*Bioinformatics*, 33(5), 2017, 784–786
doi: 10.1093/bioinformatics/btw711
Advance Access Publication Date: 6 December 2016
Applications Note

Databases and ontologies

# PhagesDB: the actinobacteriophage database

**Daniel A. Russell and Graham F. Hatfull***

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

*To whom correspondence should be addressed

Associate Editor: Jonathan Wren

Received and revised on October 5, 2016; editorial decision on November 3, 2016; accepted on November 7, 2016

## Abstract

The Actinobacteriophage Database (PhagesDB) is a comprehensive, interactive, database-backed website that collects and shares information related to the discovery, characterization and genomics of viruses that infect Actinobacterial hosts. To date, more than 8000 bacteriophages—including over 1600 with sequenced genomes—have been entered into the database. PhagesDB plays a crucial role in organizing the discoveries of phage biologists around the world—including students in the SEA-PHAGES program—and has been cited in over 50 peer-reviewed articles.

**Availability and Implementation:** http://phagesdb.org/

**Contact:** gfh@pitt.edu

## 1 Introduction

The first decade of Actinobacteriophage genomics began with the sequencing of L5 in 1993 (Hatfull and Sarkis, 1993), and culminated in the publication of a comparative analysis of 14 mycobacteriophage genomes in 2003 (Pedulla *et al.*, 2003). At a pace of just over one genome per year, it was feasible to manage the resulting data using GenBank and local spreadsheets. Two subsequent developments, however, rendered this approach untenable. First, the creation of the Phage Hunters Integrating Research and Education (PHIRE) program (Hanauer *et al.*, 2006; Hatfull *et al.*, 2006) established a path for novice high school and college scientists to isolate, purify and characterize their own novel phages, providing a sharp increase in the number of phage isolates available for sequencing (Hatfull, 2015). Second, the advent of Next-Generation Sequencing technologies allowed for faster and cheaper sequencing of those phages' genomes. The result was an exponential increase in the number of phage genomes sequenced in the following decade (Hatfull, 2015). The establishment of the Science Education Alliance-Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program in 2008 (Jordan *et al.*, 2014) further increased the pipeline for isolation and sequencing of Actinobacteriophages, and managing the data generated by these programs presented a challenge.

PhagesDB was developed to be a single, centralized repository of phage information where anyone interested in or involved with phage research could access and submit data. A web-accessible database model allowed the storage and retrieval of data in a systematic and flexible way, as well as providing easy access to anyone with an internet connection. PhagesDB launched in April 2010, and was initially only for Mycobacteriophages (phages of mycobacterial hosts). In 2015, it became the Actinobacteriophage Database to include all phages infecting hosts in the Phylum Actinobacteria. Over the past two years, PhagesDB has averaged ~6400 unique monthly users.

## 2 Design

PhagesDB was created using Django (http://djangoproject.com/) and is hosted on a WebFaction server. Django is a Python-based web-development framework, and was chosen for its versatility, accessibility to non-professional programmers, clarity of documentation, and numerous out-of-the-box features including a fully functional administrative site. Rather than host PhagesDB locally, we chose WebFaction for its easy integration with Django, its data security and its low downtime.

The underlying MySQL database began with a few simple tables—Phages, Clusters and Subclusters—but has grown to include Institutions, Hosts, Publications, Phams, Protocols, Genes, Documents and more. Django's flexibility facilitates the addition of features as needed.

## 3 Features

PhagesDB has individual phage pages for each of the more than 8000 phages that have been entered into the database. These pages contain detailed information about the phages, including discovery details (GPS coordinates, year found, isolation temperature, host bacterium, etc.), sequencing details (genome length, G + C content, type of genome termini, etc.), characterization details (morphotype, cluster/subcluster, gene list, etc.) and useful files (fasta sequence file, plaque picture, restriction digest picture, micrograph, etc.). If applicable, there are links to the GenBank entry for the phage, as well as the paper it was published in.

PhagesDB has a variety of ways the user can view and interact with phage groups. Phage lists can be generated and sorted by host (genus, species, or strain), cluster, subcluster, institution, year found, genome length, G + C content and several other criteria. The filter page (http://phagesdb.org/filter/) allows for combining criteria to target a group of phages with specific characteristics. Each phage cluster and subcluster has its own page with a list of member phages as well as meta-data about the cluster/subcluster itself, such as number of members, average genome size and G + C content and the host genera which its members infect. There is an interactive map that displays all sequenced phages/clusters with known GPS coordinates, which provides information about the geographical distribution of phage isolation locations. The compare page (http://phagesdb.org/compare/) allows users to view all plaque pictures, restriction digest pictures, or micrographs for a given group of phages.

PhagesDB contains amino-acid level information about its sequenced phage genomes via integration with Phamerator (Cresawn *et al.*, 2011). A public Phamerator database of Actinobacteriophages is maintained by the University of Pittsburgh, and its information on Phamilies (Phams) of phage genes is integrated into the PhagesDB website. Each sequenced phage has a clickable gene list, and each gene's amino acid sequence is available for download, local BLAST, or NCBI BLAST. Genes are also searchable by Pham number or gene function.

There are local blastn and blastp databases that contain many phage genomes and proteins that have not yet been published in GenBank or other public databases (http://phagesdb.org/blast/). In order to ensure that recently-sequenced genomes appear promptly in local blast results, a cronjob updates the local blastn database three times a week. A separate daily cronjob checks whether there's a new version of the Actino_Draft Phamerator database, and if so it accordingly updates the local Pham information and blastp database.

In addition to providing phage-related data, PhagesDB also houses valuable resources for Actinobacteriophage biologists. Among these are experimental protocols—organized by workflow stage—describing how to isolate, purify, sequence and annotate phages. There is a list of phage-related publications, a glossary of common phage terms, links to useful bioinformatic tools and a list of institutions that have discovered phages.

## 4 Access and rights to data

All users can freely view the data from PhagesDB. Users also have the option of registering (using a Google, Facebook, Twitter, or PhagesDB account) for the site, allowing them to add new phages they have discovered to the database. Registered users may also modify the data for phages they have entered as they learn more about their phages' characteristics.

In addition to standard web browsing, PhagesDB provides several ways to retrieve the underlying data. The data download page (http://phagesdb.org/data/) has links to download all phage genome sequences, tab-delimited text files with extensive information about each phage, and images containing all plaque pictures, restriction digest gel pictures, or micrographs for any given cluster. Each Pham page has a link to download the amino acid sequences of all members of that Pham for comparative proteomic purposes.

We have recently added an RESTful Application Programming Interface (API) to allow users to access many underlying data in a more computer-friendly way (http://phagesdb.org/api/schema/). The PhagesDB API takes requests for all phages, sequenced phages, phages by host, phages by cluster/subcluster, etc., and returns json objects with the requested information.

PhagesDB houses some data that are not yet published in any other medium, including recently-finished genome sequences. The PhagesDB Terms of Use (http://phagesdb.org/terms/) provides guidelines on how and which data may be used by third parties. For example, users who wish to use otherwise-unpublished data in their own analyses must seek permission from the owners of those data.

## 5 Discussion

PhagesDB satisfies two distinct but complementary needs. First, it provides a centralized resource for the Actinobacteriophage research community to submit, access and analyze data. It was designed to be synchronized with the pace of discovery, avoiding the time-lag between sequencing and availability of annotated genomes in GenBank. On PhagesDB, quality-controlled sequences are posted as soon as they become available, and Pham information is updated frequently. There are currently more than 600 genomes available in PhagesDB that are not yet in GenBank.

Second, PhagesDB links students from across the world who are performing authentic research via the SEA-PHAGES program (Jordan *et al.*, 2014), and demonstrates that their discoveries are indeed meaningful and relevant to the research community at large. Students can enter their own data into a legitimate scientific database, and can compare their discoveries in real time with those of students from around the world. There are more than 6400 registered PhagesDB users with xxx.edu email addresses, reflecting usage by student researchers.

PhagesDB is a critical component of the SEA-PHAGES program, and it seems likely that similar systems for data organization and distribution will be core features of other large and broadly disseminated integrated research-education programs.

## Acknowledgements

We thank all the members of the SEA-PHAGES program and the currnet and former members of the Hatfull laboratory for feedback, input and constructive suggestions.

## Funding

This work has been supported by the Howard Hughes Medical Institute [grant number 54308198]; and by the National Institutes of Health [grant numbers GM093901, GM116884].

*Conflict of Interest*: none declared.

## References

Cresawn,S.G. *et al.* (2011) Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinform.*, **12**, 395.

Hanauer,D.I. *et al.* (2006) Inquiry learning. Teaching scientific inquiry. *Science*, **314**, 1880–1881.

Hatfull,G.F. and Sarkis,G.J. (1993) DNA sequence, structure and gene expression of mycobacteriophage L5: a phage system for mycobacterial genetics. *Mol. Microbiol.*, **7**, 395–405.

Hatfull,G.F. *et al*. (2006) Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. *PLoS Genet.*, **2**, e92.

Hatfull,G.F. (2015) Innovations in undergraduate science education: going viral. *J. Virol.*, **89**, 8111–8113.

Jordan,T.C. *et al*. (2014) A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. *MBio*, **5**, e01051–e01013.

Pedulla,M.L. *et al*. (2003) Origins of highly mosaic mycobacteriophage genomes. *Cell*, **113**, 171–182.