Microbiome

METHODOLOGY

Open Access

Gauge your phage: benchmarking of bacteriophage identification tools in metagenomic sequencing data

Siu Fung Stanley Ho[1], Nicole E. Wheeler[1], Andrew D. Millard[2] and Willem van Schaik[1*]

Abstract

Background The prediction of bacteriophage sequences in metagenomic datasets has become a topic of considerable interest, leading to the development of many novel bioinformatic tools. A comparative analysis of ten state-of-the-art phage identification tools was performed to inform their usage in microbiome research.

Methods Artificial contigs generated from complete RefSeq genomes representing phages, plasmids, and chromosomes, and a previously sequenced mock community containing four phage species, were used to evaluate the precision, recall, and F1 scores of the tools. We also generated a dataset of randomly shuffled sequences to quantify false-positive calls. In addition, a set of previously simulated viromes was used to assess diversity bias in each tool's output.

Results VIBRANT and VirSorter2 achieved the highest F1 scores (0.93) in the RefSeq artificial contigs dataset, with several other tools also performing well. Kraken2 had the highest F1 score (0.86) in the mock community benchmark by a large margin (0.3 higher than DeepVirFinder in second place), mainly due to its high precision (0.96). Generally, k-mer-based tools performed better than reference similarity tools and gene-based methods. Several tools, most notably PPR-Meta, called a high number of false positives in the randomly shuffled sequences. When analysing the diversity of the genomes that each tool predicted from a virome set, most tools produced a viral genome set that had similar alpha- and beta-diversity patterns to the original population, with Seeker being a notable exception.

Conclusions This study provides key metrics used to assess performance of phage detection tools, offers a framework for further comparison of additional viral discovery tools, and discusses optimal strategies for using these tools. We highlight that the choice of tool for identification of phages in metagenomic datasets, as well as their parameters, can bias the results and provide pointers for different use case scenarios. We have also made our benchmarking dataset available for download in order to facilitate future comparisons of phage identification tools.

Keywords Bacteriophage, Metagenome, Microbiome, Machine learning, Phage, Benchmarking

Introduction

Bacteriophages (phages) and archaeal viruses are globally ubiquitous, diverse, and typically outnumber their prokaryotic hosts in most biomes [1]. Phages play a key role in microbial communities by shaping and maintaining microbial ecology by fostering coevolutionary relationships [2–4], biogeochemical cycling of essential nutrients [5–7], and facilitating microbial evolution through horizontal gene transfer [8–10]. Despite the

*Correspondence:
Willem van Schaik
w.vanschaik@bham.ac.uk
[1] Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK
[2] Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

abundance and perceived influence phages have on all microbial ecosystems, they continue to be one of the least studied and understood members of complex microbiomes [11]. Phages are obligate parasites which require their host's machinery to replicate and spread via cell lysis. They can either be lytic or temperate, and whilst the former can only follow the lytic life cycle, temperate phages can either follow the lytic or lysogenic cycle [12]. During the lytic cycle, phages hijack host cell machinery to produce new viral particles. In the lysogenic cycle, phages can integrate their genomes into their bacterial or archaeal host genome chromosome as linear DNA or as a self-replicating autonomous plasmid. In addition, a third life cycle called pseudolysogeny has been documented, in which either lytic phage infection is halted or there is no prophage formation [13].

Traditionally, phage identification and characterisation relied on isolation and culturing techniques, which are time-consuming and often require significant expertise. It is also often impractical as many hosts, and their phages, cannot be cultured under laboratory conditions [14]. The arrival of high-throughput next-generation sequencing has allowed metagenomic data from various environments to be generated routinely. Metagenomic sequencing allows direct identification and analysis of all genetic material in a sample, regardless of cultivability [15].

Metagenomic studies can opt to either sequence the whole community metagenome and then computationally isolate viral sequences or physically separate the viral fraction before library preparation to produce a metavirome. The latter approach risks eliminating a large proportion of phages owing to their association with the cellular fraction. This occurs owing to phages being integrated into their hosts' genome as prophages [16], attached to their hosts' surface [17], or when they are in a pseudolysogenic state [18–20]. Purification methods may also remove certain types of phage, e.g. chloroform can inactivate lipid enveloped and/or filamentous phages [21, 22], increasing sampling bias. This process also results in low DNA yields, leading to some metavirome studies having to use multiple displacement amplification (MDA) to achieve sufficient quantities of DNA for library generation [11]. MDA has been shown to produce significant bias into virome composition [23, 24], by preferentially amplifying small circular ssDNA phage, such as those from the family Microviridae [25]. Despite these drawbacks, the purification steps produce a metavirome with very little host contamination, although it is very difficult to produce a viral fraction that is devoid of any cellular material [26]. Metaviromes also have the advantage of being able to identify lower abundance phages at the same sequencing depth due to the approximately 100-fold larger bacterial and archaeal genomes being excluded. Alternatively, whole community metagenomic sequencing can present insights into the host and viral fractions concurrently, allowing host-phage dynamics to be analysed. Integrated phages, or prophages which have been found to be prevalent in some environments [27], can be identified since host genomes are also sequenced in this process. In this study, we focus on computationally extracting phage sequences from whole community metagenomes, as these generally make up a minority of the sequencing data compared to their hosts.

Many tools for identifying viral sequences from mixed metagenomic and virome assemblies have been developed in the last 5 years (Table 1). VirSorter [28] was one of the first of these, with previous tools focusing on prophage prediction (PhiSpy [29], Phage_Finder [30], PHAST/PHASTER [31], ProPhinder [32]) or virome analysis (MetaVir2) [33], VIROME [34]). VirSorter identifies phage sequences by detecting viral hallmark genes that have homology to reference databases and by building probabilistic models based on different metrics (viral-like genes, Pfam genes, uncharacterised genes, short genes, and strand switching) which measure the confidence of each prediction. Since VirSorter's release, other gene-based homology-based tools such as VIBRANT [35], and VirSorter2 [36], have been developed. VIBRANT uses a multilayer perceptron neural network based on protein annotation from Hidden Markov model (HMM) hits to several databases to recover a diverse array of phages infecting bacteria and archaea including integrated prophages. In addition to this, it characterises auxiliary metabolic genes and pathways after identification. VirSorter2 builds on its predecessor by incorporating five distinct random forest classifiers for five different viral groups into one algorithm to improve the diversity of viruses that it can detect accurately. MetaPhinder, unlike the tools above, uses BLAST-based homology hits to a custom database to calculate average nucleotide identity and the likelihood that a sequence is of viral origin.

VirFinder was the first machine learning, viral identification tool to utilise k-mer signatures [45]. VirFinder was shown to have considerably better rates of recovery of viral sequences than VirSorter, especially on shorter sequences (< 5 kbp), but had issues of variable performance in different environments, perhaps due to biases introduced by the reference data used for training the machine learning model [50]. DeepVirFinder [37] improves on VirFinder by applying a convolutional neural network that was trained on an enlarged dataset containing viral sequences from environmental metavirome sequencing data. DeepVirFinder boasts increased viral identification at all contig lengths over

Table 1 Overview of tools to identify and predict phage sequences in microbial ecosystems

Software	Description	Reference
DeepVirFinder	Predicts viral sequences via a k-mer-based deep learning method using convolutional neural networks (CNN). Based on VirFinder	[37]
MARVEL	Machine learning tool for predicting phage sequences in metagenomic bins	[38]
MetaPhinder	Integrates BLAST hits to multiple phage genomes in a database to identify phage sequences in assembled contigs	[39]
viralVerify (metaviralSPAdes)	ViralVerify is a module of metaviralSPAdes which classifies contigs with a Naïve Bayes classifier based on Hidden Markov models protein hits	[40]
PhaMers	Identifies phage sequences by a machine learning model based on k-mer frequencies	[41]
PPR-Meta	Deep learning CNN approach to identify both phages and plasmids	[42]
Seeker	Deep learning framework that uses long short-term memory model (LSTM) which does not depend on sequence motifs	[43]
VIBRANT	Deep learning neural network based on protein signatures which also highlights auxiliary metabolic genes and pathways	[35]
ViraMiner	Extension of DeepVirFinder that is trained to identify any virus that may colonise human samples	[44]
VirFinder	K-mer-based machine learning method for identification of viral contigs	[45]
virMine	Iterative pipeline that relies on the abundance of nonviral sequences in databases to strictly filter out unwanted contigs. Pipeline accepts both reads or assembled contigs	[46]
VirMiner	Web-based pipeline that handles genome assembly, functional annotation using a variety of databases and identification of phage contigs via a random forest algorithm	[47]
VirNet	Deep learning neural network using an attentional neural model trained on nucleotide viral fragments	[48]
VIROME	Web-based pipeline that classifies viral sequences based on homology to databases and functionally annotates them. No local version	[34]
VirSorter	Uses referenced-based and reference-free approaches in unison relying on probabilistic similarity models and referenced-based protein homology searches to increase novel virus detection	[28]
VirSorter2	Builds on VirSorter by applying machine learning to evaluate "viralness" using genomic features. Works with a wider variety of viral groups than its predecessor	[36]
VirusSeeker	Made up of two BLAST-based pipelines — virome and discovery. Virome aligns reads to a curated database to identify viral sequences and compute their abundance in the sample. Discovery focuses on contig-based analysis to aid novel virus discovery	[49]

Tools in italics were not included in this study as they were either not relevant to this study or technical difficulties were encountered during their use. MARVEL was excluded as it currently limited to detecting phages of the Caudovirales order

its predecessor VirFinder whilst mitigating the latter's bias towards phages that are easily cultivable in the laboratory. Kraken2 is a k-mer-based metagenomic taxonomic classifier [51] that can be used for viral detection [52]. It queries k-mers to a database which associates it to their lowest common ancestor taxa which is then used to assign the taxonomic label. PPR-Meta uses three convolutional neural network to identify if a sequence is of phage, plasmid, or chromosomal origin [42]. Sequence features are extracted by the network directly, instead of using pre-selected features such as k-mer signatures or genes. The three networks are also trained on three groups of different sequence lengths to improve its performance on shorter fragments, which some gene-based tools struggle with due to the low number of full-length genes available for analysis. Seeker also uses a neural network, in this case a long short-term memory (LSTM) model, which is not based on pre-selected features [43]. MetaviralSPAdes [40] uses an entirely different approach by leveraging variations in depth between viral and bacterial chromosomes in assembly graphs. The tool is split into three separate modules: a specialised assembler based on metaSPAdes (viralAssembly); a viral identification module that classifies contigs as viral/bacterial/uncertain using a Naive Bayesian classifier (viralVerify); and a module which calculates the similarity of a constructed viral contig to known viruses (viralComplete).

It is important to note that machine learning tools have the potential to identify novel species, which is especially important with the enormous diversity of phages that is theorised to still be unknown [53]. With the development of so many tools using a variety of approaches, a comprehensive comparison and benchmarking are needed to evaluate which tools are most applicable to researchers. The performance of each method can vary based on sample content, assembly method, sequence length, classification thresholds, and other custom parameters. To address these issues, we have benchmarked ten metagenomic viral identification tools using both artificial contigs, mock communities, and real samples.

Results

Benchmarking with RefSeq phage and nonviral artificial contigs

Ten commonly used tools for viral sequence identification in metagenomes were selected for evaluation: DeepVirFinder, Kraken2, MetaPhinder, PPR-Meta, Seeker, VIBRANT, viralVerify, VirFinder, VirSorter, and VirSorter2. All of these tools can be run locally without relying on a web server, accept metagenomic contigs as input, and have been published in the past decade.

We first evaluated all the programmes on the same uniform datasets. All complete phage genomes deposited in RefSeq between 1 January 2020 and 12 August 2021 were downloaded, quality controlled, and uniformly fragmented to sizes between 1 and 15 kbp to create a true-positive set of artificial contigs. A negative set was constructed from all RefSeq bacterial and archaeal chromosomes and plasmids, submitted in the same time period. Multiple steps were taken to ensure these datasets did not include false positives or false negatives. First, all bacterial, archaeal, and phage RefSeq genomes deposited prior to 2020, and the training sets of each machine learning tool were used to dereplicate the datasets used in this study to remove any similar sequences that may cause overfitting of some tools. In addition, chromosome and plasmid sequences with $\geq 30\%$ of their open reading frames having hits to the pVOG database were removed to exclude any remaining viral sequences. As the negative dataset was considerably larger than the positive dataset, we subsampled the negative set by a factor of 14.3, resulting in 253 host chromosomes, and 309 host plasmids (Table 2). This sampling rate was chosen to produce a phage:host ratio (~1:19) that was similar to what is found in human gut microbiomes [53]. Finally, we removed integrated prophages from chromosomal and plasmid sequences using two state-of-the-art prophage detection tools, Phigaro [54] and PhageBoost [55], to prevent their erroneous identification as viral contigs. In total, 2088 prophages were removed from the chromosome set and 91 from the plasmid set. The resulting sequences were then fragmented into artificial contigs and analysed using the different tools (Fig. 1).

All evaluated programmes, except Kraken2, produce thresholds or confidence ranges for viral identification. For tools (DeepVirFinder, MetaPhinder, PPR Meta, Seeker, VirFinder, and VirSorter2) that assign a continuous threshold (score, identity, or probability), a F1 curve was plotted, and an optimal threshold was determined (Additional file 1). For VirSorter and viralVerify, the categories that returned the highest F1 score were used (Additional file 2). In most tools, there was a trade-off between precision and recall. This is likely due to relaxed thresholds allowing for more viral and nonviral sequences to be detected, increasing recall, and decreasing precision simultaneously. For VIBRANT and VirSorter, the positive dataset was additionally run in virome mode and virome decontamination mode, respectively, as this improves viral recovery in samples composed mainly of viral sequences by adjusting the tools sensitivity [28, 35]. The tools we benchmarked on this dataset had highly variable performance in terms of F1 score (0.44–0.93), precision (0.47–1.00), and recall (0.46–0.96) (Fig. 2). VirSorter2, VIBRANT, and PPR-Meta achieved the highest F1 scores of 0.93, 0.93, and 0.92, respectively. VirSorter2 achieved this with high precision (0.92) as well as high recall (0.93), VIBRANT had a higher precision (0.97) and lower recall (0.89), and PPR-Meta had a slightly lower precision at

Table 2 Number of sequences at each stage of the RefSeq benchmarking workflow

Dataset	Chromosome (*n*)	Plasmid (*n*)	Phage (*n*)	Chromosome (bp)	Plasmid (bp)	Phage (bp)
1. Post-2020	7400	9960	1849	30,034,515,475	1,009,997,498	128,660,045
2. Post-2020 dereplicated	3546	4453	901	14,845,391,115	480,796,703	62,387,054
3. Subsampled	253	309	901	1,011,740,231	28,782,822	62,387,054
4. Prophage removed	2307 (2088)	400 (91)	901	965,959,872	27,323,488	62,387,054
5. pVOG removed	2065	313	901	848,361,160	24,433,971	62,387,054
6. Fragmented artificial contigs	104,003	2754	6664	830,889,456	21,783,942	53,426,665

Columns labelled with (*n*) contain the number of sequences at each step, and columns with (bp) indicate the number of base pairs at each step. Steps are numbered as follows:

1. Sequences downloaded from RefSeq which were deposited between 1 January 2020 and 12 August 2021

2. Sequences from (1) which were then dereplicated with RefSeq sequences deposited before 1.st January 2020 and training sets for DeepVirFinder, Seeker, VIBRANT, VirFinder, and VirSorter2

3. Host sequences (chromosome and plasmids) from (2) which were subsampled by a factor of 14.3

4. Host sequences from (3) with prophage removal using Phigaro and PhageBoost. The number in parentheses indicates the number of prophages removed.

5. Host sequences from (4) with sequences that have $\geq 30\%$ of their open reading frames having hits to the pVOG database removed

6. All sequences from (5) randomly and uniformly fragmented to sizes between 1 and 15 kbp for use in the benchmarking study

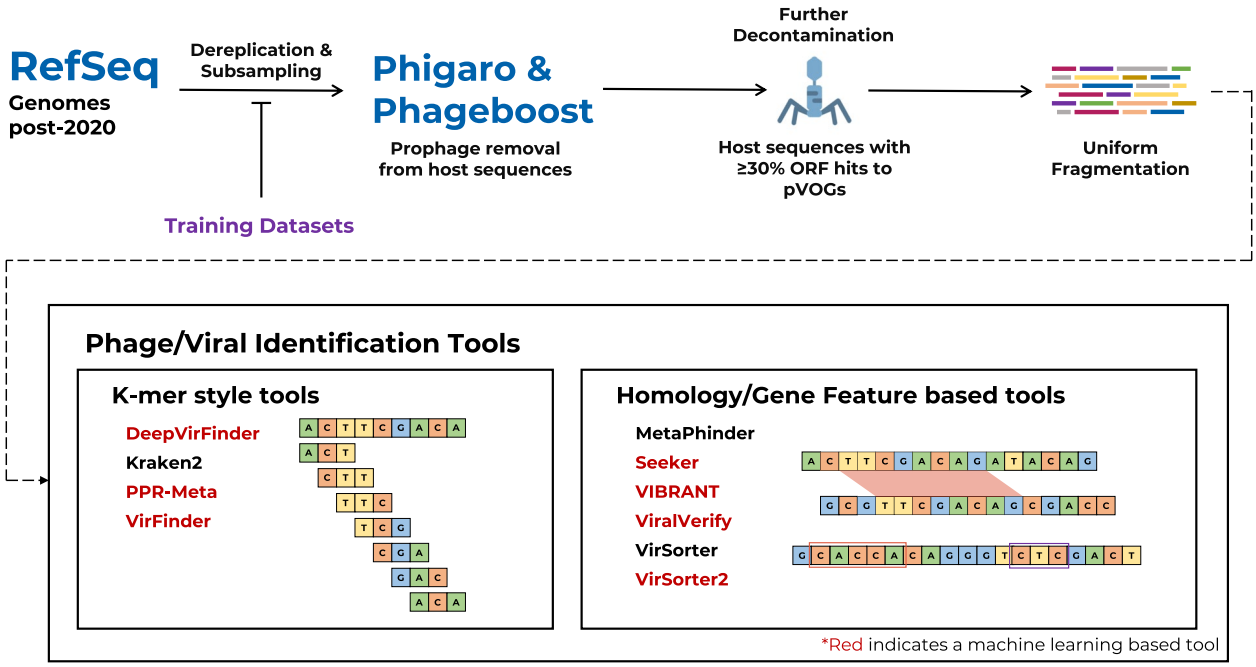

Fig. 1 Overview of RefSeq benchmarking workflow. All bacterial and archaeal chromosomes and plasmids and phage genomes that were deposited in the RefSeq database between 1 January 2020 and 12 August 2021 inclusive were downloaded. The phage genomes were used to create a positive test set and the chromosomes and plasmids for a negative set. The sequences were dereplicated with the training sets for each machine/deep learning tool that was benchmarked (highlighted in red), as well as any RefSeq sequences deposited prior to 2020. The negative set was down sampled to produce a positive:negative ratio of approximately 1:19 to replicate a typical gut microbiome. Prophages were identified and removed with Phigaro and PhageBoost. Any host sequences with greater than 30% of open read frames having hits to the Prokaryotic Virus Orthologous Groups database were then removed. All sequences were then uniformly fragmented into artificial contigs with lengths between 1 and 15 kbp. All identification tools were then run on the artificial contig sets

0.88 but higher recall at 0.96. The majority of the remaining tools performed well, with six tools (DeepVirFinder, Kraken2, MetaPhinder, VirFinder, and VirSorter) having F1 scores of over 0.83. Kraken2 had a precision score of almost 1, with only 2 chromosomal fragments and no plasmid fragments flagged as viral whilst still correctly identifying over 5000 phage fragments. These chromosomal fragments were originally part of the complete chromosomes of *Streptomyces albidoflavus* strain J1074/R2 and *Saccharolobus shibatae* strain BEU9, and Kraken2 identified them as *Streptomyces* phiC31 phage and *Sulfolobus* virus 1, respectively. Global alignment of the fragments to their references, as identified by Kraken2, revealed limited homology between them and suggests that these fragments are true- and false-positive predictions (Additional file 3). ViralVerify had a low precision score of 0.55, whilst its recall score of 0.88 was comparable to the other tools. Seeker had poor performance in both precision (0.48) and recall (0.41) compared to the other tools. Generally, k-mer-based tools performed better than reference similarity/gene-based tools, although the sample size of the investigated tools is too small to draw statistically significant conclusions. Across our benchmark, only 0.06% (4/6665) of phage contigs from our positive set were not detected by any tool, with 89.7% being identified by over half the tools, and 11.6% (775/6665) found by all 10 tools (Additional file 4). There was no significant taxonomic bias in the identification of bacteriophages between the tools, with the exception of Seeker which classifies less Ackermannviridae fragments than the other tools (Additional file 5).

The taxonomy of false-positive viral predictions varied between tools and between the chromosomal and plasmid dataset. Nearly half (44.7%) of DeepVirFinder's chromosomal false-positive hits (FPHs) belonged to the Clostridia despite the class only making up 1.27% of the dataset (Additional file 6). False-positive predictions by Seeker on the chromosomal dataset was biased towards bacilli (30.0% of FPHs versus 13.8% of the chromosomal dataset) and Actinomycetes (29.2% of FPHs versus 16.1% in chromosomal dataset). For all other tools, except for Kraken2, the false-positive taxonomic profiles roughly matched with the overall distribution in the dataset. Seeker's FPHs on the plasmid dataset were dominated by fragments belonging to the Halobacteria class (75.6% of FPHs versus 3.63% of the plasmid

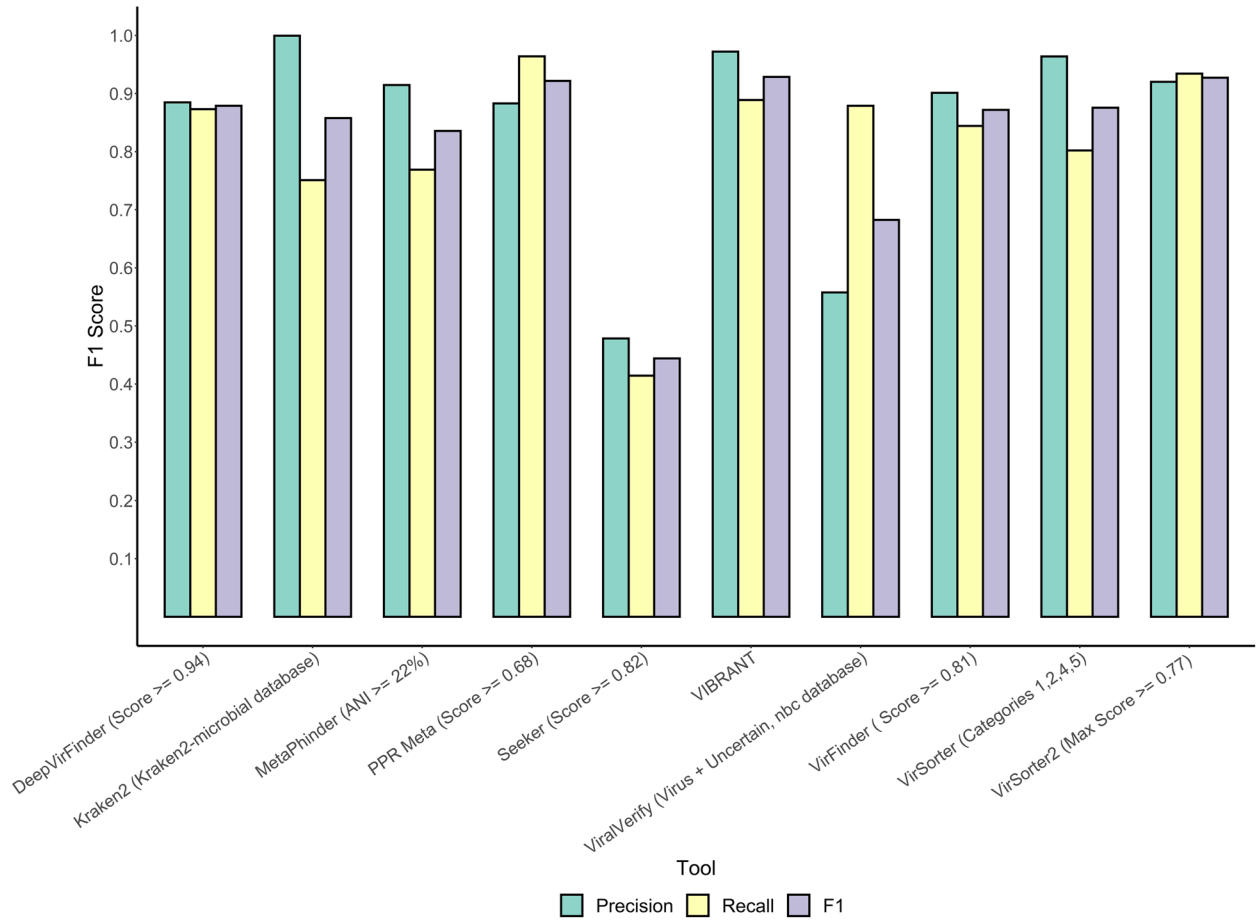

Fig. 2 Comparison of viral identification tools on artificial RefSeq contigs. Contigs were generated by randomly fragmenting complete bacterial/archaeal/phage genomes and plasmids deposited in the NCBI Reference Sequence Database (RefSeq) between 1 January 2018 and 2 July 2020, to a uniform distribution. Each tool was then separately run on the true positive (phage genome fragments) and negative (bacterial/archaeal chromosome and plasmid fragments) datasets. For tools which score/probability threshold or categories could be manually adjusted, values/categories were selected based on optimal F1 scores

dataset) (Additional file 6). VirFinder also had a bias towards Halobacteria plasmid fragments with it making up 20.0% of its FPHs. Plasmid fragments of the Mollicutes class were overrepresented in the false-positive hits of every tool except Kraken2 and ViralVerify (virus), which had zero and two false-positive hits, respectively. This overrepresentation was particularly pronounced in PPR-Meta (10.5% of its FPHs); VIBRANT (5.66% of its FPHs); VirSorter all categories (11.4% of its FPHs); VirSorter categories 1, 2, 4, and 5 (14.2% of its FPHs); and VirSorter2 (8.2% of its FPHs), whereas Mollicutes made up only 0.36% of the total plasmid fragments. These percentages must be taken in the context of each tool's overall performance as the total number of FPHs vary greatly between the tools — ranging from 2 to 4331 in the chromosomal dataset and between 0 and 313 in the plasmid dataset.

Benchmarking tools with randomly shuffled sequences

To serve as a further negative control, the positive RefSeq benchmark contigs were randomly shuffled at a nucleotide level to produce sequences that should not be identified as viral by any of the tools. Of the tools tested, four identified zero-shuffled contigs (MetaPhinder, viralVerify, VirSorter, and VirSorter2), Kraken2 classified three contigs, DeepVirFinder and VirFinder detecting 742 and 1070, respectively, and with the rest of the tools identifying over 2500 shuffled contigs as viral, including PPR-Meta which incorrectly classifying 99.2% (6608/6664) of all the shuffled contigs as viral (Table 3).

Artificial RefSeq phage contigs generated in the previous benchmark were randomly shuffled whilst preserving the dinucleotide distribution using esl-shuffle from the HMMER3 suite. Phage detection tools were then run on the shuffled contigs, and any positive hits were recorded.

Tools in bold type indicate methods utilising machine/deep learning models.

Benchmarking tools with mock community shotgun metagenomes

We next sought to compare these tools on real community shotgun metagenomic contigs. Thus, we obtained sequencing data of an uneven mock community created by Kleiner et al. [56], containing 32 species from across the tree of life, including five bacteriophages, at a large range of cell abundances (0.25–21.25%; Additional file 7). Integrated prophages were removed with Phigaro and PhageBoost to prevent them from being flagged as viral. No contigs belonging to the five bacteriophages were removed during this process.

This allowed us to assess the performance of our tools on real data whilst retaining knowledge of the ground truth (sample composition) and determine each tool's detection limit on low abundance species. The optimised parameters found in the RefSeq benchmark were used for each tool, with the exception of viralVerify and VirSorter. These tools have categorical thresholds which drastically change the profile of identified viral contigs, and the F1 scores between the thresholds were very close, so the parameters were further analysed in this benchmark. In general, the tools' F1 scores were considerably lower on this dataset than on the RefSeq artificial contigs, with F1 scores dropping by an average of 40.6%, compared to the RefSeq benchmark (Fig. 3). Kraken2 outperformed all other tools with a F1 score of 0.86, 0.3 higher than DeepVirFinder in second place. This is due to its high precision of 0.99 whilst identifying 76% of the phage contigs. DeepVirFinder had a high recall rate (0.80), but unlike in the RefSeq benchmark, had a lower precision of 0.42. Several other tools had similar results, with PPR-Meta, MetaPhinder, and VirFinder all achieving high precisions of 0.91, 0.98, and 0.83, respectively, but having comparatively low precision scores (0.24, 0.28, and 0.35, respectively). VirSorter (categories 1, 2, 4, and 5) and Seeker had comparable performances to the tools above, with them both having F1 scores of 0.44. Seeker, along with Kraken2, were the only tools to attain similar F1 scores in this benchmark to the RefSeq benchmark. VirSorter2, which performed best in the RefSeq benchmark, had a lower F1 score (0.36) than its predecessor mainly due to its high number of false-positive hits. VIBRANT, which also performed well in the previous benchmark, again had poor precision and middling recall resulting in a F1 score of 0.32. viralVerify (NBC database, virus only) had both low precision (0.24) and recall (0.22) resulting in the lowest F1 score (0.22) in this test. K-mer tools on average had a higher F1 score than the reference similarity/gene-based tools, but this difference was not statistically significant due to the small sample size of the number of tools in this study ($p = 0.094$, one-tailed Welch's *t*-test).

Out of the four DNA phage species found in the assemblies (Phage F2 was not sequenced due to being a ssRNA virus), only MetaPhinder was able to detect M13, whilst F0, the most abundant phage species, was detected by all benchmarked tools in all samples. MetaPhinder identified contigs belonging to all four phage species in two samples and three in the other sample. PPR-Meta, VIBRANT, VirSorter, and VirSorter2 were able to identify contigs belonging to three species in all three samples. viralVerify and VirFinder were able to identify the three phage species in two out of three samples, missing out on contigs belonging to phage ES18. DeepVirFinder and Kraken2 classified viral contigs belonging to three phage species in one out of three samples and detected two species in the other samples. Seeker was only able to identify contigs belonging to the most abundant phage F0. No correlation was found between F1 score and the number of phage strains detected ($R_s = -0.371$, $p = 0.29$). However, a positive, but not statistically significant, correlation, was observed between tools that identified more contigs of viral origin (true positives + false positives) and the number of phage strains identified ($R_s = 0.604$, $p = 0.06$).

Table 3 Performance of tools on randomly shuffled artificial phage contigs

Tool	False positives
DeepVirFinder	11.13%
MetaPhinder	0.00%
VIBRANT	58.76%
ViralVerify	0.00%
VirSorter	0.00%
VirSorter2	0.00%
Seeker	38.99%
PPR Meta	99.16%
VirFinder	16.06%
Kraken2	0.05%

Impact of tool prediction on diversity metric estimation

To test the impact of these tools on diversity estimations, four simulated mock community metaviromes containing an average of 719 viral genomes were retrieved from [57]. Reads were mapped to contigs (>1 kbp) that were identified as viral by each tool, and these mapped reads were then mapped to a set of population contigs to estimate their abundance in each sample. Original reads were also directly mapped to the population contigs as

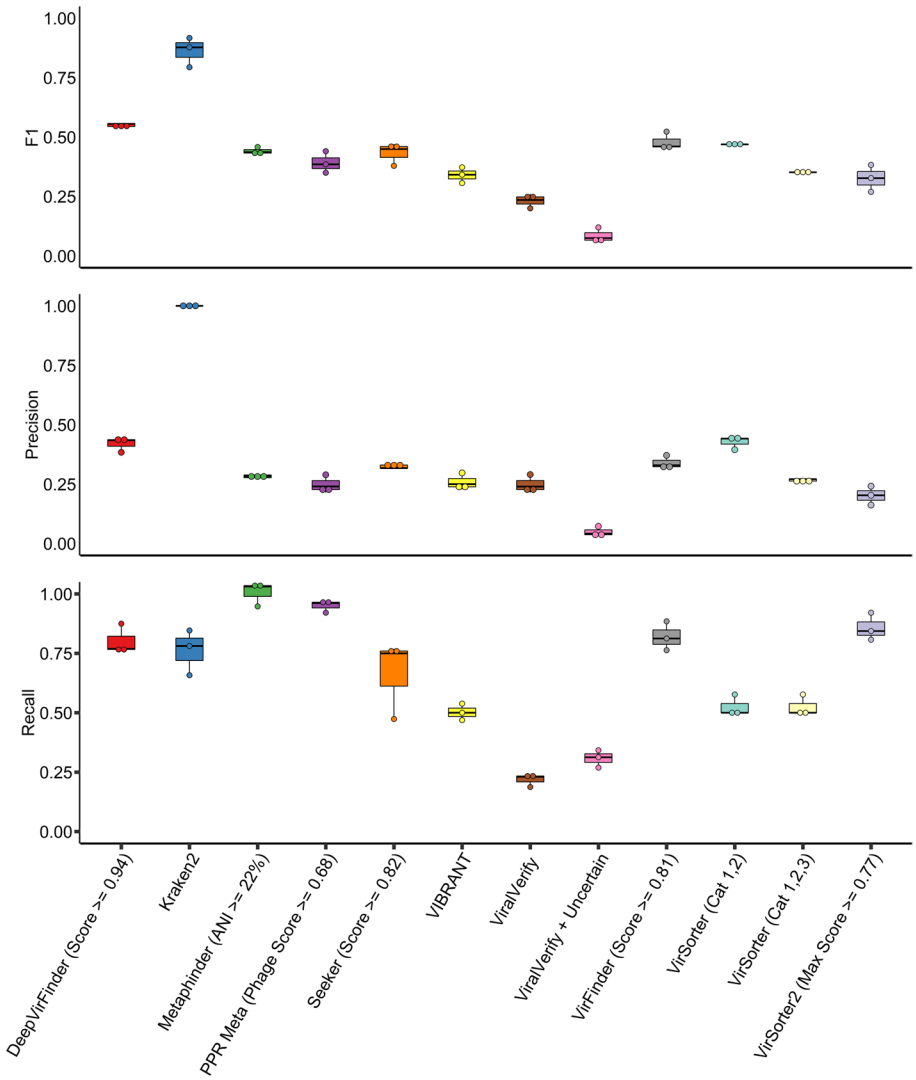

Fig. 3 Comparison of viral identification tools on uneven mock community samples. Mock community reads were retrieved from a previous study [56] and assembled with metaSPAdes. Prophages were detected and removed with Phigaro and PhageBoost before running each identification tool using optimal thresholds based on previous benchmarks except for viralVerify and VirSorter. F1 score, precision and recall metrics are displayed as separate panels. Each sample is plotted as a single point for each tool, with a boxplot indicating the interquartile ranges, extremes and mean of all three samples

a control. Read counts were then normalised by mapped contig lengths and sample library size, which Roux et al. [57] found to be reliable normalisation method. Diversity estimation metrics were then calculated using the normalised population counts. All tools returned fewer genomes per sample compared to the initial population, although there was significant variation between tools. PPR-Meta, MetaPhinder, and Kraken2 retrieved the greatest percentage of genomes with 86.8%, 89.1%, and 83.7% respectively (Fig. 4A). All other tools were able to retrieve more than 50% of the genomes with the exception of Seeker and viralVerify, which were only able to recover 32.4% and 41.3%, respectively, of the population genomes. All Shannon's alpha diversities calculated from the count matrices of each tool were within 10% of the default population with the exception of Seeker, whose *H* score was on average 27.0% lower (Fig. 4B). Simpson alpha-diversity indices showed similar performance, with all tools having a diversity score within 1% of the initial population, with the exceptions of DeepVirFinder and Seeker, who were 1.4% and 5.1% divergent, respectively (Fig. 4C). PPR-Meta was the only tool to estimate a comparatively higher alpha diversity than the default population. For beta diversity, pairwise Bray–Curtis dissimilarities within a sample were small between all tools except Seeker, whose analysis of similarity (ANOSIM)

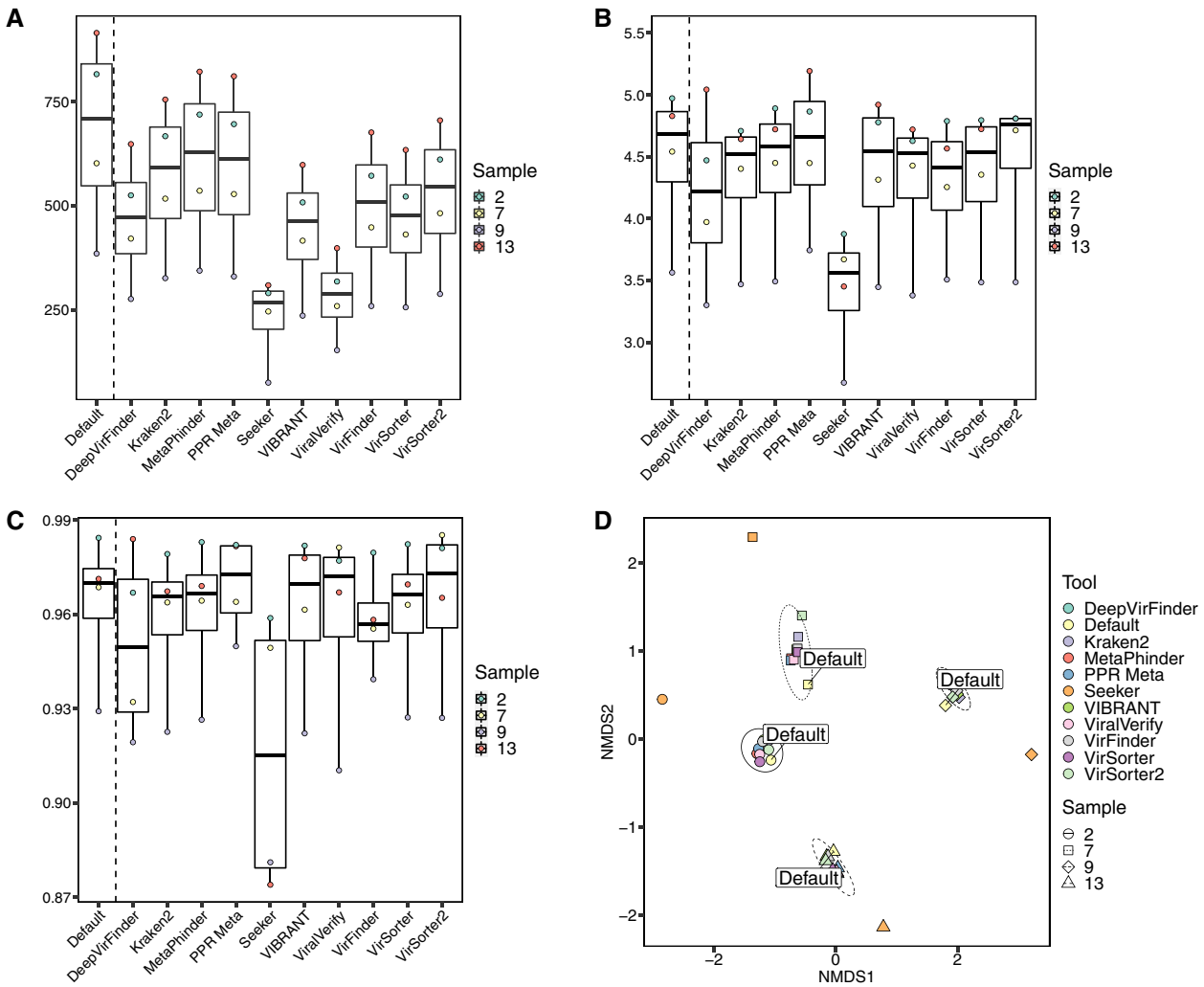

Fig. 4 Estimation of diversity metrics of tool-predicted virome populations. To assess the impact of each tool on population diversity, four simulated virome assemblies from Roux et al. [57] were downloaded. Each programme was then run to determine the subset of predicted viral contigs. Reads were mapped to these contig subsets, and mapped reads were then subsequently mapped to a pool of population contigs. All diversity metrics were computed by the R package "vegan". "Default" in each plot indicates each sample's original assembly. **A** Number of genomes observed from read mapping to predicted viral contig populations for each tool. **B** Comparison of estimated Shannon diversity indices from each tool's virome subset. Estimations are based on read counts that were normalised by contig size and sequencing depth of the virome. **C** Comparison of Simpson diversity indices from each tool's virome subset. **D** Nonmetric multidimensional scaling (NMDS) ordination plot of Bray-Curtis dissimilarity of virome subsets predicted by each viral identification tool. Ellipses indicate the 95% confidence interval for each sample cluster's centroid. Samples are represented by the same symbol and ellipse line type; tools are denoted by colour

showed significant dissimilarity when compared to other tools ($r = 0.3926$, $p = 0.0019$ with Benjamini–Hochberg correction for multiple comparisons at $FDR = 0.05$) (Fig. 4D; Additional file 8).

Runtime and computational load of each tool

We also recorded the running times of each tool on the RefSeq-positive dataset on a high-performance cluster (16 VCPU, 108-Gb RAM) (Fig. 5). Kraken2 and VirFinder were the fastest tools finishing in under 5 min. DeepVirFinder, MetaPhinder, PPR-Meta, and Seeker finished in under half an hour with viralVerify finishing just over that mark. VirSorter and VirSorter2 took the longest time to run on this dataset (2.9 h and 3.8 h to completion, respectively).

Discussion

Bacteriophages are crucial members of microbial communities in nearly every ecosystem on Earth, responsible for controlling host population size as well as having

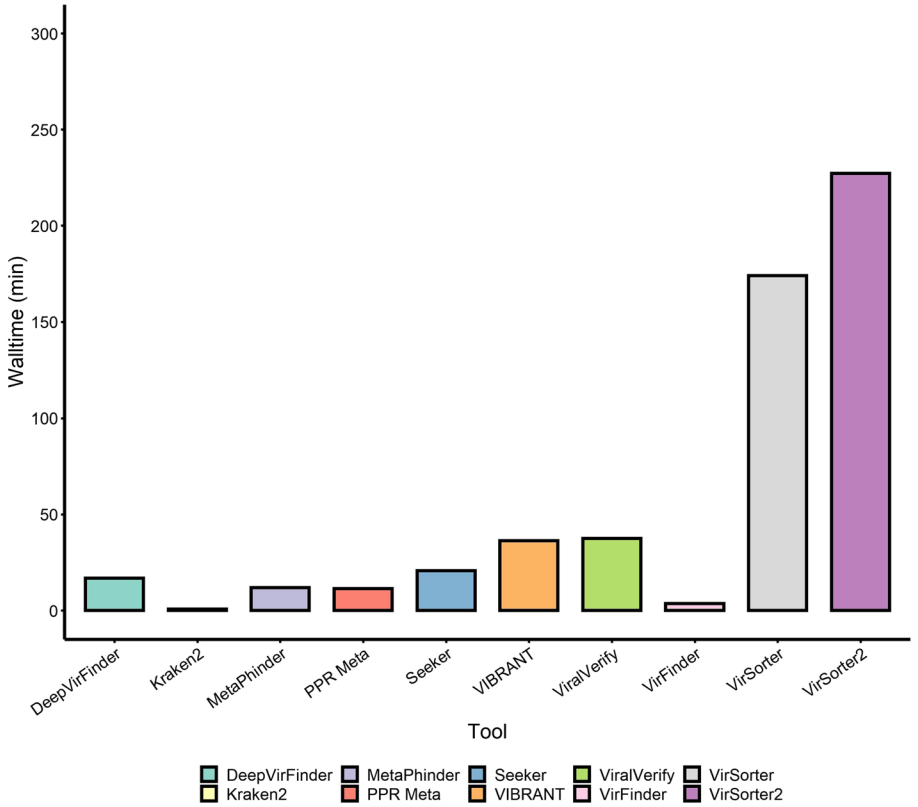

Fig. 5 Comparison of tool runtimes on the positive RefSeq artificial contig set. Wall runtime of each tool on mock community samples was recorded on a 16 VCPU, 108-GB RAM, and Linux high-performance cluster without GPU acceleration. Tools were run with 16 threads where it could be set as a parameter (all tools except MetaPhinder, PPR-Meta, and Seeker). The RefSeq-positive set contains approximately 53.4 million bp

wider impacts on community functions. Tools designed to recover viral sequences from mixed community metagenomic and virome samples are fundamental to studying the role of bacteriophages in the wider context of their environment. Advancements in this field have produced an extensive suite of viral identification tools that each claim to improve on the performance of similar tools. Selecting which tool among these is ideal for a dataset is thus not straightforward, especially as each novel tool typically only benchmarks against two or three other existing tools. Most tools developed for this purpose, especially those released in recent years, have utilised machine/deep learning to classify sequences, whereas others rely on categorising sequences based on their similarity to sequences in databases. Both these approaches have potential to improve over time with newly discovered viral genomes being added to training datasets and databases.

Here, we compare ten methods for identifying viral sequences from metagenomes across three datasets. We first benchmarked the tools on positive and negative datasets to evaluate their performance on an ideal set of contigs (size between 1 and 15 kbp, without misassemblies) and determine approximate optimal thresholds. Most tools performed well here, detecting the majority of phage sequences, whilst keeping false positives low. PPR-Meta, VIBRANT, and VirSorter2, which all use different machine learning methods, had the best performance across the tools. Generally, k-mer tools outperformed reference similarity and gene-based tools. Whilst the optimal thresholds that we determined may not necessarily be ideal for all other datasets, we believe they can be used as a basis for further usage of these tools as in each case they produced considerably better results than the default parameters. We therefore encourage researchers to apply these thresholds and parameters within the context of their prospective dataset. Phage taxonomy did not affect viral prediction of the tools, whilst chromosomal and plasmid taxonomy revealed biases in several tools. Therefore, when analysing metagenomes, it is important to consider what phage hosts can be expected to be present in the sampled microbial ecosystem and whether the bias of the selected viral prediction tool will affect downstream analyses.

A stark contrast between machine learning and more traditional tools can be seen when analysing their

identification of randomly shuffled phage sequences. Of the six tools that utilise machine/deep learning methods, five identified a significant proportion of the sequences as viral, with only VirSorter2 as the only exception, probably due to its classifier being trained on a range of sequence and gene features. Three of the other four tools returned zero false positives, and Kraken2 only returned three. This highlights that whilst these machine/deep learning algorithms have the capability to detect novel phages, their performance may be unpredictable when exposed to novel data with features that differ from those in our training sets.

When tested on real metagenomic data, most tools performed significantly worse than in the RefSeq benchmark, with the exceptions of Kraken2 and Seeker. Generally, k-mer tools had a smaller drop in F1 score from their RefSeq benchmark compared to reference similarity/gene-based tools. This is probably due to the comparatively shorter phage sequences that were assembled from the metagenomic reads, which provide less genetic context, thereby negatively impacting the algorithms of the reference similarity and gene-based tools. MetaPhinder was the only tool which was able to detect the one contig of phage M13 that was assembled in each sample. Unfortunately, this is likely due to MetaPhinder's low precision in this benchmark resulting in many false-positive calls. Most other tools were able to identify contigs belonging to the other three phages, with the exception of Seeker which could only identify the most abundant phage F0 in each sample. This suggests that when analysing metagenomic datasets where phage species are likely to be in low abundance, k-mer-based tools such as Kraken2 and DeepVirFinder are the best choices, the former being the favoured option when precision is of particular importance, whilst the latter's use is appropriate if the discovery of novel phages is of interest due to its deep learning algorithm.

We also gauged any potential biases and impact these tools may have on the diversity of its predicted viral population. Most tools performed well with alpha-diversity indices within 10% of the default population with the exception of Seeker which returned a considerably lower value due to the very low number of viral population genomes Seeker originally predicted. Some tools such as PPR-Meta predicted higher alpha diversity than default population. This is due to the tools missing some high abundance genomes from their predictions, resulting in a more even diversity distribution. When evaluating beta diversity, Seeker was the only tool that produced results that had significant dissimilarity from the other tools and did not cluster with the other programmes, again as a result of the low proportion of genomes it recovered in this dataset. Hence, beta-diversity trends of the tools examined here, with the exception of Seeker, are accurate to the original population, even when only half the genomes are recovered.

Runtime and computational load are also important factors to examine, since these can become practical limitations if large samples take many hours or days to be analysed. Most tools were reasonably fast, although VirSorter, and its successor VirSorter2, took multiple hours to complete their runs. It is important to note that VIBRANT, VirSorter, and VirSorter2 annotate the identified viral genomes and predict prophages which come at the expense of runtime, although these can be useful for some applications. Kraken2 was by far the fastest tool, taking less than a minute to run on our dataset. However, Kraken2 requires very high RAM use compared to the other benchmarked tools so it may not be feasible for researchers with limited computing power.

Although these benchmarks comprehensively compared the performance of state-of-the-art tools, there are a number of limitations with our study. First, whilst we use RefSeq genomes, and a mock metagenomic community to benchmark these tools, we do not address the tools' ability to identify viral sequences belonging to different phage families. Secondly, we used the default database(s) or the original trained model(s) that was provided with each tool. Whilst providing each tool the same database, or dataset to be trained on, may have been a fairer comparison of the underlying algorithms, this was beyond the scope of our study. We note that most routine users are also unlikely to retrain these tools prior to their use. Thirdly, we did not assess the performance of combining multiple tools, which could provide meaning insights that would be missed when only one single tool is used, as in Marquet et al. [58] where the authors combined multiple tools into a single workflow. Fourth, many of the tools have additional functionalities, which we did not benchmark here but may nevertheless impact a researcher's choice of tool such as prophage prediction (VIBRANT, VirSorter, VirSorter2); plasmid prediction (PPR-Meta); taxonomic identification (Kraken2); and functional annotation (VIBRANT). Finally, a few recently developed tools we found during our study were not included in our benchmarking either due to (1) requiring the use of its own web server and therefore not being scalable (VIROME, VirMiner), (2) lack of clear installation/running instructions (ViraMiner), or (3) errors when attempting to use the tool, which we were unable to resolve (PhaMers, VirNet, VirMine).

Conclusion

Our comparative analysis of ten currently available metagenomic virus/phage identification tools provides valuable metrics and insights for other investigators to

use and build on. Using mock communities and artifical datasets, precision, recall, and biases of these tools could be calculated. By adjusting the filtering thresholds for viral identification for each tool and comparing F1 scores, we were able to optimise performance in every case. In the artificial RefSeq contig benchmark, most tools performed well, with PPR Meta, VIBRANT and VirSorter2 having the highest F1 scores. In the mock uneven community dataset, tools generally performed worse with the exception of Kraken2, whose performance included almost perfect precision with above average recall scores. All tools except Seeker were able to produce a diversity profile with similar indices to the original virome population and are therefore suitable for phage ecology studies. We suggest that of the currently available metagenomic phage identification tools, Kraken2, should be considered when researchers are trying to identify previously characterised phages. When novel phage detection is required, Kraken2 should be used in combination with tools such as VirSorter2 and DeepVirFinder.

Materials and methods

Benchmarking with RefSeq dataset

Complete bacterial and archaeal chromosome and plasmid sequences, and phage genomes deposited in RefSeq [59] since 1 January 2020 (inclusive), were downloaded on 12 August 2021 to construct a benchmarking set. Sequences in this set with ≥ 95% identity to the pre-2020 RefSeq sequences and training datasets for DeepVirFinder, PPR Meta, Seeker, VIBRANT, VirFinder, and VirSorter were removed with dedupe.sh [60] to reduce any potential overfitting. Chromosome and plasmid sequences were then randomly down sampled by a factor of 14.3, using reformat.sh (from BBTools suite) [60], to produce a host:phage ratio of approximately 19:1. Phigaro (v2.3.0, default settings) [54] and PhageBoost (v0.1.7, default settings) [55] were run in succession on the chromosomal and plasmid sequences to remove prophage sequences. Host sequences with ≥ 30% open reading frames (ORFs) with HMM hits to the pVOG database were removed as contamination. All sequences were then uniformly fragmented to between 1 and 15 kbp, using a custom python script (available at https://github.com/sxh1136/Phage_tools), to create artificial contigs. Each viral prediction tool was then run on the three sets of contigs (chromosome, plasmid, and phage) with default settings except for VIBRANT and VirSorter where the phage-derived contig set was additionally run using their virome decontamination modes, due to their potentially improved performance in datasets consisting of mainly viral fragments. Commands used to run each tool and their version numbers can be found at https://github.com/sxh1136/Phage_tools/blob/master/manuscript_tools_script.md. RefSeq benchmarking datasets are available for download and use at https://figshare.com/articles/dataset/RefSeq_Datasets_for_benchmarking_-_Ho_et_al_/19739884.

For tools where score/probability thresholds can be manually adjusted (DeepVirFinder, MetaPhinder, PPR Meta, Seeker, and VirFinder), F1 curves were plotted (100 data points), and optimal thresholds were determined by maximal F1 score. For viralVerify and VirSorter, which have categorical thresholds for phage identification, the category sets with the highest F1-score were plotted; for VirSorter, two category sets were compared, categories 1, 2, 4, and 5 and all categories, as these are commonly used. ViralVerify was also additionally run using both Pfam-A 33.0 and a provided database of virus/chromosome-specific HMMs as these are listed on the tool's GitHub usage guide. Kraken2 was run with the pre-built kraken2-microbial database, available at https://lomanlab.github.io/mockcommunity/mc_databases.html. The reference genomes of the two chromosomal fragments which Kraken2 identified as viral were downloaded from NCBI RefSeq (NC_CP059254.1, and NZ_CP07771.1). The fragments and the reference genomes were annotated with pharokka [61] and globally aligned and visualised with clinker [62]. Taxonomy of tool true-positive and false-positive viral predictions was retrieved using the R package taxonomizr v 0.10.3 [63].

Run time of each tool on this dataset was recorded using a Linux virtual machine provided by Cloud Infrastructure for Big Data Microbial Bioinformatics (CLIMB-BIG-DATA), with the following configuration: CPU: Intel® Xeon® Processor E3-12xx v2 (8 VCPU), GPU: Cirrus Logic GD 5446, and memory: 64-GB multi-bit ECC. Tools were run with 16 threads where it could be set as a parameter (all tools except MetaPhinder, PPR-Meta, and Seeker).

Benchmarking with randomly shuffled sequences

All artificial phage contigs created in the previous benchmark were randomly shuffled whilst preserving dinucleotide distribution using esl-shuffle from the HMMER3 suite (v3.3.2, -d) [64]. Each identification tool was then run on the randomly shuffled sequences using the optimised thresholds that were determined in the RefSeq benchmark, and false positives were recorded.

Benchmarking with mock community metagenomes

Three shotgun metagenomic sequencing replicates of an uneven mock community [56] were retrieved from the European Nucleotide Archive (BioProject PRJEB19901). These communities contain five phage strains: ES18 (H1), F0, F2, M13, and P22 (HT105). The quality of the data was checked using FASTQC (v0.11.8, default settings)

[65], and overrepresented sequences were removed with cutadapt (v2.10, –max-n 0) [66]. Cleaned paired-end reads were then assembled with MetaSPAdes (v3.14.1, default settings) [67], and contigs < 1 kbp were removed. Prophages were then removed from the contigs by sequentially running Phigaro and then PhageBoost, as with the RefSeq chromosomal and plasmid datasets. Each tool was then run on the three sets of contigs using optimal parameters as determined previously with the exception of viralVerify and VirSorter where all categorical thresholds were re-evaluated. MetaQUAST (v5.0.2, default settings) [68] was used to map contigs to reference phage genomes and calculate coverage.

Benchmarking with simulated mock virome communities

Four mock communities (samples 13, 2, 7, and 9) containing between 500 and 1000 viral genomes created by Roux et al. [57] were selected for analysis. These samples belonged to four different beta diversity groups and did not share any of their 50 most abundant viruses. Each simulation of 10 million paired-end reads were quality controlled with Trimmomatic [69] and assembled with MetaSPAdes by Roux et al. [57]. The contigs were then downloaded for benchmarking. As before, contigs with length < 1 kbp were removed and then inputted into each viral identification programme. Positive viral contig sets for each tool were then extracted, and reads were mapped to these with BBMap [60] with ambiguous mapped reads assigned to contigs at random (ambiguous = random), as in Roux et al. [57]. Primary mapped reads with pairs mapping to the same contig (options -F 0×2 0×904) were then extracted with SAMtools (v1.11) and mapped to a pool of non-redundant population contigs. This pool was created by clustering all four samples with nucmer (v3.1) [70], at ≥ 95% ANI (average nucleotide identity) across ≥ 80% of their lengths. Abundance matrices for each tool were calculated by normalising read counts by contig length and total library size to produce counts per million (CPM). These abundance matrices were then used to calculate Shannon, Simpson, and Bray–Curtis dissimilarity indices using the vegan package (v2.5.7) [71]. Nonmetric multidimensional scaling (NMDS) and analysis of similarity (ANOSIM) were also computed with vegan. ANOSIM p-values were corrected with the Benjamini–Hochberg method [72]. Seed and permutations were set as 123 and 9999, respectively, where possible. All plots were generated with ggplot2 (v3.3.2) [73] and arranged with ggarrange from ggpubr (v0.4.0). [74].

Abbreviations

ANOSIM	Analysis of similarity
bp	Base pairs
CNN	Convolutional neural networks
CPM	Counts per million
FPHs	False-positive hits
GPU	Graphics processing unit
HMM	Hidden Markov model
kbp	Kilobase pairs
LSTM	Long short-term memory models
MDA	Multiple displacement amplification
NCBI	National Center for Biotechnology Information
NMDS	Nonmetric multidimensional scaling
RefSeq	NCBI Reference Sequence database
ssDNA	Single-stranded DNA
VCPU	Virtual central processing unit

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-023-01533-x.

Additional file 1: Supplementary Fig. 1. Optimisation curves for tools resulting in score/probability thresholds.

Additional file 2: Supplementary Fig. 2. F1-score plots of tools that provide categorical thresholds.

Additional file 3: Supplementary Fig. 3. Comparison of artificial RefSeq contigs with reference genomes.

Additional file 4: Supplementary Fig. 4. Upset plot of RefSeq phage artificial contigs predicted as viral by each tool.

Additional file 5: Supplementary Fig. 5. Taxonomic analysis of each tool's viral prediction on RefSeq phage fragments.

Additional file 6: Supplementary Fig. 6. Taxonomy of false positive viral predictions on RefSeq chromosomal and plasmid fragments.

Additional file 7: Supplementary Table 1. Bacteriophage strain composition of uneven mock community from Kleiner et al. [56].

Additional file 8: Supplementary Table 2. Analysis of Similarity (ANOSIM) between non-metric multidimensional scaling of tools.

Acknowledgements

We wish to thank the authors of Kleiner et al. [56] and Roux et al. [57] for making their sequencing data publicly available, allowing their use in this study.

Authors' contributions

The study was designed by all authors. SFSH analysed and interpreted the data generated in this study. All authors wrote and approved the manuscript.

Funding

SFSH was supported by the Wellcome Trust Antimicrobials and Antimicrobial Resistance Doctoral Training Programme (215154/Z/18/Z). ADM was supported by MRC (MR/T030062/1). WvS was supported by a Royal Society Wolfson Research Merit Award (WM160092) and BBSRC (BB/S017941/1).

Availability of data and materials

RefSeq dataset construction and example commands used to run each tool are available at https://github.com/sxh1136/Phage_tools. RefSeq benchmarking datasets are available for download and use at https://figshare.com/articles/dataset/RefSeq_Datasets_for_benchmarking_-_Ho_et_al_/19739884.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 11 June 2022 Accepted: 22 March 2023
Published online: 21 April 2023

References

1. Parikka KJ, Romancer ML, Wauters N, Jacquet S. Deciphering the virus-to-prokaryote ratio (VPR): insights into virus–host relationships in a variety of ecosystems. Biol Rev. 2017;92:1081–100.
2. CobiánGüemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F. Viruses as winners in the game of life. Annu Rev Virol. 2016;3:197–214.
3. Hoyles L, McCartney AL, Neve H, Gibson GR, Sanderson JD, Heller KJ, et al. Characterization of virus-like particles associated with the human faecal and caecal microbiota. Res Microbiol. 2014;165:803–12.
4. Silveira CB, Rohwer FL. Piggyback-the-winner in host-associated microbial communities. Npj Biofilms Microbiomes. 2016;2:1–5.
5. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol. 2018;3:870–80.
6. Jiao N, Herndl GJ, Hansell DA, Benner R, Kattner G, Wilhelm SW, et al. Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. Nat Rev Microbiol. 2010;8:593–9.
7. Rohwer F, Thurber RV. Viruses manipulate the marine environment. Nature. 2009;459:207–12.
8. Brown-Jaque M, Calero-Cáceres W, Muniesa M. Transfer of antibiotic-resistance genes via phage-related mobile elements. Plasmid. 2015;79:1–7.
9. Chiang YN, Penadés JR, Chen J. Genetic transduction by phages and chromosomal islands: the new and noncanonical. PLoS Pathog. 2019;15:e1007878.
10. McInnes RS, McCallum GE, Lamberte LE, van Schaik W. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. Curr Opin Microbiol. 2020;53:35–43.
11. Sutton TDS, Hill C. Gut bacteriophage: current understanding and challenges. Front Endocrinol. 2019;10:00784.
12. Campbell A. The future of bacteriophage biology. Nat Rev Genet. 2003;4:471–7.
13. Hobbs Z, Abedon ST. Diversity of phage infection types and associated terminology: the problem with 'lytic or lysogenic.' FEMS Microbiol Lett. 2016;363:fnw047.
14. Walker AW, Duncan SH, Louis P, Flint HJ. Phylogeny, culturing, and metagenomics of the human gut microbiota. Trends Microbiol. 2014;22:267–74.
15. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-M, Quick J, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. JAMA. 2013;309:1502–10.
16. Lwoff A. Lysogeny. Bacteriol Rev. 1953;17:269–337.
17. Labonté JM, Swan BK, Poulos B, Luo H, Koren S, Hallam SJ, et al. Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. ISME J. 2015;9:2386–99.
18. Cenens W, Makumi A, Mebrhatu MT, Lavigne R, Aertsen A. Phage–host interactions during pseudolysogeny. Bacteriophage. 2013;3:e25029.
19. Ripp S, Miller RV. The role of pseudolysogeny in bacteriophage-host interactions in a natural freshwater environment. Microbiology. 1997;143:2065–70.
20. Shkoporov AN, Khokhlova EV, Fitzgerald CB, Stockdale SR, Draper LA, Ross RP, et al. ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects *Bacteroides intestinalis*. Nat Commun. 2018;9:1–8.
21. Ackermann HW, Audurier A, Berthiaume L, Jones LA, Mayo JA, Vidaver AK. Guidelines for bacteriophage characterization. Adv Virus Res. 1978;23:1–24.
22. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. Nat Protoc. 2009;4:470–83.
23. Probst AJ, Weinmaier T, DeSantis TZ, Domingo JWS, Ashbolt N. New perspectives on microbial community distortion after whole-genome amplification. PLoS ONE. 2015;10:e0124158.
24. Yilmaz S, Allgaier M, Hugenholtz P. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nat Methods. 2010;7:943–4.
25. Kim K-H, Bae J-W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. Appl Environ Microbiol. 2011;77:7663–8.
26. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. Open Biol. 2013;3:130160.
27. Kim M-S, Bae J-W. Lysogeny is prevalent and widely distributed in the murine gut microbiota. ISME J. 2018;12:1127–41.
28. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.
29. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic Acids Res. 2012;40:e126.
30. Fouts DE. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. Nucleic Acids Res. 2006;34:5839–51.
31. Arndt D, Marcu A, Liang Y, Wishart DS. PHAST, PHASTER and PHAST-EST: tools for finding prophage in bacterial genomes. Brief Bioinform. 2017;20:1560–7.
32. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics. 2008;24:863–5.
33. Roux S, Tournayre J, Mahul A, Debroas D, Enault F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics. 2014;15:76.
34. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, et al. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Stand Genomic Sci. 2012;6:427–39.
35. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8:90.
36. Guo J, Bolduc B, Zayed AA, Varsani A, Dominguez-Huerta G, Delmont TO, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome. 2021;9:37.
37. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data using deep learning. Quant Biol. 2020;8:64–77.
38. Amgarten D, Braga LPP, da Silva AM, Setubal JC. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. Front Genet. 2018;9:00304.
39. Jurtz VI, Villarroel J, Lund O, Voldby Larsen M, Nielsen M. MetaPhinder—identifying bacteriophage sequences in metagenomic data sets. PLoS ONE. 2016;11:e0163111.
40. Antipov D, Raiko M, Lapidus A, Pevzner PA. MetaviralSPAdes: assembly of viruses from metagenomic data. Bioinformatics. 2020;36:4126–9.
41. Deaton J, Yu FB, Quake SR. Mini-metagenomics and nucleotide composition aid the identification and host association of novel bacteriophage sequences. Adv Biosyst. 2019;3:1900108.
42. 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. 2019;6:giz066.
43. Auslander N, Gussow AB, Benler S, Wolf YI, Koonin EV. Seeker: alignment-free identification of bacteriophage genomes by deep learning. Nucleic Acids Res. 2020;48:e121.
44. Tampuu A, Bzhalava Z, Dillner J, Vicente R. ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. PLoS ONE. 2019;14:0222271.
45. 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;5:69.
46. Garretto A, Hatzopoulos T, Putonti C. virMine: automated detection of viral sequences from complex metagenomic samples. PeerJ. 2019;7:e6695.
47. 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;7:42.
48. Abdelkareem AO, Khalil MI, Elaraby M, Abbas H, Elbehery AHA. VirNet: deep attention model for viral reads identification. 2018 13th Int Conf Comput Eng Syst ICCES. 2018. p. 623–6.

49. Zhao G, Wu G, Lim ES, Droit L, Krishnamurthy S, Barouch DH, et al. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. Virology. 2017;503:21–30.
50. Ponsero AJ, Hurwitz BL. The promises and pitfalls of machine learning for detecting viruses in aquatic metagenomes. Front Microbiol. 2019;10:00806.
51. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:257.
52. Yutin N, Benler S, Shmakov SA, Wolf YI, Tolstoy I, Rayko M, et al. Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. Nat Commun. 2021;12:1044.
53. Shkoporov AN, Hill C. Bacteriophages of the human gut: the "known unknown" of the microbiome. Cell Host Microbe. 2019;25:195–209.
54. Starikova EV, Tikhonova PO, Prianichnikov NA, Rands CM, Zdobnov EM, Ilina EN, et al. Phigaro: high throughput prophage sequence annotation. Bioinformatics. 2020;36:3882–4.
55. Sirén K, Millard A, Petersen B, Gilbert MTP, Clokie MRJ, Sicheritz-Pontén T. Rapid discovery of novel prophages using biological feature engineering and machine learning. NAR Genomics Bioinforma. 2021;3:lqaa109.
56. Kleiner M, Thorson E, Sharp CE, Dong X, Liu D, Li C, et al. Assessing species biomass contributions in microbial communities via metaproteomics. Nat Commun. 2017;8:1558.
57. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017;5:e3817.
58. Marquet M, Hölzer M, Pletz MW, Viehweger A, Makarewicz O, Ehricht R, et al. What the Phage: a scalable workflow for the identification and analysis of phage sequences. Gigascience. 2022;11:giac110.
59. O'Leary NA, Wright MW Jr, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.
60. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. 2014. Available from: https://www.osti.gov/biblio/1241166.
61. Bouras G, Nepal R, Houtak G, Psaltis AJ, Wormald P-J, Vreugde S. Pharokka: a fast scalable bacteriophage annotation tool. Bioinformatics. 2023;39:btac776.
62. Gilchrist CLM, Chooi Y-H. Clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics. 2021;37:2473–5.
63. Sherrill-Mix S. Taxonomizr: functions to work with NCBI accessions and taxonomy. 2023. Available from https://github.com/sherrillmix/taxonomizr.
64. HMMER. Available from: http://hmmer.org/.
65. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC. 2019. Available from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
66. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.
67. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
68. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2016;32:1088–90.
69. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
70. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. PLOS Comput Biol. 2018;14:e1005944.
71. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, *et al*. Vegan: Community Ecology Package. Available from: https://CRAN.R-project.org/package=vegan
72. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57:289–300.
73. Wickham H. ggplot2: elegant graphics for data analysis. Available from: https://ggplot2.tidyverse.org
74. Kassambara A. ggpubr: "ggplot2" based publication ready plots. Available from: https://CRAN.R-project.org/package=ggpubr

Publisher's Note