95% from proteins with potential domain swapping. Significance of N- and C-terminal entropy difference based on a Wilcoxon rank-sum test is shown for each plot.

the first computational tool specifically developed for the rapid and direct identification of tailspike proteins. By focusing on this critical subset of phage-encoded proteins, SpikeHunter fills a significant gap in the field of phage–host interaction research, offering a novel solution for large-scale studies.

Phage therapy has been proposed as an ideal tool for combating antibiotic resistance [13] and suppressing disease-related members of the microbiome [79, 80], but the application of phage therapy has consistently been limited by the ability to identify phages that would be effective against specific bacterial infections [13]. The predictive power shown by our data for 70% of the phage therapy–related trials, especially the tailspike protein specific to *Acinetobacter* K116 in the phage AbTP3phi1 that was used in a successful phage therapy application [72], as well as for the multiple cross-species serotype-specific tailspike proteins in the *E. coli* phage CBA120 [16], demonstrates the utility of these data in guiding the phage screening or phage protein selection in the future. Our results also demonstrate that domain swapping is common in tailspike proteins, and the modular nature of tailspike proteins has allowed them to be engineered to alter their specificity [81]. Because isolating effective natural phages can be time-consuming, engineering targeted phages could be an effective strategy for combating new antimicrobial-resistant bacteria. This process and the application of tailspike proteins as antimicrobial compounds can be significantly enhanced by these results, leading to more efficient and effective biomedical applications of phages.

Tailspike proteins also have a promising future as molecular tools in the glycosciences. The study of bacteriophages and phage–host interactions has led to the discovery of specialized polymerases [82], ligases [82], restriction enzymes [83], and the CRISPR-Cas9 [84] system, which has revolutionized molecular biology. The polysaccharide-targeting nature of tailspike proteins makes them ideal candidates as new tools to study glycans. Glycan profiling is inherently difficult due to the structural complexity, heterogeneity, lack of templates, and the limited availability of high-throughput analytical methods [85, 86]. The depolymerase activity exhibited by tailspike proteins presents the potential for their utilization analogous to restriction enzymes in DNA digestion, facilitating the breakdown of polysaccharides into smaller fragments amenable to analysis, such as linkage analysis. Moreover, tailspike proteins have demonstrated utility in the development of pathogen biosensors reliant on the recognition of distinct glycans [11, 87, 88], thus suggesting the possibility of expanding their functionality to establish novel glycan typing microarrays.

This approach to identifying tailspike protein associations has some disadvantages that highlight persistent gaps in the field. First, serotyping prediction tools are primarily focused on common human pathogens, limiting these results to these species of interest and underscoring the need for advanced serotype prediction tools that can be generalized to other species. Additionally, various other types of polysaccharides that were not considered in this analysis are presented on the cell surface and have been shown to be receptors for some phages, including enterobacterial common antigen [9] and cellulose [89]. Non-tailspike phage receptor-binding proteins, which also play roles in binding to bacterial cell surface sugars and proteins [90], are important considerations for future studies investigating phage–host interactions and likely explain the phage therapy cases that our associations did not match. Finally, it is important to note that most clinical bacterial samples and phages used in phage therapy have not been genetically sequenced. This lack of data complicates the identification of the organisms involved and limits our understanding of the genetic factors that determine phage–host interactions. These limitations highlight gaps in the understanding of both bacterial serotypes and phage biology that will be important aspects of future research. Expanding this work to include free phages and to account for other bacterial species will enhance the associations and provide important context to the results and facilitate their application in biomedical contexts. Overall, this study provides an essential foundation for the future study of bacteriophage host specificity and the future use of phages and tailspike proteins in a variety of fields.

Additional Files

Table S1. Tailspike protein clusters at different identities.
Table S2. Tailspike protein cluster to serotype associations.
Table S3. Cross-species tailspike protein associations.
Table S4. Phage host specificity prediction in pre-clinical and clinical cases in phage therapy.
Figure S1: SpikeHunter validation performance metrics per epoch.
Figure S2: ϕKp24-like jumbo prophage.
Figure S3: *E. coli* phages with 24 serotypes. Diagram of 24 similar *E. coli* from the same vOTU that have distinct tailspike proteins and are associated with distinct serotypes.
Figure S4: High serotype diversity within a tailspike protein cluster.
Figure S5: Purity of serotypes associated with 60% identity tailspike protein clusters.
Figure S6: Tailspike protein serotype networks for five species, including serotype and tailspike protein cluster labels for nodes.
Figure S7: Tailspike protein to serotype networks for *Acinetobacter* and *P. aeruginosa*.
Figure S8: Example phylogenetic distributions of tailspike proteins and serotypes.
Figure S9: Glycan structures and gene clusters in *E. coli* and *Salmonella*.
Figure S10: *Acinetobacter* and *P. aeruginosa* domain swapping.

Abbreviations

MCC: Matthew's correlation coefficient; OC: outer core; PRAUC: area under the precision–recall curve; TSP: tailspike protein.

Availability of Supporting Source Code and Requirements

Project name: SpikeHunter: A Deep Learning Tool for Identifying Phage Tailspike Proteins

Project homepage: https://github.com/nlm-irp-jianglab/SpikeHunter (SpikeHunter model and code)

Operating system(s): Linux or other Unix-like operation systems

Programming language: Python

Other requirements: See https://github.com/nlm-irp-jianglab/SpikeHunter/blob/main/environment.yml for details.

License: MIT

RRID:SCR_024831

biotools Id: spikehunter

Snapshots of our code are also archived in Software Heritage [23–25].

Acknowledgments

This work utilized the computational resources of the NIH HPC Biowulf cluster [91]. We thank Dr. Audrey Burnim at NLM/NIH for help with visualizing protein structures and Hui Yi from CenHTRO at the University of Georgia for her assistance with the statistical approaches used in the study.

Authors' Contributions

Conceptualization and supervision: X.J. Data curation and formal analysis: Y.Y., K.D., W.Y., and X.J. Methodology: Y.Y., T.C., L.X., and X.J. Software: Y.Y. and X.J. Visualization: Y.Y. Writing—original draft: Y.Y., K.D., and X.J. Writing—review & editing: Y.Y., K.D., and X.J.

Funding

Y.Y., K.D., W.Y., and X.J. are supported by the Intramural Research Program of the NIH, National Library of Medicine. T.C. and L.X. are supported by the National Science Foundation (NSF2226183 to L.X.).

Data Availability

The data underlying this article are available in the article and in its online supplementary material. The tailspike protein IDs, along with their corresponding clusters at various protein identities, are provided at https://github.com/nlm-irp-jianglab/TSP_paper/blob/main/data/TSP_ids_and_clusters.txt [23]. The tailspike protein clusters identified in our research have been compiled into a database, which can be accessed at [24]. An archival copy of the code and supporting data, also including DOME-ML annotations, are available via the *GigaScience* database, GigaDB [43].

Competing Interests

The authors declare that there are no competing interests.

References

1. Eales BM, Tam VH. Case commentary: novel therapy for multidrug-resistant Acinetobacter baumannii infection. Antimicrob Agents Chemother 2022;66:e0196–21. https://doi.org/10.1128/AAC.01996-21.
2. Khatami A, Lin RCY, Petrovic-Fabijan A, et al. Bacterial lysis, autophagy and innate immune responses during adjunctive phage therapy in a child. EMBO Mol Med 2021;13:e13936. https://doi.org/10.15252/emmm.202113936.
3. Doub JB, Ng VY, Johnson AJ, et al. Salvage bacteriophage therapy for a chronic MRSA prosthetic joint infection. Antibiotics 2020;9:241. https://doi.org/10.3390/antibiotics9050241.
4. Gainey AB, Daniels R, Burch A-K, et al. Recurrent ESBL Escherichia coli urosepsis in a pediatric renal transplant patient treated with antibiotics and bacteriophage therapy. Pediatr Infect Dis J 2023;42:43–46. https://doi.org/10.1097/INF.0000000000003735.
5. Loc-Carrillo C, Abedon ST. Pros and cons of phage therapy. Bacteriophage 2011;1:111–14. https://doi.org/10.4161/bact.1.2.14590.
6. Ross A, Ward S, Hyman P. More is better: selecting for broad host range bacteriophages. Front Microbiol 2016;7:217131. https://doi.org/10.3389/fmicb.2016.01352.
7. Gordillo Altamirano FL, Barr JJ. Unlocking the next generation of phage therapy: the key is in the receptors. Curr Opin Biotechnol 2021;68:115–23. https://doi.org/10.1016/j.copbio.2020.10.002.
8. Maffei E, Shaidullina A, Burkolter M, et al. Systematic exploration of Escherichia coli phage–host interactions with the BASEL phage collection. PLoS Biol 2021;19:e3001424. https://doi.org/10.1371/journal.pbio.3001424.
9. Beamud B, García-González N, Gómez-Ortega M, et al. Genetic determinants of host tropism in Klebsiella phages. Cell Rep 2023;42:112048. https://doi.org/10.1016/j.celrep.2023.112048.
10. Pas C, Latka A, Fieseler L, et al. Phage tailspike modularity and horizontal gene transfer reveals specificity towards E. coli O-antigen serogroups. Virol J 2023;20:174. https://doi.org/10.1186/s12985-023-02138-4.
11. Klumpp J, Dunne M, Loessner MJ. A perfect fit: bacteriophage receptor-binding proteins for diagnostic and therapeutic applications. Curr Opin Microbiol 2023;71:102240. https://doi.org/10.1016/j.mib.2022.102240.
12. Boeckaerts D, Stock M, Criel B, et al. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Sci Rep 2021;11:1467. https://doi.org/10.1038/s41598-021-81063-4.
13. Pires DP, Costa AR, Pinto G, et al. Current challenges and future opportunities of phage therapy. FEMS Microbiol Rev 2020;44:684–700. https://doi.org/10.1093/femsre/fuaa017.
14. Timoshina OY, Kasimova AA, Shneider MM, et al. Friunavirus phage-encoded depolymerases specific to different capsular types of Acinetobacter baumannii. Int J Mol Sci 2023;24:9100. https://doi.org/10.3390/ijms24109100.
15. Gencay YE, Gambino M, Prüssing TF, et al. The genera of bacteriophages and their receptors are the major determinants of host range. Environ Microbiol 2019;21:2095–111. https://doi.org/10.1111/1462-2920.14597.
16. Plattner M, Shneider MM, Arbatsky NP, et al. Structure and function of the branched receptor-binding complex of bacteriophage CBA120. J Mol Biol 2019;431:3718–39. https://doi.org/10.1016/j.jmb.2019.07.022.
17. Knecht LE, Veljkovic M, Fieseler L. Diversity and function of phage encoded depolymerases. Front Microbiol 2020;10:2949. https://doi.org/10.3389/fmicb.2019.02949.
18. Oliveira H, Pinto G, Mendes B, et al. A tailspike with exopolysaccharide depolymerase activity from a new Providencia stuartii phage makes multidrug-resistant bacteria susceptible to serum-mediated killing. Appl Environ Microb 2020;86:e00073–20. https://doi.org/10.1128/AEM.00073-20.
19. Hughes KA, Sutherland IW, Jones MV. Biofilm susceptibility to bacteriophage attack: the role of phage-borne polysaccharide depolymerase. Microbiology 1998;144:3039–47. https://doi.org/10.1099/00221287-144-11-3039.
20. Sørensen AN, Woudstra C, Sørensen MCH, et al. Subtypes of tail spike proteins predicts the host range of Ackermannviridae phages. Comput Struct Biotechnol J 2021;19:4854–67. https://doi.org/10.1016/j.csbj.2021.08.030.
21. Flemming H-C. The perfect slime. Colloids Surf B 2011;86:251–59. https://doi.org/10.1016/j.colsurfb.2011.04.025.
22. Yehl K, Lemire S, Yang AC, et al. Engineering phage host-range and suppressing bacterial resistance through phage tail fiber mutagenesis. Cell 2019;179:459–69.e9. https://doi.org/10.1016/j.cell.2019.09.015.
23. Yang Y, Dufault-Thompson K, Yan W, et al. TSP_paper (Version 1) [Computer software]. Software Heritage; 2024. https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://github.com/nlm-irp-jianglab/TSP_paper. Accessed 23 Jan 2024.

24. Yang Y, Dufault-Thompson K, Yan W, et al. TailSpikeDB (Version 1) [Computer software]. Software Heritage; 2024. https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://github.com/nlm-irp-jianglab/tailspikedb. Accessed 10 Oct 2023.
25. Yang Y, Dufault-Thompson K, Yan W, et al. SpikeHunter (Version 1) [Computer software]. Software Heritage; 2024. https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://github.com/nlm-irp-jianglab/SpikeHunter. Accessed 23 Jan 2024.
26. Cook R, Brown N, Redgwell T, et al. INfrastructure for a PHAge REference database: identification of large-scale biases in the current collection of cultured phage genomes. PHAGE 2021;2:214–23. https://doi.org/10.1089/phage.2021.0007.
27. Murzin AG, Brenner SE, Hubbard T, et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–40. https://doi.org/10.1016/S0022-2836(05)80134-2.
28. Terzian P, Olo Ndela E, Galiez C, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform 2021;3:lqab067. https://doi.org/10.1093/nargab/lqab067.
29. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res 2017;45:D491–98. https://doi.org/10.1093/nar/gkw975.
30. Moreno-Gallego JL, Reyes A. Informative regions In viral genomes. Viruses 2021;13:1164. https://doi.org/10.3390/v13061164.
31. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47:D309–14. https://doi.org/10.1093/nar/gky1085.
32. VOGDB—Virus Orthology Groups. https://vogdb.org. Accessed 24 March 2024.
33. Li W, Jaroszewski L, Godzik A. Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 2001;17:282–283. https://doi.org/10.1093/bioinformatics/17.3.282.
34. Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28:3150–52. https://doi.org/10.1093/bioinformatics/bts565.
35. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–89. https://doi.org/10.1038/s41586-021-03819-2.
36. Gaborieau B, Vaysset H, Tesson F, et al. Predicting phage-bacteria interactions at the strain level from genomes. bioRxiv 2023; doi: 10.1101/2023.11.22.567924.
37. Paszke A, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. Adv Neu Inf Process Syst. 2019;32:8024–35.
38. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30. https://doi.org/10.1126/science.ade2574.
39. Verkuil R, Kabeli O, Du Y, et al. Language models generalize beyond natural proteins. bioRxiv 2022; doi: 10.1101/2022.12.21.521521.
40. Heinzinger M, Elnaggar A, Wang Y, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinf 2019;20:723. https://doi.org/10.1186/s12859-019-3220-8.
41. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–302. https://doi.org/10.1128/AEM.00341-17.
42. NCBI Pathogen Detection database. https://www.ncbi.nlm.nih.gov/pathogens. Accessed 2 Apr 2023.
43. Yang Y, Dufault-Thompson K, Yan W, et al. Supporting data for "Large-Scale Genomic Survey with Deep Learning–Based Method Reveals Strain-Level Phage Specificity Determinants." *GigaScience Database*. 2024. http://dx.doi.org/10.5524/102504.
44. Bessonov K, Laing C, Robertson J, et al. ECTyper: in silico Escherichia coli serotype and species prediction from raw and assembled whole-genome sequence data. Microb Genom 2021;7:000728. https://doi.org/10.1099/mgen.0.000728.
45. Joensen KG, Tetzschner AMM, Iguchi A, et al. Rapid and easy *in silico* serotyping of Escherichia coli isolates by use of whole-genome sequencing data. J Clin Microbiol 2015;53:2410–26. https://doi.org/10.1128/JCM.00008-15.
46. Holt KE, Lassalle F, Wyres KL, et al. Diversity and evolution of surface polysaccharide synthesis loci in enterobacteriales. ISME J 2020;14:1713–30. https://doi.org/10.1038/s41396-020-0628-0.
47. Iguchi A, Iyoda S, Kikuchi T, et al. A complete view of the genetic diversity of the Escherichia coli O-antigen biosynthesis gene cluster. DNA Res 2015;22:101–7. https://doi.org/10.1093/dnares/dsu043.
48. Lam MMC, Wick RR, Judd LM, et al. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the Klebsiella pneumoniae species complex. Microb Genom 2022;8:000800. https://doi.org/10.1099/mgen.0.000800.
49. Thrane SW, Taylor VL, Lund O, et al. Application of whole-genome sequencing data for O-specific antigen analysis and in silico serotyping of Pseudomonas aeruginosa isolates. J Clin Microbiol 2016;54:1782–88. https://doi.org/10.1128/JCM.00349-16.
50. Zhang S, Den Bakker HC, Li S, et al. SeqSero2: rapid and improved Salmonella serotype determination using whole-genome sequencing data. Appl Environ Microb 2019;85:e01746–19 https://doi.org/10.1128/AEM.01746-19.
51. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
52. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–59. https://doi.org/10.1093/bioinformatics/btl158.
53. Gu Z, Gu L, Eils R, et al. circlize implements and enhances circular visualization in R. Bioinformatics 2014;30:2811–12. https://doi.org/10.1093/bioinformatics/btu393.
54. Phage Receptor Database (PhReD). https://portal.bio-conversion.org/home/phred. Accessed 24 March 2024.
55. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. FEMS Microbiol Lett 2016;363:fnw002. https://doi.org/10.1093/femsle/fnw002.
56. Institut Pasteur: Viral Host Range database. https://viralhostrangedb.pasteur.cloud/. Accessed 24 March 2024.
57. Lamy-Besnier Q, Brancotte B, Ménager H, et al. Viral Host Range database, an online tool for recording, analyzing and disseminating virus–host interactions. Bioinformatics 2021;37:2798–801. https://doi.org/10.1093/bioinformatics/btab070.
58. Gómez-Ochoa SA, Pitton M, Valente LG, et al. Efficacy of phage therapy in preclinical models of bacterial infection: a systematic review and meta-analysis. Lancet Microbe 2022;3:e956–68. https://doi.org/10.1016/S2666-5247(22)00288-9.
59. Adaptive Phage Therapeutics Case Studies. https://aphage.com/science/case-studies/. Accessed 20 Aug 2023.

60. Green SI, Clark JR, Santos HH, et al. A retrospective, observational study of 12 cases of expanded-access customized phage therapy: production, characteristics, and clinical outcomes. Clin Infect Dis 2023;77:1079–91. https://doi.org/10.1093/cid/ciad335.
61. Ouyang R, Costa AR, Cassidy CK, et al. High-resolution reconstruction of a jumbo-bacteriophage infecting capsulated bacteria using hyperbranched tail fibers. Nat Commun 2022;13:7241. https://doi.org/10.1038/s41467-022-34972-5.
62. Clark CG, Kropinski AM, Parolis H, et al. Escherichia coli O123 O antigen genes and polysaccharide structure are conserved in some Salmonella enterica serogroups. J Med Microbiol 2009;58:884–94. https://doi.org/10.1099/jmm.0.007187-0.
63. Pan Y-J, Lin T-L, Chen C-T, et al. Genetic analysis of capsular polysaccharide synthesis gene clusters in 79 capsular types of Klebsiella spp. Sci Rep 2015;5:15573. https://doi.org/10.1038/srep15573.
64. Pieroni P, Rennie RP, Ziola B, et al. The use of bacteriophages to differentiate serologically cross-reactive isolates of Klebsiella pneumoniae. J Med Microbiol 1994;41:423–29. https://doi.org/10.1099/00222615-41-6-423.
65. Nanayakkara BS, O'Brien CL, Gordon DM. Diversity and distribution of Klebsiella capsules in Escherichia coli. Environ Microbiol Rep 2019;11:107–17. https://doi.org/10.1111/1758-2229.12710.
66. Liu B, Perepelov AV, Li D, et al. Structure of the O-antigen of Salmonella O66 and the genetic basis for similarity and differences between the closely related O-antigens of Escherichia coli O166 and Salmonella O66. Microbiology 2010;156:1642–49. https://doi.org/10.1099/mic.0.037325-0.
67. Wang L, Reeves PR. The *Escherichia coli* O111 and *Salmonella enterica* O35 gene clusters: gene clusters encoding the same colitose-containing O antigen are highly conserved. J Bacteriol 2000;182:5256–61. https://doi.org/10.1128/JB.182.18.5256-5261.2000.
68. Liu B, Knirel YA, Feng L, et al. Structural diversity in Salmonella O antigens and its genetic basis. FEMS Microbiol Rev 2014;38:56–89. https://doi.org/10.1111/1574-6976.12034.
69. MacLean LL, Liu Y, Vinogradov E, et al. The structural characterization of the O-polysaccharide antigen of the lipopolysaccharide of Escherichia coli serotype O118 and its relation to the O-antigens of Escherichia coli O151 and Salmonella enterica O47. Carbohydr Res 2010;345:2664–69. https://doi.org/10.1016/j.carres.2010.10.004.
70. Gage MJ, Robinson AS. C-terminal hydrophobic interactions play a critical role in oligomeric assembly of the P22 tailspike trimer. Protein Sci 2003;12:2732–47. https://doi.org/10.1110/ps.0315030 3.
71. Boeckaerts D, Stock M, De Baets B, et al. Identification of phage receptor-binding protein sequences with hidden Markov models and an extreme gradient boosting classifier. Viruses 2022;14:1329. https://doi.org/10.3390/v14061329.
72. Liu M, Hernandez-Morales A, Clark J, et al. Comparative genomics of Acinetobacter baumannii and therapeutic bacteriophages from a patient undergoing phage therapy. Nat Commun 2022;13:3776. https://doi.org/10.1038/s41467-022-31455-5.
73. 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. Antimicrob Agents Chemother 2017;61:e00954–17. https://doi.org/10.1128/AAC.00954-17.
74. Shashkov AS, Cahill SM, Arbatsky NP, et al. Acinetobacter baumannii K116 capsular polysaccharide structure is a hybrid of the K14 and revised K37 structures. Carbohydr Res 2019;484:107774. https://doi.org/10.1016/j.carres.2019.107774.
75. Briers Y, Boeckaerts D, Stock M, et al. Actionable prediction of Klebsiella phage-host specificity at the subspecies level. Research Square 2023; doi: 10.21203/rs.3.rs-3101607/v1.
76. Magill DJ, Skvortsov TA. DePolymerase Predictor (DePP): a machine learning tool for the targeted identification of phage depolymerases. BMC Bioinf 2023;24:208. https://doi.org/10.1186/s12859-023-05341-w.
77. Vieira M, Duarte J, Domingues R, et al. PhageDPO : phage depolymerase finder. bioRxiv 2023; doi: 10.1101/2023.02.24.529883.
78. Latka A, Leiman PG, Drulis-Kawa Z, et al. Modeling the architecture of depolymerase-containing receptor binding proteins in Klebsiella phages. Front Microbiol 2019;10:2649. https://doi.org/10.3389/fmicb.2019.02649.
79. Gan L, Feng Y, Du B, et al. Bacteriophage targeting microbiota alleviates non-alcoholic fatty liver disease induced by high alcohol-producing Klebsiella pneumoniae. Nat Commun 2023;14:3215. https://doi.org/10.1038/s41467-023-39028-w.
80. Federici S, Kredo-Russo S, Valdés-Mas R, et al. Targeted suppression of human IBD-associated gut microbiota commensals by phage consortia for treatment of intestinal inflammation. Cell 2022;185:2879–98.e24. https://doi.org/10.1016/j.cell.2022.07.003.
81. Gil J, Paulson J, Brown M, et al. Tailoring the host range of Ackermannviridae bacteriophages through chimeric Tailspike proteins. Viruses 2023;15:286. https://doi.org/10.3390/v15020286.
82. Abril AG, Carrera M, Notario V, et al. The use of bacteriophages in biotechnology and recent insights into. Antibiotics 2022;11:653. https://doi.org/10.3390/antibiotics11050653.
83. Loenen WAM, Dryden DTF, Raleigh EA, et al. Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Res 2014;42:3–19. https://doi.org/10.1093/nar/gkt990.
84. Ran FA, Hsu PD, Wright J, et al. Genome engineering using the CRISPR-Cas9 system. Nat Protoc 2013;8:2281–308. https://doi.org/10.1038/nprot.2013.143.
85. Wells L, Hart GW. Glycomics: building upon proteomics to advance glycosciences. Mol Cell Proteomics 2013;12:833–5. https://doi.org/10.1074/mcp.E113.027904.
86. Gray CJ, Migas LG, Barran PE, et al. Advancing solutions to the carbohydrate sequencing challenge. J Am Chem Soc 2019;141:14463–79. https://doi.org/10.1021/jacs.9b06406.
87. Singh A, Arya SK, Glass N, et al. Bacteriophage tailspike proteins as molecular probes for sensitive and selective bacterial detection. Biosens Bioelectron 2010;26:131–38. https://doi.org/10.1016/j.bios.2010.05.024.
88. Born Y, Fieseler L, Thöny V, et al. Engineering of bacteriophages Y2:: dpoL1-C and Y2:: luxAB for Efficient control and rapid detection of the fire blight pathogen, Erwinia amylovora. Appl Environ Microb 2017;83:e00341–17. https://doi.org/10.1128/AEM.00341-17.
89. Knecht LE, Heinrich N, Born Y, et al. Bacteriophage S6 requires bacterial cellulose for Erwinia amylovora infection. Environ Microbiol 2022;24:3436–50. https://doi.org/10.1111/1462-2920.15973.
90. German GJ, Misra R. The TolC protein of Escherichia coli serves as a cell-surface receptor for the newly characterized TLS bacteriophage. J Mol Biol 2001;308:579–85. https://doi.org/10.1006/jmbi.2001.4578.
91. NIH HPC Biowulf cluster. http://hpc.nih.gov. Accessed 24 March 2024.

Received: September 28, 2023. **Revised:** January 23, 2024. **Accepted:** March 24, 2024