PhageScope: a well-annotated bacteriophage database with automatic analyses and visualizations

Ruo Han Wang 1,2, Shuo Yang 1,2, Zhixuan Liu 1, Yuanzheng Zhang 1, Xueying Wang 1,2, Zixin Xu 1, Jianping Wang 1,* and Shuai Cheng Li 1,*

1 Department of Computer Science, City University of Hong Kong, Hong Kong
2 City University of Hong Kong (Dongguan), Dongguan, China

To whom correspondence should be addressed. Tel: +852 34427737; Fax: +852 34420503; Email: jianwang@cityu.edu.hk
Correspondence may also be addressed to Shuai Cheng Li. Tel: +852 34429412; Fax: +852 34420503; Email: shuaicli@cityu.edu.hk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Bacteriophages are viruses that infect bacteria or archaea. Understanding the diverse and intricate genomic architectures of phages is essential to study microbial ecosystems and develop phage therapy strategies. However, the existing phage databases are short of meticulous annotations. To this end, we propose PhageScope (https://phagescope.deepomics.org), an online phage database with comprehensive annotations. PhageScope harbors a collection of 873,718 phage sequences from various sources. Applying fifteen state-of-the-art tools to perform systematic annotations and analyses, PhageScope provides annotations on genome completeness, host range, lifestyle information, taxonomy classification, nine types of structural and functional genetic elements, and three types of comparative genomic studies for curated phages. Additionally, PhageScope incorporates automatic analyses and visualizations for curated and customized phages, serving as an efficient platform for phage study.

Graphical abstract

Introduction

Viruses infecting bacteria or archaea, i.e. bacteriophages or phages, are the most abundant and diverse biological entities on Earth (1). Phages play essential roles in maintaining species diversity and driving bacterial co-evolution. Given the threat of multi-drug resistance, phage therapy, which uses phages to treat bacterial infections, is considered an alternative to traditional antibiotic therapy. In-depth investigation of microbial systems and effective exploration of phages as therapeutic agents rely on meticulous studies on phage genomes.

Exhaustive investigations of phage genomes rely on extensive collections and diligent annotations of phage sequences. In recent years, the accumulation of next-generation sequencing (NGS) data and the development of phage detection methods have facilitated the computational extraction of numerous phage sequences from bacterial and metagenomic NGS data (2–8). However, these phage sequences derived from NGS data often lack accurate and detailed annotations, since phage annotation workflows, which require manually curated references and well-designed pipelines, are tedious and laborious.
A systematic workflow for phage annotation should encompass completeness assessment, phenotype and taxonomy determination, structural and functional annotations, as well as comparative genomic studies (9). First, the completeness assessment reports the quality of the assembled phage genomes. Then, the phenotype and taxonomy determination characterize the observable traits, such as morphology, lifestyles and host ranges. Furthermore, structural and functional annotations identify the genetic features and functional elements of phage genomes. Last, comparative genomic studies provide insight into evolutionary relationships, genetic diversity and functional variations for multiple phage genomes.

The existing bacteriophage databases and web servers do not offer a comprehensive provision of the aforementioned information and functionality. PhagesDB (10) stores actinobacteriophage sequences with phenotype annotated, but lack genome annotations. MVP (11) provides phage-host interaction information, with other information unavailable. PHROG (12) stores annotated phage protein families, but lacks other functional elements and phage information. PhANNs (13) and PhaGAA (14) are phage annotation web servers. However, they simply establish partial workflows and lack curated data.

To fill the gap, we propose PhageScope (https://phagescope.deeomics.org), a bacteriophage database with comprehensive annotations. PhageScope stores 873,718 phage sequences from multiple public repositories and published datasets. According to the workflows described above, we have applied fifteen state-of-the-art tools to provide annotations and analyses for curated phages, encompassing genome completeness, host range, lifestyle information, taxonomy classification and genetic element annotations, such as open reading frames (ORFs) and proteins, transcriptional terminators, tRNA and tmRNA genes, Anti-CRISPR proteins, CRISPR arrays, virulent factors, antimicrobial resistance genes and transmembrane proteins. Comparative genomic studies among multiple sequences, including genome clustering, sequence alignment and comparative tree construction, are also available. In addition, to streamline the workflow for users to analyze their phage genomes, PhageScope also supports automatic analyses and interactive visualizations for curated and customized phages.

Materials and methods
Phage sequence collection
We first searched for phage sequences across multiple public repositories, including RefSeq (15), Genbank (16), EMBL (17) and DDBJ (18), with ‘phage’, ‘bacteriophage’ and the bacterial names from NCBI taxonomy database (19) along with ‘virus’ as keywords. Furthermore, we incorporated phages from various published datasets, including PhagesDB (10), GOV2 (2), GVD (3) GPD (4), MGV (5), CHVD (6), STV (20), IGVD (21), IMG/VR (8), as well as 66,823 phage sequences from TemPhD, mined with our temperate phage detection method (7). As a result, we collected a dataset comprising 873,718 phage sequences (Supplementary Table S1). Subsequently, we applied multiple analysis tools to provide exhaustive genome annotations and sequence comparison information for the phages (Supplementary Table S2).

Genome annotation
Completeness assessment
We applied CheckV v0.8.1 (22) to assess the completeness of the phage genomes. CheckV categorized each genome into four distinct quality tiers, which are complete, high-quality, medium-quality and low-quality.

Phenotype annotation
We determined the phenotypic characteristics of the phages with respect to the host range and lifestyle. Regarding phage host assignment, the host information of 530,085 phages was available from the data source through a systematic search, which serves as the reference. To annotate the remaining phages, we incorporated homology search and DeepHost (23) to infer host taxonomies (Supplementary Methods S1.1). Regarding lifestyle prediction, the phages from TemPhD dataset are temperate phages according to their phage mining method. For the remaining phages, we utilized Graphpe (24) to discern between virulent and temperate phages.

Structural annotation
We identified structural components, including coding regions and transcriptional terminators within phage genomes. First, we adopted Prodigal v2.6.3 (25) with a meta option to identify the ORFs on the phage genomes. Then we employed EggnoMapper v2.1.10 (26) to conduct orthology assignments and transfer annotations from the assigned ortholog groups for the resulting coding sequences. To refine the annotation, for proteins lacked hits, we iteratively applied mmseqs (27) to detect homology from the PHROG database (12), adopting a threshold of e-value <1e-5 with a sensitive mode and then annotated the proteins (Supplementary Methods S1.2 and Supplementary Figure S1). We further categorized the annotated proteins into ten functional classes (Supplementary Methods S1.3 and Supplementary Table S3). Additionally, we adopted TransTermHP v2.09 (28) to identify the terminators on the phage genomes.

Taxonomic annotation
To determine the taxonomy of each phage, we adopted Nayfach et al.’s approach by searching the phage proteins against a taxonomically representative HMMs database (22). Specifically, 30,553 taxonomy-specific VOGs from eight taxonomical groups were selected as marker genes (Supplementary Table S4). For each phage, we applied HMMSearch to align its encoded proteins with the VOGs and assigned them to the taxonomical group with the most HMM hits.

Functional annotation
To assign functional annotations to genomic elements, we employed a series of analysis tools. First, we utilized tRNAscan-SE v2.0 with a bacterial option (29) and ARAGORN v1.2.41 with the bacteria genetic code (30) to detect tRNA and tmRNA genes within phage genomes. Additionally, we incorporated AcrRanker (31) and homolog-based search, using Anti-CRISPR database (32) as the reference, to identify anti-CRISPR proteins (Supplementary Methods S1.4). We utilized CRISPRCasFinder v4.2.20 (33) to detect CRISPR arrays on phage genomes. Furthermore, we employed mmseqs (27) to conduct a homology search for phage proteins against VFDB (34) and CARD (35), which allowed us to identify virulence factor and antimicrobial resistance genes on phage.
genomes if the match met the thresholds of > 80% identity and > 40% coverage (36). Finally, we utilized TMHMM v2.09 (37) to predict the topology of transmembrane proteins.

**Genome comparison**

**Sequence clustering**

Clustering homologous phages enables comparative analyses of genomes. We performed a two-step procedure to assign the phages into subclusters and clusters. First, following the suggested criteria (38), we applied mmseqs (27), with a threshold of identity >0.9 and coverage >0.9, to generate subclusters along with their representative sequences. Subsequently, we took the representative sequences as the inputs to another round of mmseqs, with a threshold of identity >0.6 and coverage >0.75, thereby generating clusters.

**Sequence alignment**

To compare CDS sequences in multiple phage genomes, we adopted BLASTP (39) to perform a pairwise alignment between encoded proteins derived from the annotation process. The alignment coverage and identity values from the BLAST outputs were showcased in the alignment visualizations. The order in which the phages are presented in the visualizations was automatically determined to ensure an optimal arrangement of the alignments (Supplementary Methods S1.5).

**Comparative tree construction**

We demonstrated the hierarchies among multiple phages with a tree structure. To construct a comparative tree for multiple phages, we first applied Alfp, which is an alignment-free sequence comparison method, to calculate the genomic distance between the phage sequences. Alfp has demonstrated its superiority for various sequence comparison tasks and exhibited scalability with large datasets (40). Then we utilized neighbor-joining algorithm (41) to construct a comparative tree (Supplementary Methods S1.6).

**Platform development**

PhageScope is hosted on an Ubuntu 20.04.6 LTS server, which is outfitted with 1 TB of memory and 90 TB of storage. The platform’s backend functionality is supported by an in-house framework (42,43) consisting of Apache, Django, PostgreSQL and Typescit+Vue3. All online data visualizations are implemented with Oviz (44). We provide detailed tutorials on the platform to facilitate usage.

**Results**

**PhageScope database**

PhageScope database holds a vast collection of 873 718 phages sourced from diverse public repositories and published databases, consisting of 767 797 nonredundant sequences (Figure 1). The sequence length and GC content distributions are depicted in Supplementary Figure S2. The phage sequences, accompanied by their respective source links, are provided on PhageScope. Annotations and metadata of the phages are frequently lacking in the original data sources. To augment the database, we applied multiple state-of-the-art tools to endow the curated phages with systematic and comprehensive annotations.

The completeness levels for the curated phages are available in PhageScope, allowing users to assess the quality of the sequences. Among the phages, 72 668 sequences are complete, 300 137 with high quality, 212 175 with medium quality, 267 050 with low quality and the remaining 21 688 sequences not-determined (Supplementary Figure S3).

PhageScope provides the phenotype information for the phages, including the host taxonomy and lifestyle. Host information from 530 085 phages is available from the data source, and the remaining are identified through the aforementioned pipeline, with 124 446 from the homology search and 219 187 from DeepHost (23). Consequently, the curated phages are assigned to bacteria of 4723 species, 1649 genera, 435 families, 196 orders, 94 classes and 57 phyla. The complete host taxonomies, accompanied by the information sources, are accessible within PhageScope. Regarding lifestyle, the phages in PhageScope are categorized into 553 688 virulent phages and 320 030 temperate phages, based on information from data submitters or predictions generated by Graphage (24). The host taxonomy and lifestyle distributions are shown in Supplementary Figure S4.

The phage genomes within PhageScope are meticulously annotated with their genomic structures, providing detailed information about the locations of ORFs and transcription terminators. There are 43 088 582 proteins and 6 462 417 terminators detected from stored phages. For proteins, the encoding products, functional classifications, physicochemical properties and annotation sources are provided. The proteins are categorized into 10 functional classes, including lysis, integration, replication, tRNA-related, regulation, packaging, assembly, infection, immune and hypothetical protein (Supplementary Table S5). For the terminators, region types and confidence scores are available.

Additionally, PhageScope yields taxonomical annotations for the phages via homology search, resulting in phages of Cauloviricetes, Microviridae, Inoviridae, Riboviria, Cressnaviricota or Parvoviridae, Ampullaviridae, Bicaudaviridae or Turriviridae, Ligamenvirales, and Autolykiviridae, Fuselloviridae or Guttaviridae. Taxonomical classification enables users to explore genetic and functional traits within specific taxonomies.

PhageScope also equips the phages with exhaustive annotations of the functional elements associated with RNA molecules, CRISPR systems, host interaction and protein topology. By screening the phage genomes with the pipelines mentioned above, we have identified 691 091 tRNA genes, 11 516 tmRNA genes, 307 329 anti-CRISPR proteins, 56 652 CRISPR arrays, 41 609 virulent factors, 2602 antimicrobial resistance genes and 4 020 770 transmembrane proteins (Supplementary Table S6). All of these functional elements, along with their corresponding supplementary information, are curated within PhageScope.

Furthermore, we have clustered the phage sequences in PhageScope according to the sequence similarities to facilitate comparative analysis. For the resultant 555 901 clusters and 669 183 subclusters, PhageScope provides multiple sequence alignment and hierarchical comparative results with visualizations. The sequence alignment results present homologous CDSs among the sequences, and the hierarchical comparative results exhibit sequence similarity with a hierarchical structure, allowing researchers to comprehend the evolutionary connections and diversity among the phages.

All of the phage sequences, completeness reports, phenotype information, taxonomy categories, protein sequences and
functional annotations, along with their sources, are ready for download in PhageScope.

**Automatic analyses and visualizations**

PhageScope supplies automatic analysis workflows for users to study their customized phages efficiently (Figure 2). The workflow encompasses the aforementioned pipelines for genome annotation and genome comparison analyses (Supplementary Methods S1.7). Users can select curated phages or upload their customized phages to perform partial or complete analyses in the workflows tailored to their needs. Additionally, users have the option to compare their phage genomes with the PhageScope database in genome comparison pipelines. Once the submitted tasks are completed, users can easily download the resultant documents and visualizations (Supplementary Figures S5–S12). The visualizations, which support informative tooltips to deliver detailed introductions and information, are prepared in an optional format with high quality, making them seamlessly incorporated into academic publications.

**Comparison to the existing databases and webservers**

We compare PhageScope with existing online bacteriophage databases, including PHROG (12), MVP (11) and PhagesDB (10) and webservers, including PhaGAA (14) and PhANNs (13). PhageScope demonstrates a multitude of distinct advantages, attributed to the extensive collection of phage sequences, comprehensive annotations, integrated automatic analyses and informative visualizations (Supplementary Table S7).

**Discussion**

PhageScope serves as a well-annotated bacteriophage database, enriched with advanced features such as automated analyses and visualizations. The extensive repertoire of phages, coupled with the comprehensive genome annotations and sequence comparison results from systematic analyses assisted with fifteen state-of-the-art tools, provide details about genetic features and comparative genomics for phage study.

Accurate genome annotations for phages offer valuable insights into their survival mechanisms, host interactions, roles in horizontal gene transfer, as well as their potential and threat to be utilized as therapeutic interventions. PhageScope provides comprehensive annotations for curated phages, encompassing genome completeness, phenotype information, taxonomy classifications, genetic features and functional elements. This wealth of information is crucial to comprehending the landscape of phage diversity and discerning unique characteristics among different phages. For instance, in PhageScope database, temperate phages exhibit a higher abundance of virulent factors and phages rarely encode antimicrobial resistance genes (Supplementary Table S6), which agrees with previous findings (36,45).

Additionally, PhageScope provides comparative genomics to establish informative and intricate evolutionary dynamics within curated phages. The results of sequence clustering, multiple sequence alignment and comparative tree construction assist researchers in elucidating shared genomic features, the diverse genetic contents and the potential interrelationships between phages.

The incorporation of automatic analyses and interactive visualizations within PhageScope reinforces its purpose of
serving the scientific community as a user-friendly and effective tool. Recently, a vast number of phages have been mined from NGS data using computational methods without reliable annotations, restricting researchers from gleaning meaningful deductions from the data. PhageScope enables users to conveniently explore the details of their custom phages and gain insight and understanding of the unique genomic characteristics.

Data availability
All the data are freely available at https://phagescope.deepomics.org.

Supplementary data
Supplementary Data are available at NAR Online.

Acknowledgements
The authors thank Dr Yiqi Jiang for her constructive suggestions for the PhageScope platform.

Funding
Hong Kong Innovation and Technology Fund [GHX/002/19SZ (CityU: 9440262 to S.C.L.)]. Funding for open access charge: Hong Kong Innovation and Technology Fund [GHX/002/19SZ (CityU: 9440262 to S.C.L.)].

Conflict of interest statement
None declared.

References


