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. Identification from sequence data is a crucial first step to understand their impact on the environment. A variety of bacteriophage identification 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 identification combined with an annotation and classification downstream strategy, thus, supporting the

user's decision-making process when the phage identification tools are not in agreement to each other. WtP is reproducible and scales to thousands of datasets through the use of 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 metabolism of the host (1). 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 have a direct impact on the microbiome e.g. the human gut and can influence human health (5). However, despite having considerable impacts on microbial ecosystems such as the human gut, they remain one of the least understood members of complex communities (6).

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 small genome size of phages (5 kbp to 300 kbp) (7), their entire genome can be sequenced assembly-free via long-read technologies (e.g., Oxford Nanopore Technologies or PacBio) (8). This facilitates phage genome recovery in their natural habitat, without the need to culture their hosts to isolate the phages (2). However, the identification of phages from metagenomes in general and their differentiation from prophages remains a challenge as there is no established computational gold standard (9).

Existing identification tools rely on direct comparison of sequence similarity (10,11), sequence composition (12,13), and models based on these features derived through learning algorithms (10,11,14,15).

The performance of each identification method varies greatly 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 and based on different software dependencies and databases. While working with these phage identification 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 identification tools in parallel to detect and annotate phages.

Design and Implementation

WtP was implemented in Nextflow, a portable, scalable, and parallelizable workflow manager (16). At the time of writing, twelve different approaches to identify 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 steps which are executed subsequently or, if specified, individually (see Figure 1):

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

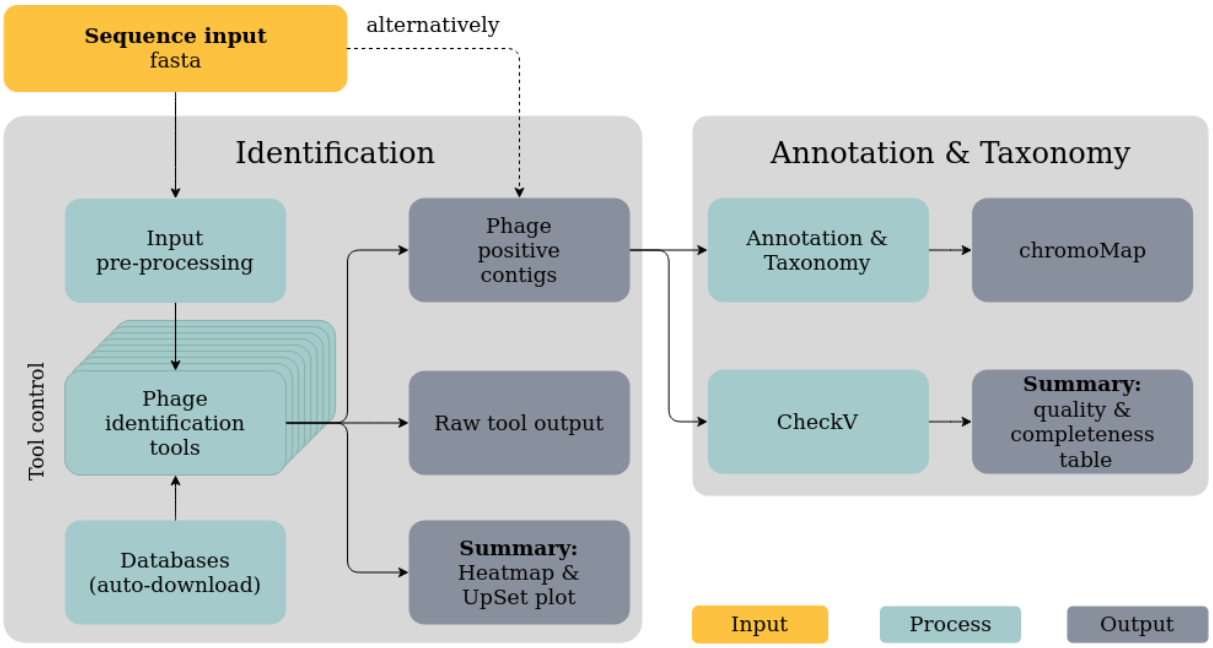

Figure 1: Simplified DAG chart of the “What the Phage” workflow. Sequence input (yellow) can either be first-run through the “Identification” and subsequently “Annotation & Taxonomy” as a whole or used directly as an input for the “Annotation & Taxonomy” only. Each of the multiple phage identification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

Identification and Visualization

The first step 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 (17). Sequences which are too small usually generate false-positive hits as observed by Gregory *et al*. (18). The phage identification process is

performed by twelve different tools in parallel: VirFinder v1.1 (13), PPR-Meta v1.1 (15), VirSorter v1.0.6 (with and without virome mode) (11), DeepVirFinder v1.0 (19), Metaphinder with no release version (using default database and own database) (20), MARVEL v0.2 (14), sourmash v2.0.1 (12), Vibrant v1.2.1 (with and without virome mode) (10), VirNet v0.1 (21) Phigaro v2.2.6, Virsorter 2 v2.0.beta and Seeker with no release version. Positive identifications are collected, filtered by adjustable parameters (Table 1), and the results are summarized via a detailed heat map (Figure 3 A) that serves as a general identification performance overview of each input sample. In addition, and because standard Venn and Euler diagrams are an inadequate solution for quantitative visualization of multiple (n > 4) intersections, we used an R package to generate UpSet plots (22) as a scalable alternative for visualizing intersecting sets and their properties (Figure 2).

Table 1: Overview of the default criteria for “What the Phage” to determine a phage positive contig by the raw output results of each tool. VirSorter and Vibrant are executed in default and virome mode. MetaPhinder is executed with the default database and an own database based on Zheng *et al.* (23)

Tool	**Criteria**	**Filter**
MARVEL	probability according to Random Forest algorithm	> 75 %
VirFinder	p-value	> 0.9
PPR-Meta	contig classification	“Phage”
VirSorter & VirSorter_virome	Category of detection (1, 2 or 3: intact, incomplete or questionable)	Category 1 & 2
MetaPhinder & MetaPhinder-own-DB	A) contig classification & B) average nucleotide identity %	A) Phage & B) > 50
DeepVirFinder	p-value	> 0.9

Vibrant & Vibrant_virome	contig classification	Virus
Virnet	p-value (as median across all hits per contig)	> 0.5
Sourmash	Similarity score	> 0.5
Phigaro	Indicator function	Minimum score threshold: 45.39 Maximum score threshold: 46.0
Virsorter 2 (beta)	dsDNA phage score	> 0.9
Seeker	Score	> 0.75

Annotation & Taxonomy

For this step, phage positive contigs are used and either automatically retrieved from the identification step or directly via user input. Prodigal v2.6.3-1 (22) is used in metagenome mode to predict ORFs and HMMER v3.3 (24) to identify homologs via the pVOG-database (25). All annotations are summarized in an interactive HTML file via chromoMap (26) (see Figure 3 B). Additionally, WtP classifies positive matches via sourmash and thus provides a taxonomic classification of already known phages.

Other features

All mandatory databases and containers are automatically downloaded when the workflow is started and stored for following executions. Additionally, the workflow can be pre-setup in order to subsequently analyse sequences offline. To support a transparent and reproducible mode of operation, the raw output of each tool is provided. Maximum execution stability is ensured by automatically excluding phage identification tools that cannot analyse the input data without failing the workflow (e.g. file too large, not the scope of an individual tool).

Dependencies and version control

WtP requires only the workflow management software Nextflow (16) and either Docker or Singularity (27) 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 https://github.com/replikation/What_the_Phage. 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 always 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 analysed a described metagenome data set (ENA Study PRJEB6941) using a local desktop machine (24 threads, 60 GB RAM, Ubuntu 18.04.4 LTS) and WtP release v1.0.0. In this study (28), Kleiner *et al*. sequenced an artificial microbiome sample which was produced via bacteria and phage cultures in mice faeces (germ-free C57BL/6 J mice). The samples contained six different phages: P22, T3, T7, ϕ6, M13 and ϕVPE25.

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 (29). The resulting eight assembly files (available at https://github.com/mult1fractal/WtP_test-data/tree/master/01.Phage_assemblies) were

analysed with WtP (release v1.0.0, default settings). As WtP uses multiple tools for phage identification, an UpSet plot summarizes for each sample the performance of all approaches executed successfully (see Figure 2 for sample ERR575692 and Supplementary Dataset for all samples).

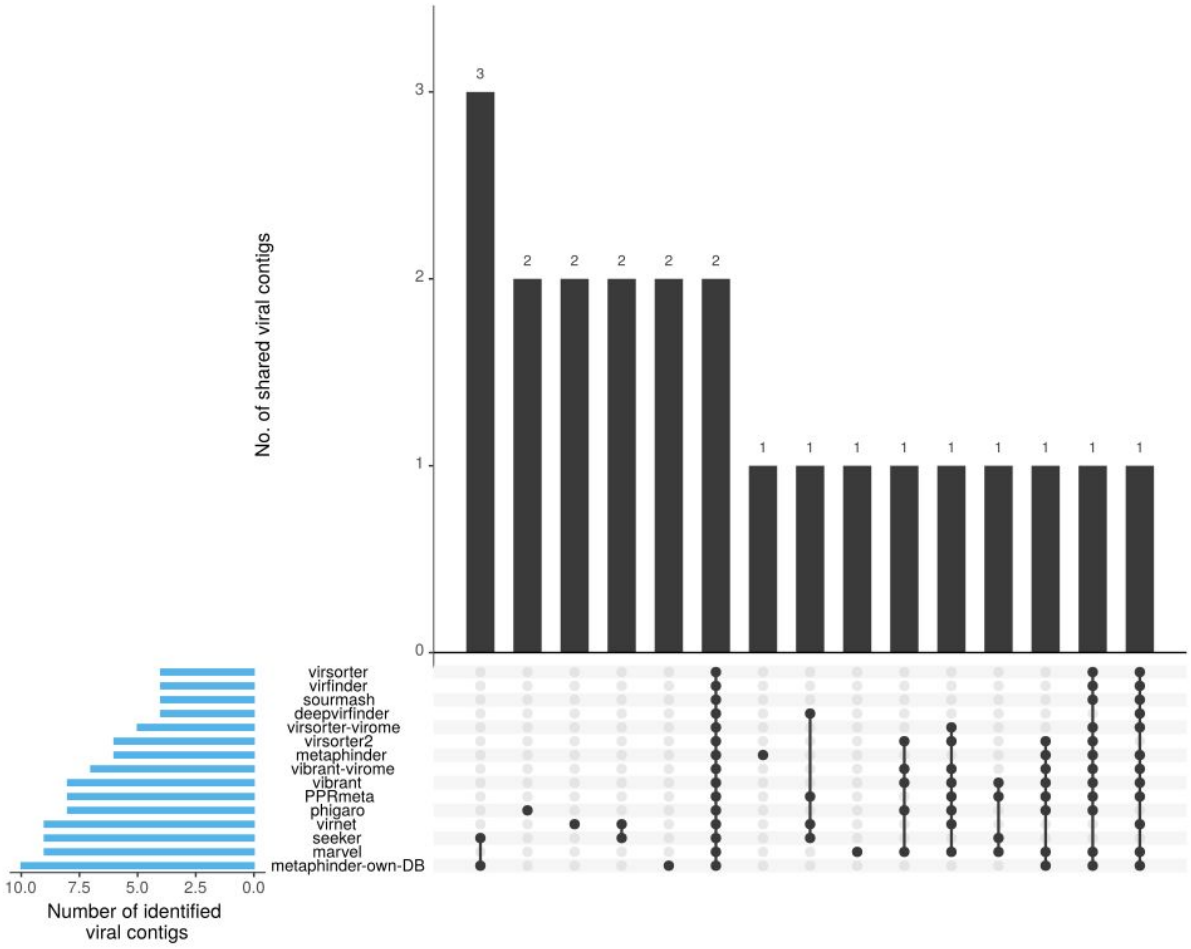

Figure 2: UpSet plot summarizing the identification performance of each tool for the sample ERR575692. The total amount of identified phage-contigs per tool is shown in blue bars on the left. Black bars visualize the number of contigs that each tool or tool combination has uniquely identified. Each tool combination is shown below the barplot as a dot matrix.

The workflow was able to detect contigs that correspond to the phages P22, T3, T7 and VPE25 in all eight samples. In addition, the phage for the internal Illumina control (phiX174)

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 Φ6 which was not detectable by Kleiner *et. al* (27).

Unknown or novel phages, false-positive hits and tool disagreements are all plausible results during the phage identification step. Therefore, WtP generates for each positive contig a visual plot highlighting the identified phage genes and additionally a heatmap to visualize tool agreements for each contig. While for some contigs all tools agreed, in many cases, only a few (e.g. 8 out of 15 in case of VPE25) could identify the phage positive contig (Figure 3 A). VPE25 was initially not taxonomically classified by WtP as it was not represented in the taxonomic database at this time, however, the corresponding positive contig was annotated with multiple different essential phage genes (Figure 3 B). Therefore, this unclassified but positive contig was compared via blastn and matched against the genome sequence of VPE25 (PRJEB13004).

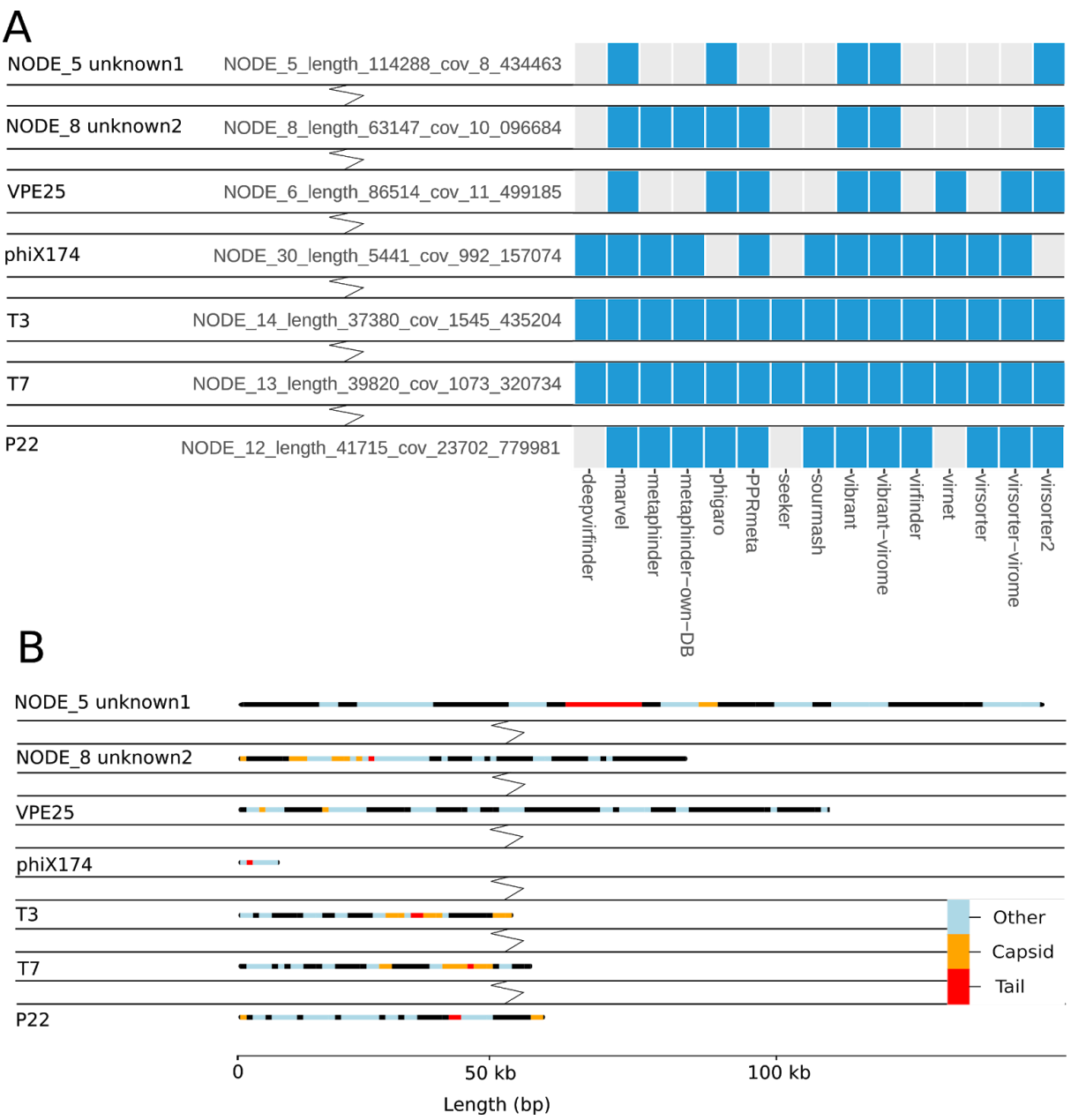

Figure 3: (A) Modified heatmap for assembly ERR575692 visualising the tool agreements per phage positive contig and (B) a visual annotation of phage contigs and annotated protein-coding genes via chromoMap. Annotations are coloured based on the categories of capsid genes (orange), tail genes (red) and other phage genes (blue). Other contigs without either capsid or tail genes have been removed for better readability. All unedited figures for each sample can be found in the Supplementary Dataset.

WtP streamlines the detection of phage sequences across multiple tools and thus balancing some drawbacks of tools (e.g. relying on updated databases, only identifying known phages). This issue is best highlighted in Figure 3 A: P22 could not be identified by either VirNet, Seeker or DeepVirFinder and VPE25 could not be identified by DeepVirFinder, MetaPhinder (both databases), Seeker, Sourmash, VirFinder and VirSorter without the virome option. However, besides other phage genes (e.g. DNA ligase, helicase), capsid and tail genes could be clearly annotated for both (Figure 3 B: P22 and VPE25). Furthermore, CheckV determined a phage completeness score of over 99.99 for both P22 and VPE25 (Table 2).

In addition to the above-mentioned phages, WtP identified two more large contigs with capsid and tail annotations and a taxonomic assignment pointing to a phage of *Salmonella enterica* (contig NODE_5 and NODE_8). However, both contigs are labelled as prophages via CheckV with an estimated completeness of over 99.99 %. These phage positive contigs match to prophages *Salmonella enterica* (additionally confirmed via blastn search on NCBI) and they were identified in six out of eight samples.

Table 2: Summary of the CheckV output for the sample ERR575692. All contigs with a completeness > 10 % 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 specialised phage assembly strategy or any cleanup step were 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

Conclusion

With the rise of metagenomics and the application of machine learning principles for virus detection, several phage identification tools have been released over the last few years. All these tools utilize different identification approaches, all with advantages and limitations. The choice of the user of a certain tool often depends strongly on its usability 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 and different filter parameters, 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 identify positive phage sequences via a comprehensive and extendable multitool approach that is easy to use across different platforms. After a WtP run, the user is provided with sufficient processed data (such as tool performance comparisons, taxonomic assessments, and annotation maps) to reliably work with the identified sequences. The results support the decision-making process of the user if different identification tools are not in agreement with each other (e.g.: see reported results for VPE25 phage). Thus, WtP streamlines the identification 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 containerisation simplify the integration of new tools. Besides the intended main application of the workflow - the identification of phages - the workflow can be used to benchmark current and novel virus detection tools in a continuous manner. 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 raw long 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 identification of phage sequences.

Declarations

Availability

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

Supplementary Dataset:
https://github.com/mult1fractal/WtP_test_profile_results/tree/master/release_test_candidate_v1.0.0

The databases WtP uses are stored here: 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.

Acknowledgements

We thank Michael Shamash for his help in properly testing and validating WtP on a Slurm-based HPC utilizing Singularity. We also thank Polina Tikhonova and Nikos P. for their help in implementing their phage identification tools Phigaro and Seeker.

References

1. G MTP, B LS, L MES. Bacteriophage Genome Sequencing: A New Alternative to Understand Biochemical Interactions between Prokaryotic Cells and Phages. J Microb Biochem Technol. 2017;9(4):1–5.
2. Clokie MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. Bacteriophage. 2011;1(1):31–45.
3. De Sordi L, Lourenço M, Debarbieux L. The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract. Cell Host Microbe. 2019 13;25(2):210–8.
4. 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.
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. Sutton TDS, Hill C. Gut Bacteriophage: Current Understanding and Challenges. Front Endocrinol [Internet]. 2019 [cited 2020 May 23];10. Available from: https://www.frontiersin.org/articles/10.3389/fendo.2019.00784/full
7. Hatfull GF, Hendrix RW. Bacteriophages and their Genomes. Curr Opin Virol. 2011 Oct 1;1(4):298–303.
8. 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
9. 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.
10. 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.
11. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015 May 28;3:e985.
12. Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. J Open Source Softw. 2016 Sep 14;1(5):27.
13. 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.
14. 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
15. 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/
16. 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.
17. 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/
18. 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.
19. 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
20. 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.
21. 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.
22. 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.
23. Zheng T, Li J, Ni Y, Kang K, Misiakou M-A, Imamovic L, et al. Mining, analyzing, and integrating viral signals from metagenomic data. Microbiome. 2019 Mar 19;7(1):42.
24. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013 Oct 1;29(19):2487–9.
25. 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.
26. 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
27. Boettiger C. An introduction to Docker for reproducible research. ACM SIGOPS Oper

Syst Rev. 2015 Jan 20;49(1):71–79.
28. 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.
29. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017 May;27(5):824–34.