Viral Host Range database, an online tool for recording, analyzing and disseminating virus-host interactions

Quentin Lamy-Besnier, Bryan Brancotte, Hervé Ménager, Laurent Debarbieux

Title

Viral Host Range database, an online tool for recording, analyzing and disseminating virus-host interactions

Authors

Quentin Lamy-Besnier[1,2*], Bryan Brancotte[3*], Hervé Ménager[3], Laurent Debarbieux[1]

Affiliations

1 Bacteriophage, Bacterium, Host Laboratory, Department of Microbiology, Institut Pasteur, Paris F-75015 France.

2 Université de Paris, Paris, France

3 Bioinformatics and Biostatistics Hub, Institut Pasteur, Paris F-75015 France.

Corresponding author

Laurent Debarbieux, Bacteriophage, Bacterium, Host Laboratory, Department of Microbiology, Institut Pasteur, Paris F-75015 France

laurent.debarbieux@pasteur.fr

* These authors have equally contributed

Abstract

Motivation

Viruses are ubiquitous in the living world, and their ability to infect more than one host defines their host range. However, information about which virus infects which host, and about which host is infected by which virus, is not readily available.

Results

We developed a web-based tool called the Viral Host Range database to record, analyze and disseminate experimental host range data for viruses infecting archaea, bacteria and eukaryotes.

Availability and implementation

The ViralHostRangeDB application is available from https://viralhostrangedb.pasteur.cloud. Its source code is freely available from the Gitlab hub of Institut Pasteur (https://gitlab.pasteur.fr/hub/viralhostrangedb).

Introduction

Viral genomic data is expanding, and their *in silico* analysis poses many challenges, including how to predict the likely host of a given virus (de Jonge, et al., 2020; Dzunkova, et al., 2019; Kieft, et al., 2020; Li, et al., 2020; Santiago-Rodriguez and Hollister, 2019). The gold standard for host identification remains the experimental evidence, which can take a long time and considerable effort to obtain. Four years passed between the prediction of *Bacteroidetes* as the putative host for crAssphage (the most abundant human gut bacteriophage) and the first experimental evidence that the strain *Bacteroidetes intestinalis* APC919/174 serves as a host for ϕcrAss001 (Dutilh, et al., 2014; Shkoporov, et al., 2018).

The GenBank (Sayers, et al., 2019) database might be expected to provide information about the host of a virus, but these records mostly identify the host only to genus or species level, which is insufficient. For instance, the host indicated for bacteriophage T4 is the bacterium *Escherichia coli,* with no identification of a strain, which is as imprecise as indicating that human cells are the host for HIV-1. For a non-expert, such information suggests that any *E.*

coli strain can be infected by bacteriophage T4, or that any human cell can be infected by HIV-1. Another public resource that could be used is the International Committee on Taxonomy of Viruses (ICTV) (Lefkowitz, et al., 2018). However, host is not indicated in the data available from the ICTV website (talk.ictvonline.org). Finally, it is possible to search in microbial collections (ATCC; www.atcc.org, DSMZ; www.dsmz.de) the host associated with a deposited virus, but, unfortunately, these resources contain data for only limited numbers of published virus/host pairs.

Over and above the identification of a single host for virus propagation, virus host range is another characteristic that is not readily available from public data sources. For viruses infecting multicellular organisms, including humans, in particular, the determination of host range is limited by the ability to grow cell lines. By contrast, for unicellular organisms, the number of hosts to be tested is very large, but unfortunately data is rarely published under an exploitable format. Interestingly, bacteriophage host range data is as old as the first article naming these viruses, published in 1917 by F. d'Herelle, in which bacteriophages infecting a *Shiga* strain were reported to be unable to infect *Flexner* or *Hiss* strains (d'Herelle, 1917).

For decades, viral host range tests were routinely performed for the typing of bacteria (Sabat, et al., 2013; Sechter, et al., 2000). Nowadays, host ranges are being determined for an increasing number of bacteriophages to identify candidates for phage therapy. This treatment for bacterial infections was originally proposed in 1917, and is used regularly in some countries (Georgia, Poland) (d'Herelle, 1917; Kutateladze, 2015). Its use is now expanding worldwide to treat infections caused by antibiotic-resistant pathogens (Corbellino, et al., 2019; Dedrick, et al., 2019; Jennes, et al., 2017; Schooley, et al., 2017). Consequently, semi-automated systems for high-throughput host range tests have been developed (http://www.aphage.com/the-science/). However, only the small number of positive outputs from these tests are finally used, with the bulk of the information obtained discarded and, thus, unavailable.

Another major challenge is the integration of host range data into a single searchable and analysis tool. Viral host range data is, by definition, a variable, which should be regenerated dynamically following the acquisition of new data.

Results

We circumvented the challenges associated with virus host range analysis, by designing the Viral Host Range database (VHRdb, https://viralhostrangedb.pasteur.cloud/), which compiles experimental host range data provided by contributors. This open web-based resource can be used to explore and analyze publicly accessible data with a powerful search engine that scans data and metadata (virus or host names, contributor name, location, GenBank accession number, etc.). Not only can users find a virus, but they can also immediately identify the set of hosts on which it has been tested, across all the available data. Filters, analysis and display settings can facilitate rapid visualization of the most relevant information, such as the highest host range score or the most susceptible host (Figure 1). Importantly, when discrepancies between datasets are detected, they are highlighted and direct access is provided to the source data, for further investigations.

We designed a user-guided process for uploading data compatible with the VHRdb mapping tool, to facilitate comparisons of datasets. This mapping tool is the cornerstone of VHRdb, translating the contributor's original (numerical) data into a unified ranking system. The mapping tool was designed to allow each contributor to classify the results of virus-host interaction tests into a maximum of three responses: "0", for "no infection"; "2" for "infection"; and "1" for "intermediate", corresponding to any interaction that is different from "0" and "2". Then, contributors can readily compare their results with publicly available datasets (curated by administrators to ensure that the database remains homogeneous). If kept private, data are neither accessible to, nor curated by administrators. Analysis across a restricted number of datasets is also possible, to focus on specificities associated with one or several viruses or hosts.

Another issue affecting the accurate appreciation of a virus host range is the lack of precise characterizations of tested hosts. In particular, most of clinical isolates used to determine the host range of bacteriophages for phage therapy applications are not sequenced.. In addition, viruses themselves evolve over time and adapt their host range to the available hosts (Rothenburg and Brennan, 2020).. The VHRdb therefore handles GenBank accession numbers for both viruses and hosts, as a solution to provide unique identifiers.

In addition to the identification of suitable hosts for viruses and the cross-analysis of experimental tests, we anticipate that the VHRdb will become a resource for the development of machine learning approaches, which require large amounts of data, to improve the prediction of the host of a virus, or even the receptor that it uses (Leite, et al., 2018; Young, et al., 2020). It could also be used more directly by clinicians, who will increasingly have access to the genome sequences of pathogens. If the strain infecting a patient is closely related to a tested strain present in the VHRdb, candidate bacteriophages are immediately identified, shortening the time required to develop an appropriate treatment. The VHRdb will also provide opportunities to address fundamental questions in virology, from ecological dynamics to the molecular mechanisms underlying virus-host interactions.

The VHRdb is a unique, publicly accessible resource for the community of microbial virologists, for the rapid identification, characterization and dissemination of data for virus-host interactions of broad interest to the educational, scientific and medical communities, and to private sector entities developing applications.

At the time of publication, the VHRdb holds 15,753 interactions obtained from 739 viruses infecting 1,664 archaeal, bacterial or protist hosts, including the entire Felix d'Herelle collection of bacteriophages.

Methods

Data Availability

The ViralHostRangeDB application is available from https://viralhostrangedb.pasteur.cloud. Its source code is freely available from the Gitlab hub of Institut Pasteur (https://gitlab.pasteur.fr/hub/viralhostrangedb), under the terms of the MIT license, together with detailed documentation (https://hub.pages.pasteur.fr/viralhostrangedb/) including instructions for use, deployment and administration purposes. A demonstration server can be run directly from a docker image (https://hub.docker.com/r/viralhostrangedb/demo), providing a way of testing all features of the application, including the privileges and (in)visibility of private data sources.

Architecture

The architecture of the ViralHostRangeDB web application is based on the Django Web Framework, and the PostgreSQL database. Data are displayed, on the server side, in the Django REST framework. This environment provides efficient and safe data storage as well as tight control access. The application, its database and routine processes (backup, email notifications, virus/host identifier analysis, etc.), are hosted on a Kubernetes cluster (https://kubernetes.io/), providing high availability, scalability and fail-over. The global software quality of the application is ensured through unit test scenarios covering 99% of the code base.

Importing data

Any authenticated user can contribute datasets via the top menu. Datasets can be uploaded as Excel files as detailed in the online documentation (https://hub.pages.pasteur.fr/viralhostrangedb/compatible_file.html). Excel data files are imported with the Pandas and xlrd Python packages (McKinney, 2017). During the mapping of the responses of a file onto the global scheme, the thresholds suggested to users are calculated with the NumPy (Oliphant, 2006) and Scikit-learn (Pedregosa, et al., 2011) packages. The NCBI identifiers describing the host and virus strains are validated with Entrez web services (Sayers, et al., 2020) which are queried with the BioPython (Cock, et al., 2009) package.

Privacy

The access to uploaded datasets can be finely controlled, by restricting it to the uploader only, sharing it with a specific set of other users, or making it public. It is also possible to set permissions for the edition of a dataset for each user. Private data sources can be accessed only by explicitly authorized users, regardless of whether the user is a curator or a privileged administrator. To secure edition operations on the datasets, all modifications are logged and stored in histories, to allow rollback.

Search tool

The web interface allows the interrogation of datasets. A `search module`, accessible either through a quick search box or through a specific advanced search page, can be used to discover datasets through full text and specific filters (e.g. Host or Virus names, contributor, publication…). The exploration module, accessible from the top menu or from the search

results, provides the main functionality of the application: the ability to compare the responses of any number of hosts to any number of viruses, across all the datasets accessible.

Acknowledgments

We warmly thank Roger Carlson, David Dunigan, Mart Krupovic, Sylvain Moineau, Marie-Agnès Petit, Catherine Schouler and Denise Tremblay, and all current and former members of the laboratory of L. Debarbieux for contributing data to the VHRdb. We thank the IT department of Institut Pasteur, including Thomas Menard, in particular, for providing access to the Kubernetes cluster and initial training. We thank Jean-François Charles for assistance in designing the figure. Q.L.-B. is funded by École Doctorale FIRE - Programme Bettencourt.

References

Cock, P.J., *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25(11):1422-1423.

Corbellino, M., *et al.* Eradication of a multi-drug resistant, carbapenemase-producing Klebsiella pneumoniae isolate following oral and intra-rectal therapy with a custom-made, lytic bacteriophage preparation. *Clin Infect Dis* 2019.

d'Herelle, F. Sur un microbe invisible antagoniste des bacilles dysentériques. *C. R. Acad. Sci. Paris* 1917;165:373-375.

de Jonge, P.A., *et al.* Adsorption Sequencing as a Rapid Method to Link Environmental Bacteriophages to Hosts. *iScience* 2020;23(9):101439.

Dedrick, R.M., *et al.* Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. *Nat Med* 2019;25(5):730-733.

Dutilh, B.E., *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 2014;5:4498.

Dzunkova, M., *et al.* Defining the human gut host-phage network through single-cell viral tagging. *Nat Microbiol* 2019;4(12):2192-2203.

Jennes, S., *et al.* Use of bacteriophages in the treatment of colistin-only-sensitive Pseudomonas aeruginosa septicaemia in a patient with acute kidney injury-a case report. *Crit Care* 2017;21(1):129.

Kieft, K., Zhou, Z. and Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;8(1):90.

Kutateladze, M. Experience of the Eliava Institute in bacteriophage therapy. *Virol Sin* 2015;30(1):80-81.

Lefkowitz, E.J., *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* 2018;46(D1):D708-D717.

Leite, D.M.C., *et al.* Computational prediction of inter-species relationships through omics data analysis and machine learning. *BMC Bioinformatics* 2018;19(Suppl 14):420.

Li, M., *et al.* A Deep Learning-Based Method for Identification of Bacteriophage-Host Interaction. *IEEE/ACM Trans Comput Biol Bioinform* 2020;PP.

McKinney, W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media; 2017.

Oliphant, T.E. A Guide to NumPy. Trelgol Publishing; 2006.

Pedregosa, F., *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011;12:2825-2830.

Rothenburg, S. and Brennan, G. Species-Specific Host-Virus Interactions: Implications for Viral Host Range and Virulence. *Trends Microbiol* 2020;28(1):46-56.

Sabat, A.J., *et al.* Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro Surveill* 2013;18(4):20380.

Santiago-Rodriguez, T.M. and Hollister, E.B. Human Virome and Disease: High-Throughput Sequencing for Virus Discovery, Identification of Phage-Bacteria Dysbiosis and Development of Therapeutic Approaches with Emphasis on the Human Gut. *Viruses* 2019;11(7).

Sayers, E.W., *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2020;48(D1):D9-D16.

Sayers, E.W., *et al.* GenBank. *Nucleic Acids Res* 2019;47(D1):D94-D99.

Schooley, R.T., *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(10).

Sechter, I., Mestre, F. and Hansen, D.S. Twenty-three years of Klebsiella phage typing: a review of phage typing of 12 clusters of nosocomial infections, and a comparison of phage typing with K serotyping. *Clin Microbiol Infect* 2000;6(5):233-238.

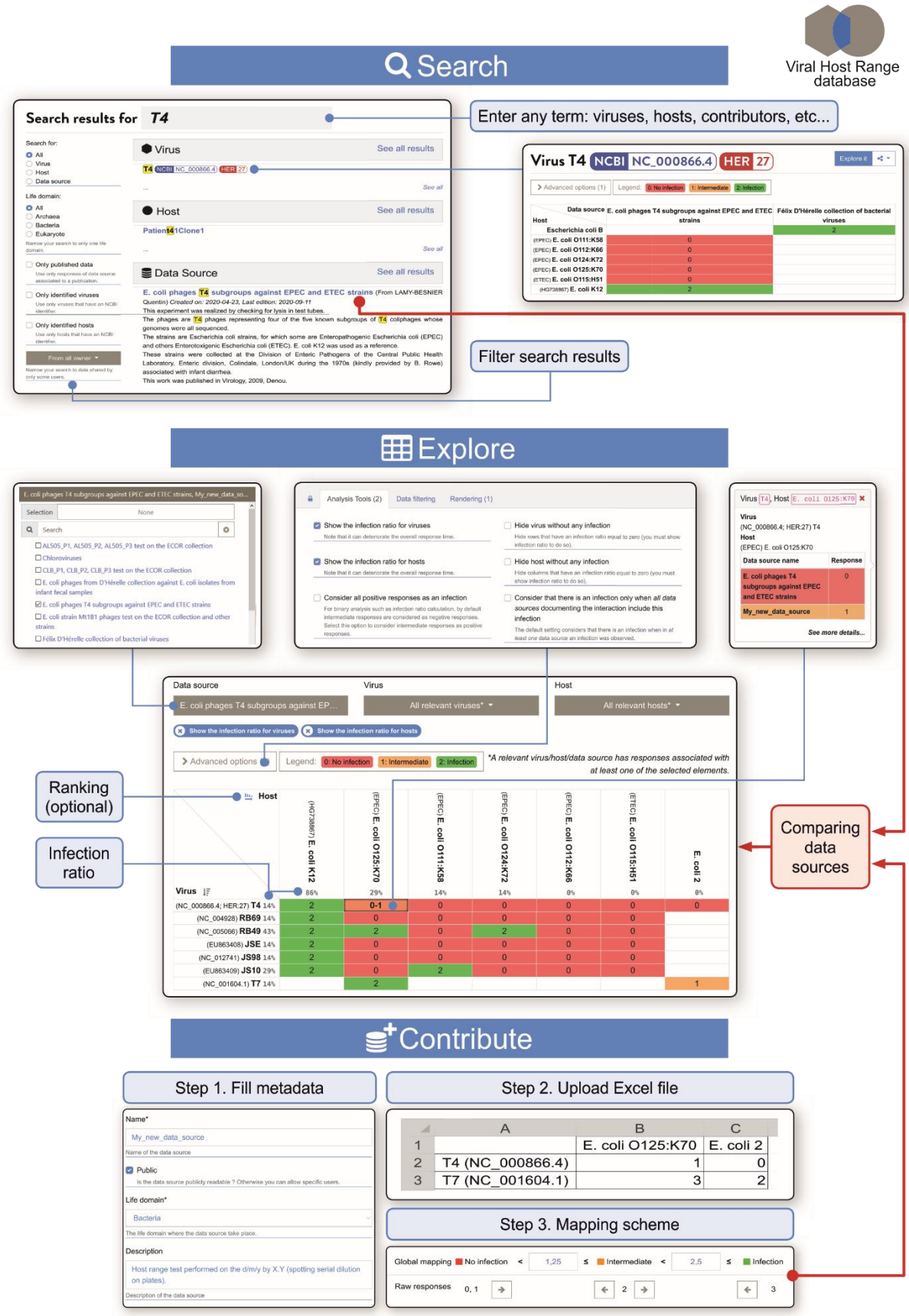
229 Shkoporov, A.N*., et al.* PhiCrAss001 represents the most abundant bacteriophage family in the
230 human gut and infects Bacteroides intestinalis. *Nat Commun* 2018;9(1):4781.
231 Young, F., Rogers, S. and Robertson, D.L. Predicting host taxonomic information from viral genomes:
232 A comparison of feature representations. *PLoS Comput Biol* 2020;16(5):e1007894.

233

234

Legend to Figure 1. Diagram presenting the main functionalities of the Viral Host Range database. The top panel (Search) introduces the search tool and links to subsequent information. The bottom panel (Contribute) presents the main steps that contributors must achieve to record new data. Shown in the middle panel (Explore) is an example of results obtained from dataset comparison, using the datasets selected from the searched results displayed in the top panel and the newly contributed data displayed in the bottom panel (red arrows). Main tools and options to select, rank and display data are also indicated.

243 Figure 1



244