What the Phage: A scalable workflow for the identification and analysis of phage sequences

Mike Marquet[1,2], Martin Hölzer[3,4], Mathias W. Pletz[1], Adrian Viehweger[5], Oliwia Makarewicz[1], Ralf Ehricht[6,7,8], Christian Brandt[1]

[1] Jena University Hospital, Jena, 07747, Germany

[2] Center of Sepsis Control and Care (CSCC), Jena, Germany

[3] RNA Bioinformatics and High-Throughput Analysis, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany

[4] MF1 Bioinformatics, Robert Koch Institute, 13353 Berlin, Germany

[5] Institute for Medical Microbiology and Epidemiology of Infectious Diseases, University Hospital Leipzig, Leipzig, 04103, Germany

[6] Leibniz Institute of Photonic Technology (Leibniz-IPHT), Jena, Germany

[7] InfectoGnostics Research Campus, Jena, Germany

[8] Institute of Physical Chemistry, Friedrich-Schiller-University Jena, Jena, Germany

Abstract

Phages are among the most abundant and diverse biological entities on earth. Phage prediction from sequence data is a crucial first step to understanding their impact on the environment. A variety of bacteriophage prediction tools have been developed over the years. They differ in algorithmic approach, results, and ease of use. We, therefore, developed “What the Phage” (WtP), an easy-to-use and parallel multitool approach for phage prediction combined with an annotation and classification downstream strategy, thus, supporting the user’s decision-making process by summarizing the results of the different prediction tools in charts and tables. WtP is reproducible and scales to thousands of datasets through a workflow manager (Nextflow). WtP is freely available under a GPL-3.0 license (https://github.com/replikation/What_the_Phage).

Introduction

Bacteriophages (phages) are viruses that infect prokaryotes and replicate by utilizing the host's metabolism (1,2). They are among the most abundant and diverse organisms on the planet and inhabit almost every environment (2). Phages drive and maintain bacterial diversity by perpetuating the coevolutionary interactions with their bacterial prey, facilitating horizontal gene transfer and nutrient turnover through continuous cycles of predation and coevolution (3,4). They directly impact the microbiome, e.g., the human gut, and can influence human health (5). At the same time, phages in aquatic habitats are responsible for the mortality of nearly 20–40% of prokaryotes every day (6). However, despite having considerable impacts on microbial ecosystems, they remain one of the least understood members of complex communities (7).

The sequencing of the entire DNA of environmental samples (metagenomics) is an essential approach to gain insights into the microbiome and functional properties.

It should be noted that due to the genome size of phages between 5 kbp to 500 kbp (8), their entire genome can be sequenced assembly-free via long-read technologies (e.g., Oxford Nanopore Technologies or PacBio) (9). They facilitate phage genome recovery in their natural habitat without the need to culture their hosts to isolate the phages (2) and lead to a rapid increase in human gut virome studies (10). This development demands reliable and easy-to-use phage prediction tools and workflows that can be directly used on assembled sequencing data.

However, predicting phages from metagenomes in general and their differentiation from prophages remains a challenge as there is no established computational gold standard (11).

Existing prediction tools rely on direct comparison of sequence similarity (12,13), sequence composition (14,15), and models based on these features derived through learning algorithms (12,13,16,17).

The performance of each prediction method varies (18) depending on the sample type or material, the sequencing technology, and the assembly method, which makes the correct choice for any given sample difficult without having to install and test several tools.

To further complicate matters, the user can choose from many tools based on different calculation strategies, software dependencies, and databases. While working with these phage prediction tools, we observed various installation issues and conflicts, making a multi-tool screening approach unnecessary complex, and time-consuming.

To overcome these obstacles and issues, we developed “What the Phage” (WtP), a reproducible, accessible, and scalable workflow utilizing the advantages of multiple prediction tools in parallel to detect and annotate phages.

Design and Implementation

WtP was implemented in Nextflow, a portable, scalable, and parallelizable workflow manager (19). At the time of writing, twelve different approaches to predict phage sequences are included in WtP besides other programs for further annotation and classification. WtP uses so-called containers (Docker or Singularity) for an installation-free workflow execution without dependency or operating system conflicts for each of the currently over 21 programs included. All containers are pre-build, version-controlled, online available at dockerhub.com, and automatically downloaded if used. Additionally, all nine different databases/datasets used by the workflow are automatically managed. The modular code structure and functionalities of Nextflow and Docker/Singularity allow easy integration of other phage prediction tools and additional analysis steps in future releases of the pipeline. The workflow consists of two main phases, which are executed subsequently or, if specified, individually (Figure 1):

1. Prediction: The prediction of putative phage sequences
2. Annotation & Taxonomy: The gene annotation and taxonomic classification of phage sequences

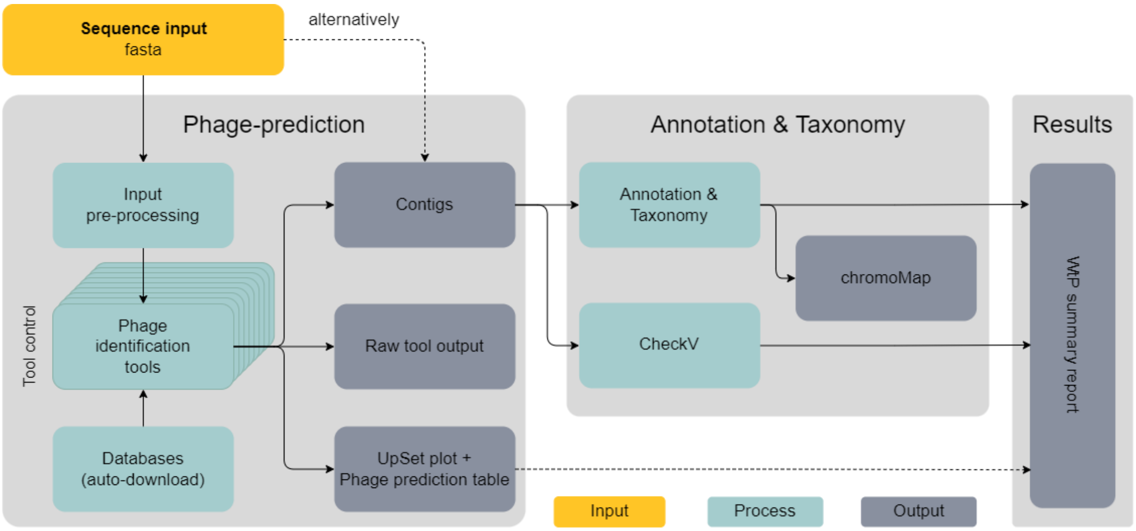


Figure 1: Simplified DAG chart of the "What the Phage" workflow. Sequence input (yellow) can either be first-run through the "prediction" and subsequently "Annotation & Taxonomy" as a whole or used directly as an input for the "Annotation & Taxonomy" only. Each of the multiple phage prediction tools can be individually controlled if needed (tool control).

Prediction and Visualization

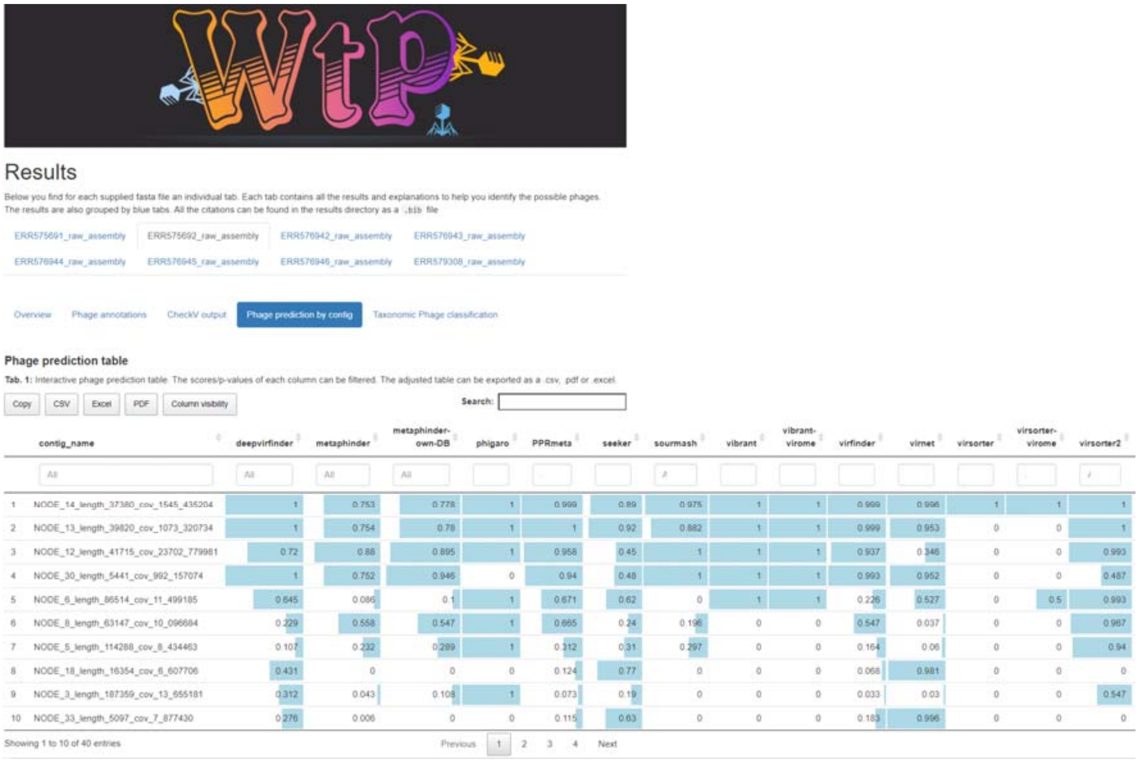
The first stage takes a multi-fasta file as input (e.g., a metagenome assembly), formats it to the demands of each tool, and filters sequences below a user-defined length threshold (1,500 bp by default) via SeqKit v0.10.1 (20). Sequences that are too small usually generate false-positive hits, as Gregory *et al*. (21) observed. The phage prediction process is performed by eleven different tools in parallel: VirFinder v1.1 (15), PPR-Meta v1.1 (17), VirSorter v1.0.6 (with and without virome mode) (13), DeepVirFinder v1.0 (22), Metaphinder with no release version (using default database and own database (Zheng *et al.*)) (23), Sourmash v2.0.1 (14), Vibrant v1.2.1 (with and without virome mode) (12), VirNet v0.1 (24) Phigaro v2.2.6 (25), Virsorter2 v2.0 and Seeker (26) with no release version. Tool outputs are collected in a detailed result report (See section: Result report, Figure 2, https://replikation.github.io/What_the_Phage/).

Functional annotation & Taxonomy

For this step, phage positive contigs are used and either automatically retrieved from the prediction step or directly via user input. Prodigal v2.6.3-1 (27) is used in metagenome mode to predict ORFs and HMMER v3.3 (cutoff: -E 1e-30) (28) to identify homologs via the pVOG-database (29). All annotations are summarized in an interactive HTML file via chromoMap (30) (see Figure 3). Additionally, WtP classifies all contigs via sourmash and provides a probability score to the corresponding taxonomic classification based on Zheng *et al*. database.

Result report

WtP streamlines the detection of phage sequences across multiple tools in their default settings, thus balancing some drawbacks (e.g., relying on updated databases, only predicting phages available in databases). To ease the data interpretation for the user, WtP collects the results in a detailed summary report HTML file for simplified interpretation (Figure 2, full report on: https://replikation.github.io/What_the_Phage/). The report contains an UpSet plot summarizing the prediction performance of each tool (Figure 2). Finally, the “phage prediction by contig table” (Figure 2) summarizes tool outputs for each contig based on the tool's output. WtP assigns numeric values to tools that do not generate p-values or scores between 0 and 1 (see result report, Phage prediction by contig section) and sorts them based on phage likelihood. All the results are individually filterable so the user can consider additional insights or information provided by community platforms such as IMG/VR, iVirus, or VERVE-NET.



Results

Below you find for each supplied fasta file an individual tab. Each tab contains all the results and explanations to help you identify the possible phages. The results are also grouped by blue tabs. All the citations can be found in the results directory as a .bib file

Phage prediction table

Tab. 1: Interactive phage prediction table. The scores/p-values of each column can be filtered. The adjusted table can be exported as a .csv, .pdf or .excel.

	contig_name	deepvirfinder	metaphinder	metaphinder-own-DB	phigaro	PPRmeta	seeker	sourmash	vibrant	vibrant-virome	virfinder	virnet	virsorter	virsorter-virome	virsorter2
	All	All	All	All											
1	NODE_14_length_37380_cov_1545_435204	1	0.753	0.778	1	0.999	0.89	0.975	1	1	0.999	0.996	1	1	1
2	NODE_13_length_39820_cov_1073_320734	1	0.754	0.78	1	1	0.92	0.882	1	1	0.999	0.953	0	0	1
3	NODE_12_length_41715_cov_23702_779981	0.72	0.88	0.895	1	0.958	0.45	1	1	1	0.937	0.346	0	0	0.993
4	NODE_30_length_5441_cov_992_157074	1	0.752	0.946	0	0.94	0.48	1	1	1	0.993	0.952	0	0	0.487
5	NODE_6_length_86514_cov_11_499185	0.645	0.086	0.1	1	0.671	0.62	0	1	1	0.226	0.527	0	0.5	0.993
6	NODE_8_length_63147_cov_10_096684	0.229	0.558	0.547	1	0.665	0.24	0.196	0	0	0.547	0.037	0	0	0.967
7	NODE_5_length_114288_cov_8_434463	0.107	0.232	0.289	1	0.312	0.31	0.297	0	0	0.164	0.06	0	0	0.94
8	NODE_18_length_16354_cov_6_607706	0.431	0	0	0	0.124	0.77	0	0	0	0.068	0.981	0	0	0
9	NODE_3_length_187359_cov_13_655181	0.312	0.043	0.108	1	0.073	0.19	0	0	0	0.033	0.03	0	0	0.547
10	NODE_33_length_5097_cov_7_877430	0.276	0.006	0	0	0.115	0.63	0	0	0	0.183	0.996	0	0	0

Showing 1 to 10 of 40 entries

Figure 2: Example of the final report showing the analyzed sample ERR575692 *with the "Phage prediction by contig table" section opened.*

Other features

All mandatory databases and containers are automatically downloaded when the workflow is started and stored for the following executions. Additionally, the workflow can be pre-setup to analyze sequences offline subsequently. WTP provides the raw output of each tool to support a transparent and reproducible mode of operation. Maximum execution stability is ensured by automatically excluding phage prediction tools that cannot analyze the input data without failing the workflow (e.g., file too large, not the scope of an individual tool). We also provide a detailed user manual that is regularly updated and available at www.mult1fractal.github.io/wtp-documentation/.

Dependencies and version control

WtP requires only the workflow management software Nextflow (19) and either Docker or Singularity (31) installed and configured on the system. The pipeline was tested on Ubuntu 16.04 LTS, Ubuntu 18.04 LTS, and Windows 10 (via Windows Subsystem for Linux 2 using Docker). The installation process is described in detail at mult1fractal.github.io/wtp-documentation/. Each workflow release specifies the Nextflow version the code was tested on to avoid any version conflicts between the workflow code and the workflow manager at any time. A specific Nextflow version can be directly downloaded as an executable file from https://github.com/nextflow-io/nextflow/releases.

Additionally, each container used in the workflow is tagged by the accompanying tool version, pre-build, and stored on hub.docker.com.

Results

To demonstrate the utility and performance of WtP, we analyzed a described metagenome data set (ENA Study PRJEB6941, ERR575692) using a local desktop machine (24 threads, 60 GB RAM, Ubuntu 18.04.4 LTS) and WtP release v1.1.0. In this study (32), Kleiner *et al*. sequenced an artificial microbiome sample which was produced via bacteria and phage cultures in mice feces (germ-free C57BL/6 J mice). The group added six different phages: P22, T3, T7, □6, M13, and □VPE25 and two bacteria (*Listeria monocytogenes and Bacteroides thetaiotaomicron*) to germ-free feces. We, therefore, expect the prediction of the six known phages and possibly other phage sequences related to both bacteria strains. Still, false-positive hits and tool disagreements are plausible results during the phage prediction step.

The raw read data set composed of eight samples was downloaded from the ENA server and individually assembled via metaSPAdes v3.14 using the default settings (33). The resulting assembly files (available at https://github.com/mult1fractal/WtP_test-data/tree/master/01.Phage_assemblies) were analyzed with WtP (release v1.1.0, default settings). As WtP uses multiple tools for phage prediction, an UpSet plot summarizes for each sample the performance of all approaches executed successfully (Figure 3 for sample ERR575692).

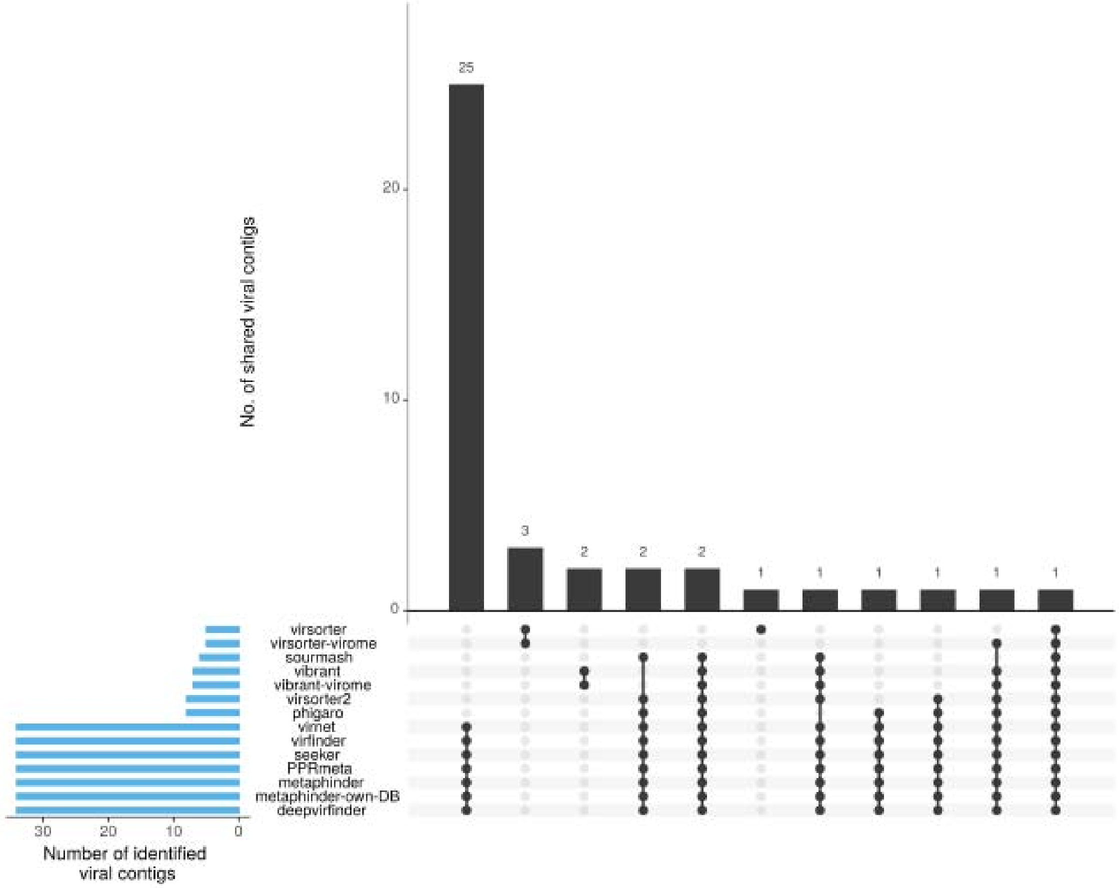


Figure 3: UpSet plot summarizing the prediction performance of each tool for the sample ERR575692. The total amount of identified phage-contigs per tool is shown in blue bars on the left. Black bars visualize the number of contigs that each tool or tool combination has uniquely identified. Each tool combination is shown below the barplot as a dot matrix.

In general, the prediction values are high (>0.7) for the first four sequences (NODE_14, NODE_13, NODE_12, NODE_30), indicating high consensus among the prediction tools, although in some cases tools prediction values were below 0.5 (Phigaro: NODE_30, Seeker:

NODE_12 and NODE_30, Virnet: NODE_12 and Virsorter2: NODE_30) (Supplementary Figure S1).Prediction values for NODE_6 are below 0.67, and Virsorter2 and Phigaro show high values > 0.99. The same applies to NODE_8 and NODE_5, indicating dissonance for these three contigs. Surprisingly, Virsorter and Virsorter-virome only predict contig NODE_14 as phage. In case of dissonance and when tools coincide, validation of contigs via phage annotations and CheckV simplifies the assessment of each contig as a phage or not. In our case, phage genes (like tail and capsid genes) were annotated on all seven contigs (Figure 4).

The workflow was able to detect contigs that correspond to the phages P22 (NODE_12), T3 (NODE_14), T7 (NODE_13). In addition, the phage for the internal Illumina control (phiX174: NODE_30) was also identified. The M13 phage (27) could not be identified as it was not assembled via metaSPAdes due to the low read-abundance and low coverages (below 0.55x, determined by Kleiner *et al.* (27)). The same applies to phage □6, which was not detectable by Kleiner *et al.* (27). However, VPE25 (NODE_6) was initially not taxonomically classified by WtP as it was not represented in the taxonomic database (Zheng *et al.*) at this time; however, the corresponding contig was annotated with essential phage genes (Figure 4). Therefore, this unclassified contig was compared manually via blastn (nr/nt database) and resulted in the genome sequence of VPE25 (PRJEB13004).

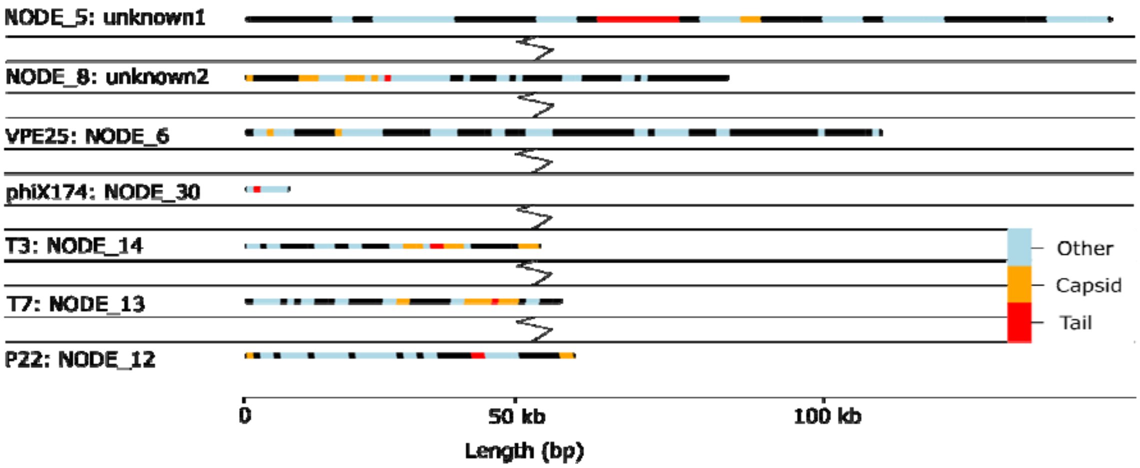


Figure 4: Visual annotation of phage contigs and annotated protein-coding genes via chromoMap. Annotations are colored based on the categories of capsid genes (orange), tail genes (red), and other genes (blue). For better readability, other contigs without either capsid or tail genes have been removed from this figure.

Furthermore, CheckV determined a phage completeness score of over 89% for all seven contigs (Table 1). In addition to the phages mentioned above, two more large contigs with capsid and tail gene annotations indicate prophage(s) of *Salmonella enterica* (contig NODE_5 and NODE_8). Both contigs showed tail and capsid genes and were labeled as prophages via CheckV with estimated completeness of over 99.99 %. These results were manually confirmed using NCBI's blastn (nr/nt database). The sequences matched with 100% identity to *Salmonella enterica (Salmonella enterica strain FDAARGOS_768 chromosome, complete genome)*, but not to prophage sequences.

Table 1: Summary of the CheckV output for the sample ERR575692. All contigs with a completeness > 89 % and a length > 5,000 bp are shown.

Phage name	**Contig_id**	**Gene count**	**CheckV quality**	**Completeness**	**contig length [bp]**
unknown1	NODE_5	107	Complete	100.0	114,288
unknown2	NODE_8	71	High-quality	100.0	63,147
VPE25	NODE_6	137	High-quality	99.99	86,514

phiX174	NODE_30	8	Medium-quality	89.35	5,441
T3	NODE_14	43	High-quality	93.34	37,380
T7	NODE_13	53	Complete	99.48	39,820
P22	NODE_12	67	Complete	100.0	41,715

Some limitations must be noted. No specialized phage assembly strategy or any cleanup step was included during the assembly step. Therefore, some smaller mice host contigs (below 5,000 bp) produced false positive hits. However, these contigs were clearly distinguishable after the “Annotation & Taxonomy” step both in CheckV and due to the lack of typical genes related to, e.g., capsid or tail proteins, showing the application of WtP also for contaminated datasets. WtP does not filter the output of phage prediction tools for prophages, although the CheckV output indicates if a contig could be a prophage.

Conclusion

With the rise of metagenomics and the application of machine learning principles for virus detection, several phage prediction tools have been released over the last few years. All these tools utilize a variety of prediction approaches, all with advantages and limitations. The user's choice for using certain tools often depends strongly on its usability and accessibility and less on its performance. While some tools already come with a packaging system such as Conda or a containerized environment, there exists no general framework for their execution database dependencies, and installation issues prevent many potential users from using certain tools. At least one multitool approach was implemented on a smaller scale by Ann C. Gregory *et al*. (comprising only VirFinder and VirSorter) (20).

The overarching goal of WtP is to make phage prediction tools more accessible for a broader user spectrum and non-bioinformaticians, as culture-free sequencing led to the rapid increase of phage studies (10). WtP acts as an ideal all-encompassing starting point for any given assembly while providing a searchable and filterable report of the analyzed data. The user is provided with sufficient processed data (such as tool performance comparisons, taxonomic assessments, and annotation maps) to work reliably with the predicted sequences and support the decision-making process if different prediction tools are not in agreement with each other. For this, further information and guides are provided either via the report or the hosted manual. WtP streamlines the prediction of phage sequence recognition across multiple tools in a reproducible and scalable workflow to allow researchers to concentrate on their scientific questions instead of software implementations.

Future directions

WtP is a workflow project that will be improved and extended as the modular approach and containerization simplify the integration of new tools. The predictive scope of WtP can be extended to other viruses (such as RNA viruses) and prophages by including future tools specifically designed for such use cases and by adjusting filter and annotation steps. Furthermore, we plan to support the input of long raw reads as an alternative to assemblies. The versioning of WtP represents a well-functioning approach with tested and up-to-date versions of the workflow. Thus, the correct functioning of the workflow is always guaranteed and allows a reliable and fast prediction of phage sequences.

Declarations

Availability

Source code: https://github.com/replikation/What_the_Phage

Result Report: https://replikation.github.io/What_the_Phage/

WtP result data storage: https://osf.io/kuc96/

WtP databases: https://osf.io/wtfrc/

Sequence data used in this work is available at: https://github.com/mult1fractal/WtP_test-data

Competing interest

None to declare.

Funding

This study was supported by the Federal Ministry of Education and Research (BMBF), Germany, grant numbers 01EO1502 and 13GW0423B.

Acknowledgments:

We thank Michael Shamash for his help in properly testing and validating WtP on a Slurm-based HPC utilizing Singularity, Luiz Irber, to improve the sourmash integration. We also thank Polina Tikhonova and Nikos P. for their help in implementing their phage prediction tools Phigaro and Seeker.

References

1. Tulio Pardini G M, Silva B L, Aguiar A LA, Elisa Soto L M. Bacteriophage Genome Sequencing: A New Alternative to Understand Biochemical Interactions between

Prokaryotic Cells and Phages. J Microb Biochem Technol [Internet]. 2017 [cited 2022 Feb 4];09(04). Available from: https://www.omicsonline.org/open-access/bacteriophage-genome-sequencing-a-new-alternative-to-understandbiochemical-interactions-between-prokaryotic-cells-and-phages-1948-5948-1000362.php?aid=92421

2. Clokie MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. Bacteriophage. 2011;1(1):31–45.
3. Reyes A, Semenkovich NP, Whiteson K, Rohwer F, Gordon JI. Going viral: next-generation sequencing applied to phage populations in the human gut. Nat Rev Microbiol. 2012 Sep;10(9):607–17.
4. De Sordi L, Lourenço M, Debarbieux L. The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract. Cell Host Microbe. 2019 Feb 13;25(2):210–8.
5. Divya Ganeshan S, Hosseinidoust Z. Phage Therapy with a Focus on the Human Microbiota. Antibiotics [Internet]. 2019 Aug 27 [cited 2020 May 25];8(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6783874/
6. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol. 2007 Oct;5(10):801–12.
7. Sutton TDS, Hill C. Gut Bacteriophage: Current Understanding and Challenges. Front Endocrinol [Internet]. 2019 [cited 2022 Feb 4];10. Available from: https://www.frontiersin.org/article/10.3389/fendo.2019.00784
8. Hatfull GF, Hendrix RW. Bacteriophages and their Genomes. Curr Opin Virol. 2011 Oct 1;1(4):298–303.
9. Beaulaurier J, Luo E, Eppley J, Uyl PD, Dai X, Turner DJ, et al. Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities [Internet]. Microbiology; 2019 Apr [cited 2020 Jun 10]. Available from: http://biorxiv.org/lookup/doi/10.1101/619684
10. Garmaeva S, Sinha T, Kurilshikov A, Fu J, Wijmenga C, Zhernakova A. Studying the

gut virome in the metagenomic era: challenges and perspectives. BMC Biol. 2019 Oct 28;17(1):84.

11. Overholt WA, Hölzer M, Geesink P, Diezel C, Marz M, Küsel K. Inclusion of Oxford Nanopore long reads improves all microbial and phage metagenome-assembled genomes from a complex aquifer system. bioRxiv. 2019 Dec 19;2019.12.18.880807.
12. Kieft K, Zhou Z, Anantharaman K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of virome function from genomic sequences. bioRxiv. 2019 Nov 26;855387.
13. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015 May 28;3:e985.
14. Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. J Open Source Softw. 2016 Sep 14;1(5):27.
15. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017 Jul 6;5(1):69.
16. Amgarten D, Braga LPP, da Silva AM, Setubal JC. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. Front Genet [Internet]. 2018 [cited 2020 May 1];9. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2018.00304/full
17. Fang Z, Tan J, Wu S, Li M, Xu C, Xie Z, et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. GigaScience [Internet]. 2019 Jun 20 [cited 2020 May 1];8(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6586199/
18. Ho SFS, Millard AD, Schaik W van. Comprehensive benchmarking of tools to identify phages in metagenomic shotgun sequencing data [Internet]. bioRxiv; 2021 [cited 2022 Feb 16]. p. 2021.04.12.438782. Available from: https://www.biorxiv.org/content/10.1101/2021.04.12.438782v1
19. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow

enables reproducible computational workflows. Nat Biotechnol. 2017 11;35(4):316–9.

20. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS ONE [Internet]. 2016 Oct 5 [cited 2020 Jun 10];11(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5051824/

21. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell. 2019 May;177(5):1109-1123.e14.

22. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data by deep learning. ArXiv180607810 Q-Bio [Internet]. 2018 Jun 20 [cited 2020 May 1]; Available from: http://arxiv.org/abs/1806.07810

23. Jurtz VI, Villarroel J, Lund O, Larsen MV, Nielsen M. MetaPhinder—Identifying Bacteriophage Sequences in Metagenomic Data Sets. PLOS ONE. 2016 Sep 29;11(9):e0163111.

24. Abdelkareem AO, Khalil MI, Elaraby M, Abbas H, Elbehery AHA. VirNet: Deep attention model for viral reads identification. In: 2018 13th International Conference on Computer Engineering and Systems (ICCES). 2018. p. 623–6.

25. Starikova EV, Tikhonova PO, Prianichnikov NA, Rands CM, Zdobnov EM, Govorun VM. Phigaro: high throughput prophage sequence annotation. bioRxiv. 2019 Apr 22;598243.

26. Auslander N, Gussow AB, Benler S, Wolf YI, Koonin EV. Seeker: Alignment-free identification of bacteriophage genomes by deep learning. bioRxiv. 2020 Apr 6;2020.04.04.025783.

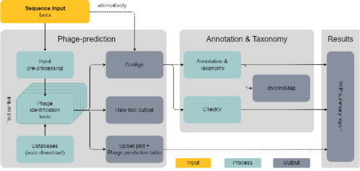
27. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Mar 8;11(1):119.

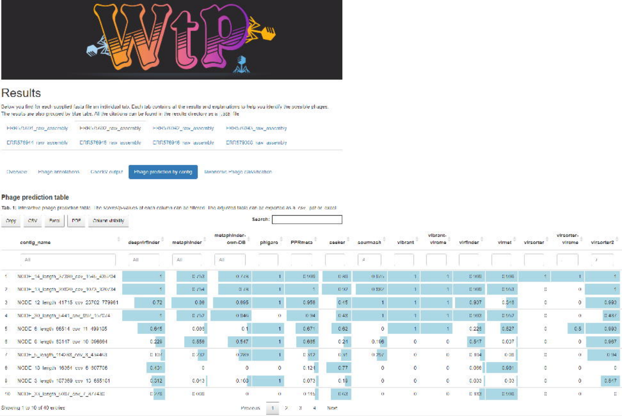
28. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013 Oct 1;29(19):2487–9.

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 Jan 4;45(D1):D491–8.

30. chromoMap-An R package for Interactive Visualization and Annotation of Chromosomes [Internet]. [cited 2020 May 7]. Available from: https://cran.r-project.org/web/packages/chromoMap/vignettes/chromoMap.html
31. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Oper Syst Rev. 2015 Jan 20;49(1):71–9.
32. Kleiner M, Hooper LV, Duerkop BA. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics. 2015 Jan 22;16:7.
33. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017 May;27(5):824–34.


Sequence input
fasta
alternatively
Phage-prediction
Input
pre-processing
Tool control
Phage
identification
tools
Databases
(auto download)
Contigs
Raw tool output
UpSet plot +
Phage prediction table
Annotation & Taxonomy
Annotation &
taxonomy
chromoMap
CheckV
Results
WtP summary report
Input
Process
Output



Results

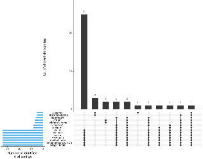
Below you find for each supplied fasta file an individual tab. Each tab contains all the results and explanations to help you identify the possible phages. The results are also grouped by blue tabs. All the citations can be found in the results directory as a .bib file

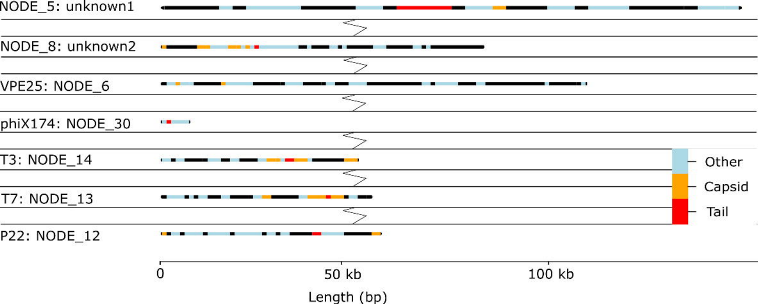
Phage prediction table

Tab. 1: Interactive phage prediction table. The scores/p-values of each column can be filtered. The adjusted table can be exported as a .csv, .pdf or .excel

	contig_name	deepvirfinder	metaphinder	metaphinder-own-DB	phigaro	PPRmeta	seeker	sourmash	vibrant	vibrant-virome	virfinder	virnet	virsorter	virsorter-virome	virsorter2
	All	All	All	All											
1	NODE_14_length_37380_cov_1545_435704	1	0.753	0.778	1	0.998	0.88	0.875	1	1	0.998	0.998	1	1	1
2	NODE_13_length_39820_cov_1073_320734	1	0.754	0.78	1	1	0.92	0.882	1	1	0.988	0.953	0	0	1
3	NODE_12_length_41715_cov_20702_779981	0.72	0.98	0.895	1	0.959	0.45	1	1	1	0.937	0.348	0	0	0.993
4	NODE_30_length_6441_cov_892_157874	1	0.752	0.846	0	0.94	0.48	1	1	1	0.993	0.957	0	0	0.487
5	NODE_6_length_66514_cov_11_499105	0.645	0.008	0.1	1	0.671	0.62	0	1	1	0.226	0.527	0	0.5	0.993
6	NODE_8_length_63147_cov_10_096664	0.229	0.559	0.547	1	0.685	0.21	0.196	0	0	0.547	0.007	0	0	0.967
7	NODE_5_length_114288_cov_8_434463	0.107	0.232	0.289	1	0.312	0.31	0.257	0	0	0.164	0.08	0	0	0.94
8	NODE_10_length_16354_cov_6_807706	0.401	0	0	0	0.121	0.77	0	0	0	0.036	0.991	0	0	0
9	NODE_3_length_107359_cov_10_655161	0.312	0.013	0.103	1	0.073	0.18	0	0	0	0.003	0.03	0	0	0.517
10	NODE_33_length_5087_cov_7_877430	0.276	0.006	0	0	0.115	0.62	0	0	0	0.182	0.998	0	0	0

Showing 1 to 10 of 40 entries


25
3
2
2
2
1
1
1
1
1
1
0
10
20
Number of identified
viral contigs


NODE_5: unknown1
NODE_8: unknown2
VPE25: NODE_6
phiX174: NODE_30
T3: NODE_14
T7: NODE_13
P22: NODE_12
Other
Capsid
Tail
0
50 kb
100 kb
Length (bp)