



HAL
open science

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

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

► To cite this version:

Quentin Lamy-Besnier, Bryan Brancotte, Hervé Ménager, Laurent Debarbieux. Viral Host Range database, an online tool for recording, analyzing and disseminating virus-host interactions. *Bioinformatics*, 2021, 60 (4), pp.921-925. 10.1093/bioinformatics/btab070 . pasteur-03166905

HAL Id: pasteur-03166905

<https://pasteur.hal.science/pasteur-03166905>

Submitted on 11 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

1 **Title**

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

4

5 **Authors**

6 Quentin Lamy-Besnier^{1,2*}, Bryan Brancotte^{3*}, Hervé Ménager³, Laurent Debarbieux¹

7

8 **Affiliations**

9 1 Bacteriophage, Bacterium, Host Laboratory, Department of Microbiology, Institut Pasteur,
10 Paris F-75015 France.

11 2 Université de Paris, Paris, France

12 3 Bioinformatics and Biostatistics Hub, Institut Pasteur, Paris F-75015 France.

13

14 **Corresponding author**

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

17 laurent.debarbieux@pasteur.fr

18 * These authors have equally contributed

19

20 **Abstract**

21 **Motivation**

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

25 **Results**

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

29 **Availability and implementation**

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

33

34 **Introduction**

35 Viral genomic data is expanding, and their *in silico* analysis poses many challenges, including
36 how to predict the likely host of a given virus (de Jonge, et al., 2020; Dzunkova, et al., 2019;
37 Kieft, et al., 2020; Li, et al., 2020; Santiago-Rodriguez and Hollister, 2019). The gold standard
38 for host identification remains the experimental evidence, which can take a long time and
39 considerable effort to obtain. Four years passed between the prediction of *Bacteroidetes* as
40 the putative host for crAssphage (the most abundant human gut bacteriophage) and the first
41 experimental evidence that the strain *Bacteroidetes intestinalis* APC919/174 serves as a host
42 for ϕ crAss001 (Dutilh, et al., 2014; Shkoporov, et al., 2018).

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

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

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

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

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

78

79 **Results**

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

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

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

108 In addition to the identification of suitable hosts for viruses and the cross-analysis of
109 experimental tests, we anticipate that the VHRdb will become a resource for the
110 development of machine learning approaches, which require large amounts of data, to
111 improve the prediction of the host of a virus, or even the receptor that it uses (Leite, et al.,
112 2018; Young, et al., 2020). It could also be used more directly by clinicians, who will
113 increasingly have access to the genome sequences of pathogens. If the strain infecting a
114 patient is closely related to a tested strain present in the VHRdb, candidate bacteriophages
115 are immediately identified, shortening the time required to develop an appropriate
116 treatment. The VHRdb will also provide opportunities to address fundamental questions in
117 virology, from ecological dynamics to the molecular mechanisms underlying virus-host
118 interactions.

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

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

126 **Methods**

127 **Data Availability**

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

136 **Architecture**

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

145 **Importing data**

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

155 **Privacy**

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

162 **Search tool**

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

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

170

171 **Acknowledgments**

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

178

179 **References**

- 180 Cock, P.J., *et al.* Biopython: freely available Python tools for computational molecular biology and
181 bioinformatics. *Bioinformatics* 2009;25(11):1422-1423.
- 182 Corbellino, M., *et al.* Eradication of a multi-drug resistant, carbapenemase-producing *Klebsiella*
183 pneumoniae isolate following oral and intra-rectal therapy with a custom-made, lytic bacteriophage
184 preparation. *Clin Infect Dis* 2019.
- 185 d'Herelle, F. Sur un microbe invisible antagoniste des bacilles dysentériques. *C. R. Acad. Sci. Paris*
186 1917;165:373-375.
- 187 de Jonge, P.A., *et al.* Adsorption Sequencing as a Rapid Method to Link Environmental
188 Bacteriophages to Hosts. *iScience* 2020;23(9):101439.
- 189 Dedrick, R.M., *et al.* Engineered bacteriophages for treatment of a patient with a disseminated drug-
190 resistant *Mycobacterium abscessus*. *Nat Med* 2019;25(5):730-733.
- 191 Dutilh, B.E., *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human
192 faecal metagenomes. *Nat Commun* 2014;5:4498.
- 193 Dzunkova, M., *et al.* Defining the human gut host-phage network through single-cell viral tagging. *Nat*
194 *Microbiol* 2019;4(12):2192-2203.
- 195 Jennes, S., *et al.* Use of bacteriophages in the treatment of colistin-only-sensitive *Pseudomonas*
196 aeruginosa septicaemia in a patient with acute kidney injury-a case report. *Crit Care* 2017;21(1):129.
- 197 Kieft, K., Zhou, Z. and Anantharaman, K. VIBRANT: automated recovery, annotation and curation of
198 microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*
199 2020;8(1):90.
- 200 Kutateladze, M. Experience of the Eliava Institute in bacteriophage therapy. *Viol Sin* 2015;30(1):80-
201 81.
- 202 Lefkowitz, E.J., *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of
203 Viruses (ICTV). *Nucleic Acids Res* 2018;46(D1):D708-D717.
- 204 Leite, D.M.C., *et al.* Computational prediction of inter-species relationships through omics data
205 analysis and machine learning. *BMC Bioinformatics* 2018;19(Suppl 14):420.
- 206 Li, M., *et al.* A Deep Learning-Based Method for Identification of Bacteriophage-Host Interaction.
207 *IEEE/ACM Trans Comput Biol Bioinform* 2020;PP.
- 208 McKinney, W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly
209 Media; 2017.
- 210 Oliphant, T.E. A Guide to NumPy. Trelgol Publishing; 2006.
- 211 Pedregosa, F., *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning*
212 *research* 2011;12:2825-2830.
- 213 Rothenburg, S. and Brennan, G. Species-Specific Host-Virus Interactions: Implications for Viral Host
214 Range and Virulence. *Trends Microbiol* 2020;28(1):46-56.
- 215 Sabat, A.J., *et al.* Overview of molecular typing methods for outbreak detection and epidemiological
216 surveillance. *Euro Surveill* 2013;18(4):20380.
- 217 Santiago-Rodriguez, T.M. and Hollister, E.B. Human Virome and Disease: High-Throughput
218 Sequencing for Virus Discovery, Identification of Phage-Bacteria Dysbiosis and Development of
219 Therapeutic Approaches with Emphasis on the Human Gut. *Viruses* 2019;11(7).
- 220 Sayers, E.W., *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic*
221 *Acids Res* 2020;48(D1):D9-D16.
- 222 Sayers, E.W., *et al.* GenBank. *Nucleic Acids Res* 2019;47(D1):D94-D99.
- 223 Schooley, R.T., *et al.* Development and Use of Personalized Bacteriophage-Based Therapeutic
224 Cocktails To Treat a Patient with a Disseminated Resistant *Acinetobacter baumannii* Infection.
225 *Antimicrob Agents Chemother* 2017;61(10).
- 226 Sechter, I., Mestre, F. and Hansen, D.S. Twenty-three years of *Klebsiella* phage typing: a review of
227 phage typing of 12 clusters of nosocomial infections, and a comparison of phage typing with K
228 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

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

242



Viral Host Range database

Search

Search results for T4

Virus See all results

T4 (NCBI NC_000866.4) HER 27

Host See all results

Patition1Clone1

Data Source See all results

E. coli phages T4 subgroups against EPEC and ETEC strains (From LAMY-BESNIER Query) Created on: 2020-04-23; Last edited: 2020-09-11

This experiment was realized by checking for lysis in test tubes.

The phages are T4 phages representing four of the five known subgroups of T4 coliphages whose genomes were all sequenced.

The strains are Escherichia coli strains, for which some are Enteropathogenic Escherichia coli (EPEC) and others Enterotoxigenic Escherichia coli (ETEC). E. coli K12 was used as a reference.

These strains were collected at the Division of Enteric Pathogens of the Central Public Health Laboratory, Enteric division, Colindale, London UK during the 1970s (kindy provided by B. Rowe) associated with infant diarrhea.

This work was published in Virology, 2009, Denou.

Enter any term: viruses, hosts, contributors, etc...

Virus T4 **NCBI NC_000866.4** **HER 27** Explore

Advanced options (1) Legend: 0: No infection 1: Intermediate 2: Infection

Host	Data source E. coli phages T4 subgroups against EPEC and ETEC strains	Félix D'Hérelle collection of bacterial viruses
Escherichia coli B	0	2
(EPEC) E. coli O111:K58	0	
(EPEC) E. coli O112:K66	0	
(EPEC) E. coli O124:K72	0	
(EPEC) E. coli O125:K70	0	
(ETEC) E. coli O15:H51	0	
(ETEC) E. coli K12	2	

Explore

Selection

Search

- ALS05_P1, ALS05_P2, ALS05_P3 test on the ECOR collection
- Chloroviruses
- CLB_P1, CLB_P2, CLB_P3 test on the ECOR collection
- E. coli phages from D'Hérelle collection against E. coli isolates from infant fecal samples
- E. coli phages T4 subgroups against EPEC and ETEC strains
- E. coli strain M181 phages test on the ECOR collection and other strains
- Félix D'Hérelle collection of bacterial viruses

Analysis Tools (2) Data filtering Rendering (1)

Show the infection ratio for viruses

Show the infection ratio for hosts

Consider all positive responses as an infection

Hide virus without any infection

Hide host without any infection

Consider that there is an infection only when all data sources documenting the interaction include this infection

Virus [T4], Host [E. coli O125:K70]

Data source name	Response
E. coli phages T4 subgroups against EPEC and ETEC strains	0
My_new_data_source	1

See more details...

Data source: E. coli phages T4 subgroups against EP...

Virus: All relevant viruses*

Host: All relevant hosts*

Advanced options Legend: 0: No infection 1: Intermediate 2: Infection

Host	(HGT38887) E. coli K12	(EPEC) E. coli O125:K70	(EPEC) E. coli O111:K58	(EPEC) E. coli O124:K72	(EPEC) E. coli O112:K66	(ETEC) E. coli O15:H51	E. coli 2
Virus [F]	85%	29%	14%	14%	8%	8%	8%
(NC_000866.4) HER:27 T4	2	0-1	0	0	0	0	0
(NC_04929) RB69	2	0	0	0	0	0	0
(NC_00506) RB49	2	2	0	2	0	0	0
(EU863408) JSE	2	0	0	0	0	0	0
(NC_012741) JS98	2	0	0	0	0	0	0
(EU863409) JS10	2	0	2	0	0	0	0
(NC_001604.1) T7		2					1

Step 1. Fill metadata

Step 2. Upload Excel file

Step 3. Mapping scheme

Name*

My_new_data_source

Name of the data source

Public

Life domain*

Bacteria

Description

Host range test performed on the d/mly by X.Y (spotting serial dilution on plates).

	A	B	C
1		E. coli O125:K70	E. coli 2
2	T4 (NC_000866.4)	1	0
3	T7 (NC_001604.1)	3	2

Global mapping

Raw responses 0, 1 → 2 → 3

Ranking (optional)

Infection ratio

Comparing data sources