Comprehensive benchmarking of tools to identify phages in metagenomic shotgun sequencing data

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

[1] Institute of Microbiology and Infection, College of Medical and Dental Sciences, University of Birmingham, Birmingham, United Kingdom

[2] Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom

[*] Corresponding author: w.vanschaik@bham.ac.uk.

Abstract

Background

As the relevance of bacteriophages in shaping diversity in microbial ecosystems is becoming increasingly clear, the prediction of phage sequences in metagenomic datasets has become a topic of considerable interest, which has led to the development of many novel bioinformatic tools. A comprehensive comparative analysis of these tools has so far not been performed.

Methods

We benchmarked ten state-of-the-art phage identification tools. We used artificial contigs generated from complete RefSeq genomes representing phages, plasmids, and chromosomes, and a previously sequenced mock community containing four phage strains to evaluate the precision, recall and F1-scores of the tools. In addition, a set of previously simulated viromes was used to assess diversity bias in each tool's output.

Results

DeepVirFinder performed best across the datasets of artificial contigs and the mock community, with the highest F1-scores (0.98 and 0.61 respectively). Generally, machine learning-based tools performed better on the artificial contigs, while reference and machine learning based tool performed comparably on the mock community. Most tools produced a viral genome set that had similar alpha and beta diversity patterns to the original population with the notable exception of Seeker, whose metrics differed significantly from the diversity of the underlying data.

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.

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 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 bacterial host's machinery to replicate, and subsequently spread via cell lysis. They can either be lytic or temperate, and while 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 the genome of the bacterial host as linear DNA or as a self-replicating autonomous plasmid. In addition, an alternative life cycle, termed pseudolysogeny, has been documented, during which neither phage genome replication nor prophage formation occurs [13].

Traditionally, phage identification and characterisation relied on isolation and culturing techniques, which are time-consuming and often fail to capture the full repertoire of phages in an ecosystem 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]. In metagenomic studies, researchers 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, due to their integration into their hosts' genome as prophages [16], attachment to their hosts' surface [17], or their presence 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. The isolation of viral particles frequently results in low DNA yields, leading to 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]. Ideally, the purification of viral particles will lead to a metavirome with very little host contamination.

However, it is very difficult to produce a viral fraction that is devoid of DNA originating from microbes that are present in the ecosystem [26]. 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.

Many tools for identifying viral sequences from mixed metagenomic and virome assemblies have been developed in the last five years (Table 1) and we will shortly discuss these here. 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 homology based tools (MetaPhinder [35] and VirusSeeker [36]) have been developed.

VirFinder was the first machine learning, reference-free viral identification tool, utilising k-mer signatures [37]. VirFinder had considerably better performance in recovering viral sequences than VirSorter, especially on shorter sequences (<5,000 bp). However, it displayed variable performance in different environments, perhaps due to biases introduced by the reference data used for training the machine learning model [38]. A number of machine/deep learning tools have since been published, including DeepVirFinder [39], which boasts increased viral identification at all contig lengths over its predecessor VirFinder, whilst mitigating the latter's biases by including various metavirome datasets that contain uncultivated viral sequences.

Other recent tools have started utilising alternative approaches. MARVEL [40], integrates the two approaches described above, using a random forest model to leverage sequence features (gene density and strand shifts) and homologies (hits to pVOGs (Prokaryotic Virus Orthologous Groups) database [41]). This allows the tool to identify metagenomic bins that resemble phages, with comparable specificity but improved sensitivity to VirSorter and VirFinder. This detection is however currently limited to phages of the *Caudovirales* order. VIBRANT also employs a hybrid machine learning and protein similarity approach but is able to recover a diverse array of phages infecting bacteria and archaea, including integrated prophages. In addition, it characterises auxiliary metabolic genes and pathways after phage identification.

MetaviralSPAdes [42] 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).

With the development of so many tools using a variety of approaches, a comprehensive comparison and benchmarking is needed to evaluate which tools are most useful 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 non-viral artificial contigs

Ten commonly used tools for viral sequence identification in metagenomes were selected for evaluation: DeepVirFinder; MARVEL; MetaPhinder; PPR Meta; Seeker; VIBRANT; ViralVerify; VirFinder; VirSorter; and VirSorter2. All of these tools can be run locally without the use of a web server; accept metagenomic contigs as input, except MARVEL which requires bins to be created first; and have been published in the past decade.

We first evaluated all the programs on the same uniform datasets. All complete phage genomes deposited in RefSeq between 1 January 2018 and 2 July 2020 later were downloaded, quality controlled, and fragmented to create a true positive set of artificial contigs. A negative set was constructed from all RefSeq bacterial and archaeal plasmids, and a random 1:10 subsample of all RefSeq bacterial and archaeal chromosomes, submitted in the same time frame. As chromosomes often have prophages integrated within them, which would cause tools to falsely identify some contigs as viral, we removed these with two state-of-the-art prophage detection tools, Phigaro [43] and PhageBoost [44]. The negative dataset is considerably larger than the true positive dataset as we wanted to consider the performance of these tools in metagenomic shotgun sequencing datasets which are typically dominated by non-viral sequences. All evaluated programs, except MARVEL, 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 VIBRANT, VirSorter and ViralVerify, the categories that returned the highest F1 score were used. In most tools there was a trade-off between precision and recall. This is likely due to relaxed thresholds allowing for more viral and non-viral sequences to be detected, increasing recall, and decreasing precision simultaneously. Additionally, for VIBRANT and VirSorter, the true positive dataset was 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, 45]. The optimal settings for each of these two tools found determined using this dataset were then used for subsequent analyses The tools we benchmarked on this dataset had highly variable performance in terms of their F1-score (0.36 – 0.99), precision (0.23 – 0.98), and recall (0.46 – 1.00) (Figure 1). DeepVirFinder and its predecessor VirFinder achieved the highest F1-scores of 0.99 and 0.98 respectively. These tools identified the majority of the true positive dataset as viral, and classified markedly less of the bacterial chromosome and plasmids fragments as viral, compared to other tools. PPR Meta, another machine learning based classifier, also performed well with an F1-score of 0.89. The homology-based tool ViralVerify similarly performed well (F1=0.81) with almost perfect recall (0.98) but a larger proportion of false positives resulted in a lower precision score (0.69). MetaPhinder, VIBRANT, VirSorter and VirSorter2 performed similarly (F1-scores of 0.60, 0.71, 0.63 and 0.68 respectively) with Seeker and MARVEL achieving relatively low scores due to their poor precision (0.39 and 0.36 respectively). Generally, pure machine learning based tools (DeepVirFinder, Seeker, PPR Meta, VirFinder) outperformed both mixed methods (MARVEL, VIBRANT, VirSorter2) and reference-based methods (MetaPhinder, ViralVerify, VirSorter) with average F1-scores of 0.81, 0.58, and 0.68 respectively, although these differences are not statistically significant due to the small sample sizes. Across our benchmark, every true positive phage contig was found by at least one tool, with 19.6% (1648/8411) found by all 10 tools (Additional File 2).

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.* [46], 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 3). 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. In general, the tools' F1-scores were considerably lower on this dataset than on the RefSeq artificial contigs, with F1-scores of machine learning-based tools dropping by 42% and reference-based tools by 33%, compared to the RefSeq benchmark (Figure 2). DeepVirFinder again outperformed all other tools despite a lower F1-score (0.61) and was closely followed by MetaPhinder which obtained a similar score to the previous dataset (0.56). However, DeepVirFinder achieved this score by having a lower recall (0.51) but the best precision (0.76), with MetaPhinder generating the opposite result, with the highest recall (0.94) and poorer precision (0.40). VirSorter, PPR Meta, and Seeker achieved comparable scores of 0.47, 0.47, and 0.46 respectively, with the latter being the only tool that performed better on this dataset than the first. VirSorter2 attained a lower F1-score (0.35) than its predecessor in this experiment, as a result of predicting 26.1% more true positive contigs but returning 166% more false positives, resulting in a low precision score (0.23). MARVEL, despite now having real bins as input, had a 39% lower F1-score than its RefSeq benchmark score, which is almost identical to the average decrease in F1-score across tools (40%). VirFinder, one of the tools that performed very well on the previous dataset, only identified nine out of 96 viral contigs across the three replicate samples, resulting in a very low recall

(0.09) and thus the lowest F1-score of all the tools (0.15). Similarly, ViralVerify performed poorly with a F1-score of 0.25, as a result of relatively low recall (0.42) and very low precision (0.17). Unlike in the previous analyses using the RefSeq benchmark dataset, machine learning based tools and reference-based tools performed almost identically with average F1-scores of 0.42 and 0.43 respectively, with mixed methods having a lower average F1-score of 0.30.

Out of the four phage species found in the assemblies, no tool was able to identify M13. PPR Meta, MetaPhinder, ViralVerify, and VirSorter2 were able to identify contigs belonging to the other three species – F0, ES18, and P22. VirSorter and VIBRANT were able to identify the three phage strains in two out of three samples and one out of three samples respectively, missing out contigs belonging to phage ES18. MARVEL predicted two phages across all samples, F0 and P22, with DeepVirFinder, VirFinder and Seeker only picking up the most abundant phage strain, F0. No correlation was found between F1-score and the number of phage strains detected (R_s = 0.146, p = 0.69) but a positive 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 = 0.726, p = 0.02).

We also recorded the running times of each tool on this dataset on a high-performance cluster (8 VCPU) (Figure 3). DeepVirFinder, MetaPhinder, PPR Meta, Seeker, and VirFinder were the fastest tools finishing each sample in under twenty minutes. Vibrant and ViralVerify performed their analyses in ~35 mins/sample and 1 hr/sample respectively. VirSorter required just under two hours to run each sample, with both MARVEL and VirSorter2 taking over four hours; MARVEL's runtime was over five hours/sample in total, if binning time is included.

Impact of tool prediction on diversity metric estimation

To test the impact of these tools on diversity estimates, four simulated mock community metaviromes containing an average of 719 viral genomes were retrieved from Roux *et al.* [47]. Reads were mapped to contigs (>1 kb) 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 a control. Read counts were then normalised by their length and sequencing depth, which Roux *et al.* [47] found to be reliable normalisation method. Diversity estimation metrics were then calculated using the normalised population counts. All tools returned less genomes per sample compared to the initial population, although there was significant variation between tools. DeepVirFinder, MetaPhinder, PPR Meta, and VirFinder retrieved the greatest percentage of genomes with 90.8%, 88.5%, 80.9%, and 88.1% respectively (Figure 4A). All other tools were able to retrieve more than 50% of the genomes with the exception of Seeker, which was only able to recover 28.7% of the population genomes. All Shannon's alpha diversities calculated from the count matrices of each tool were within 3% of the initial population with the exception of Seeker, whose *H* score was on average 27.2% lower (Figure 4B). Similarly, all but two tools identified populations with a Simpson alpha diversity index that was <1% different from the initial population, with MARVEL and Seeker's being 1.8% and 6.2% divergent, respectively (Figure 4C). MARVEL and ViralVerify, which only predicted ~50% of the total genomes in the initial population, were the only tools to estimate a comparatively higher alpha diversity than the initial population. For beta diversity, pairwise Bray-Curtis dissimilarities within a sample were small between all tools except for Seeker (Additional File 4), whose Analysis of Similarity (ANOSIM) showed significant dissimilarity when compared to other tools (r = 0.495, p = 0.0002 with Benjamini–Hochberg correction for multiple comparisons) (Figure 4D; Additional File 5).

Discussion

Bacteriophages are crucial members of microbial communities in nearly every ecosystem on Earth and are responsible for controlling host population size as well as having 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 certain 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 direct sequence similarity to 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 ≥ 1kb, without mis-assemblies), and determine approximate optimal thresholds. There was no significant difference in performance between machine learning and similarity-based classifiers, although the variance within these categories were high. DeepVirFinder and VirFinder, which were the only tools benchmarked that rely on k-mer frequencies, outperformed all other tools, exemplifying the power of this method. Machine learning tools performed better than reference and mixed-method tools, although the relatively low number of tools compared here may not mean that this is a generalisable observation. Prior to this study, we expected that mixed-method tools would gain an edge by providing the benefits of both machine learning and reference-based methods and minimising their weaknesses, but this does not seem to have been realised in the current generation of 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.

When tested on real metagenomic data, most tools performed significantly worse compared to the RefSeq dataset, although DeepVirFinder had the best performance, along with MetaPhinder. Generally, reference similarity tools had a lower drop in F1-score compared to the RefSeq dataset than deep learning tools, which is probably due to the presence of only four phage strains in each sample, all of which are widely available in the public databases used by these programs. This suggests that when studying metagenomic datasets where viral species are expected to be present at low frequencies, or where only a few phage strains of interest are being searched for, reference-based tools such as MetaPhinder or VirSorter are reasonable choices, especially when recall is more important than precision. When high precision is preferred, DeepVirFinder is the ideal choice, whilst also producing the best F1-score overall. 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 a few took multiple hours to complete their run. Generally, tools that only relied of machine learning prediction were considerably faster. It is important to note that VIBRANT, VirSorter, and VirSorter2 annotate the identified viral genomes and predict prophages and MARVEL's pipeline produces metagenomic bins at the expense of runtime, although these can be useful for some applications.

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 MARVEL and ViralVerify 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 programs, 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.

Although these benchmarks comprehensively compared the performance of state-of-the-art tools, there are a number of limitations with our study. First, whilst an effort was made to benchmark the machine learning based tools on data that it was not trained on, not all tools segregated their datasets by date or were not trained on NCBI RefSeq genome data (PPR Meta, VIBRANT, VirSorter2, ViralVerify), so there may be instances of overlap between a tool's training dataset and our RefSeq benchmark contig set. Second, whilst we use RefSeq genomes, and a mock metagenomic community to benchmark these tools, we did not address the tools' ability to identify viral sequences belonging to different phage families. Some tools such as MARVEL are specifically designed to detect certain families (those within the order *Caudovirales)*, and would therefore not perform as well on other phage families such as the *Microviridae* [11]. Third, we used the default database or the original trained model 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. Fifth, 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.* [48] where the authors combined multiple tools into a single workflow. 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 that we were unable to resolve (PhaMers, VirNet, VirMine).

Conclusion

Our comprehensive, comparative analysis of 10 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. Among the tested tools, DeepVirFinder performed best, with the highest F1-score in both the artifical RefSeq contig and mock uneven community datasets, whilst displaying similar diversity indices to the original population. All tools, except Seeker, were able to produce a diversity profile with similar indices to the original population, and are therefore suitable for phage ecology studies. DeepVirFinder was also one of the fastest tools in our study as were all other solely machine learning based tools such as PPR Meta, and Seeker. Generally, we suggest that of currently available tools DeepVirFinder should be considered the as an optimal solution in most cases, although this will depend on the type of sample that is analysed, whether precision or recall is more valued, and whether the additional functionality of other tools is required.

Materials and Methods

Benchmarking with RefSeq dataset

Complete bacterial and archaeal chromosomes and plasmids, and phage genomes deposited in RefSeq [49] since January 2018 (inclusive) were downloaded (2 July 2020). Chromosomes were then randomly down-sampled by a factor of 10, using reformat.sh from BBTools suite [50], to reduce downstream computational load. Phigaro [43] (default settings) and PhageBoost [44] (default settings), with genomes being split and run individually before being concatenated back together, were run in succession on the chromosomal sequences to remove prophage sequences. All sequences were then uniformly fragmented to between 1 kb and 15 kb, using a python script (available at [51]), 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 the virome mode, due to their improved performance in datasets consisting of mainly viral fragments [28, 45]. 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.

Benchmarking with mock community metagenomes

Three shotgun metagenomic sequencing replicates of an uneven mock community [46] was retrieved from the European Nucleotide Archive (BioProject PRJEB19901). These communities contain five phage strains: the DNA viruses ES18 (H1), F0, M13, and P22 (HT105), and the RNA virus F2. The quality of the data was checked using FASTQC (v11.8) [52] and overrepresented sequences were removed with Cutadapt (--max-n 0) (v2.10) [53]. Cleaned paired-end reads were then assembled with MetaSPAdes (with default settings) (v3.14.1) [54] and contigs <1 kb were removed. Each tool was then run on the three sets of contigs using optimal parameters as determined previously. MetaQUAST (v5.0.2) [55] was used to map contigs to reference phage genomes and calculate coverage. Run time of each tool 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; Memory: 64GB Multi-Bit ECC.

Benchmarking with simulated mock virome communities

Four mock communities (samples 2, 7, 9, and 13) containing between 500 and 1000 viral genomes created by Roux *et al.* were selected for analysis [47]. 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 (v.0.38) [56] and assembled with MetaSPAdes by Roux *et al.* [47]. The contigs were then downloaded for benchmarking. As before, contigs with length <1 kbp were removed, and then inputted into each viral identification program. Positive viral contig sets for each tool were then extracted and reads were mapped to these with BBMap [57] with ambiguous mapped reads assigned to contigs at random, as in Roux *et al*. [47]. Primary mapped reads with pairs mapping to the same contig (options -F 0x2 0x904) were then extracted with Samtools [58] and mapped to a pool of non-redundant population contigs. This pool was created by clustering all four samples with nucmer (v3.1) [59], 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 (counts per million) calculation commonly used in RNA-seq analyses. These abundance matrices were then used to calculate Shannon, Simpson, and Bray-Curtis dissimilarity indices using the vegan package [60]. Non-metric multidimensional scaling (NMDS) and analysis of similarity (ANOSIM) were also computed with vegan. ANOSIM *p*-values were corrected with the Benjamini–Hochberg method. Seed and permutations were set as 123 and 9999 respectively, where possible. All plots were generated with ggplot2 [61] and arranged with ggarrange from ggpubr [62].

List of Abbreviations

ANI	Average Nucleotide Identity
ANOSIM	Analysis of Similarity
GPU	Graphics Processing Unit
kb	Kilobases
MDA	Multiple Displacement Amplification
NCBI	National Center for Biotechnology Information
NMDS	Non-metric Multidimensional Scaling
RefSeq	NCBI Reference Sequence Database
ssDNA	Single-stranded DNA
VCPU	Virtual Central Processing Unit

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All RefSeq genomes used in this study can be downloaded from NCBI - https://www.ncbi.nlm.nih.gov/genome/. Relevant accession numbers for virus and host genomes can be found in Additional file 6.

Sequencing data from other studies first described in other man are available as described in the relevant articles. Scripts and commands created for this study are available at our GitHub repository – https://github.com/sxh1136/Phage_tools.

Competing interests

The authors declare that they have no competing interests

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).

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.

Acknowledgements

We wish to thank the authors of Kleiner *et al.* [46] and Roux *et al.* [47] for making their sequencing data publicly available, allowing their use in this study.

Tables

Software	Description	Reference
DeepVirFinder	Predicts viral sequences via a k-mer based deep learning method using convolutional neural networks (CNN). Based on VirFinder.	[39]
MARVEL	Machine learning tool for predicting phage sequences in metagenomic bins.	[40]
MetaPhinder	Integrates BLAST hits to multiple phage genomes in a database to identify phage sequences in assembled contigs.	[35]
metaviralSPAdes (ViralVerify)	Identifies viral sequences by leveraging metagenomic assembly graphs and analyzing the variations in depth of coverage between viral and bacterial genomes. Made of three modules, it also calculates the completeness of predicted viral sequences.	[42]
PhaMers	*Identifies phage sequences by a machine learning model based on k-mer frequencies.*	[63]
PPR-Meta	Deep learning CNN approach to identify both phages and plasmids	[64]
Seeker	Deep learning framework that uses Long Short-Term Memory models which does not depend on sequence motifs.	[65]
VIBRANT	Deep learning neural network based on protein signatures which also highlights auxiliary metabolic genes and pathways.	[45]
ViraMiner	*Extension of DeepVirFinder that is trained to identify any virus that may colonise human samples.*	[66]
VirFinder	K-mer based machine learning method for identification of viral contigs.	[37]
virMine	*Iterative pipeline that relies on the abundance of non-viral sequences in databases to strictly filter out unwanted contigs. Pipeline accepts both reads or assembled contigs.*	[67]
VirMiner	*Web-based pipeline that handles genome assembly,*	[68]

	functional annotation using a variety of databases and identification of phage contigs via a random forest algorithm	
VirNet	*Deep learning neural network using an attentional neural model trained on nucleotide viral fragments.*	[69]
VIROME	*Web-based pipeline that classifies viral sequences based on homology to databases and functional 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.	[70]
VirusSeeker	*Consists 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.*	[36]

Table 1. Overview of tools to identify and predict phage sequences in microbial ecosystems. Tools in italics were not included as they were either irrelevant to this study or insurmountable technical difficulties were encountered during their use.

Sequence group	Number of sequences	Number of artificial contigs
Bacterial and archaeal chromosomes	719 (6,963)	326,595
Bacterial and archaeal plasmids	5,664	100,296
Bacteriophage and archaeal virus genomes	1,039	8,411

Table 2. Sequences included in the RefSeq-based dataset. Numbers in parenthesis indicate the number of sequences before random down sampling. All sequences were randomly fragmented into artificial contigs of lengths between 1 kb and 15 kb. Identities of the included sequences are provided in Additional file 6.

Figure legends

Figure 1: 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. As MARVEL accepts bins as input, each contig was treated as a separate bin.

Figure 2: Comparison of viral identification tools on uneven mock community samples

Mock community reads were retrieved from a previous study [46]. and assembled with metaSPAdes before running each identification tool using optimal thresholds based on previous benchmarks. 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. Where no contigs were identified as viral by the tool, precision was set as zero.

Figure 3: Comparison of tool runtimes on uneven mock community samples

Wall runtime of each tool on mock community samples was recorded on an 8 VCPU, 64GB RAM, Linux high performance cluster without GPU acceleration. Each assembly contains ~50 million bp. For MARVEL, the lower bar indicates wall time of the prediction tool itself, and the top bar indicates the wall time for binning each sample.

Figure 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.*[47] were downloaded. Each program 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” [60]. ‘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. Non-metric 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.

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. Cobián Güemes 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. doi:10.1371/journal.ppat.1007878.

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. doi:10.3389/fendo.2019.00784.

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. doi:10.1093/femsle/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. doi:10.4161/bact.25029.

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 M-S, Bae J-W. Lysogeny is prevalent and widely distributed in the murine gut microbiota. ISME J. 2018;12:1127–41.

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 PHASTEST: 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. Bioinforma Oxf Engl. 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. Jurtz VI, Villarroel J, Lund O, Voldby Larsen M, Nielsen M. MetaPhinder—Identifying Bacteriophage Sequences in Metagenomic Data Sets. PLoS ONE. 2016;11. doi:10.1371/journal.pone.0163111.

36. 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.

37. 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.

38. Ponsero AJ, Hurwitz BL. The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes. Front Microbiol. 2019;10. doi:10.3389/fmicb.2019.00806.

39. 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.

40. Amgarten D, Braga LPP, da Silva AM, Setubal JC. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. Front Genet. 2018;9. doi:10.3389/fgene.2018.00304.

41. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45 Database issue:D491–8.

42. Antipov D, Raiko M, Lapidus A, Pevzner PA. MetaviralSPAdes: assembly of viruses from metagenomic data. Bioinformatics. 2020;36:4126–9.

43. Starikova EV, Tikhonova PO, Prianichnikov NA, Rands CM, Zdobnov EM, Ilina EN, et al. Phigaro: high throughput prophage sequence annotation. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa250.

44. 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. doi:10.1093/nargab/lqaa109.

45. 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.

46. 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. doi:10.1038/s41467-017-01544-x.

47. 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.

48. 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. bioRxiv. 2020;:2020.07.24.219899.

49. Na O, Mw W, Jr B, S C, D H, R M, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2016;44. doi:10.1093/nar/gkv1189.

50. Bushnell B. BBMap. 2020. https://sourceforge.net/projects/bbmap/.

51. Ho SFS. Benchmarking phage predection tool GitHub repository. 2021. https://github.com/sxh1136/Phage_tools.

52. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC. 2019.

53. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2.

54. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 2017;27:824–34.

55. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinforma Oxf Engl. 2016;32:1088–90.

56. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.

57. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. 2014. https://www.osti.gov/biblio/1241166.

58. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

59. 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.

60. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: Community Ecology Package. 2020. https://CRAN.R-project.org/package=vegan. Accessed 1 Feb 2021.

61. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. https://ggplot2.tidyverse.org.

62. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots. 2020. https://CRAN.R-project.org/package=ggpubr.

63. Deaton J, Yu FB, Quake SR. PhaMers identifies novel bacteriophage sequences from thermophilic hot springs. bioRxiv. 2017;:169672.

64. 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;8. doi:10.1093/gigascience/giz066.

65. 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–e121.

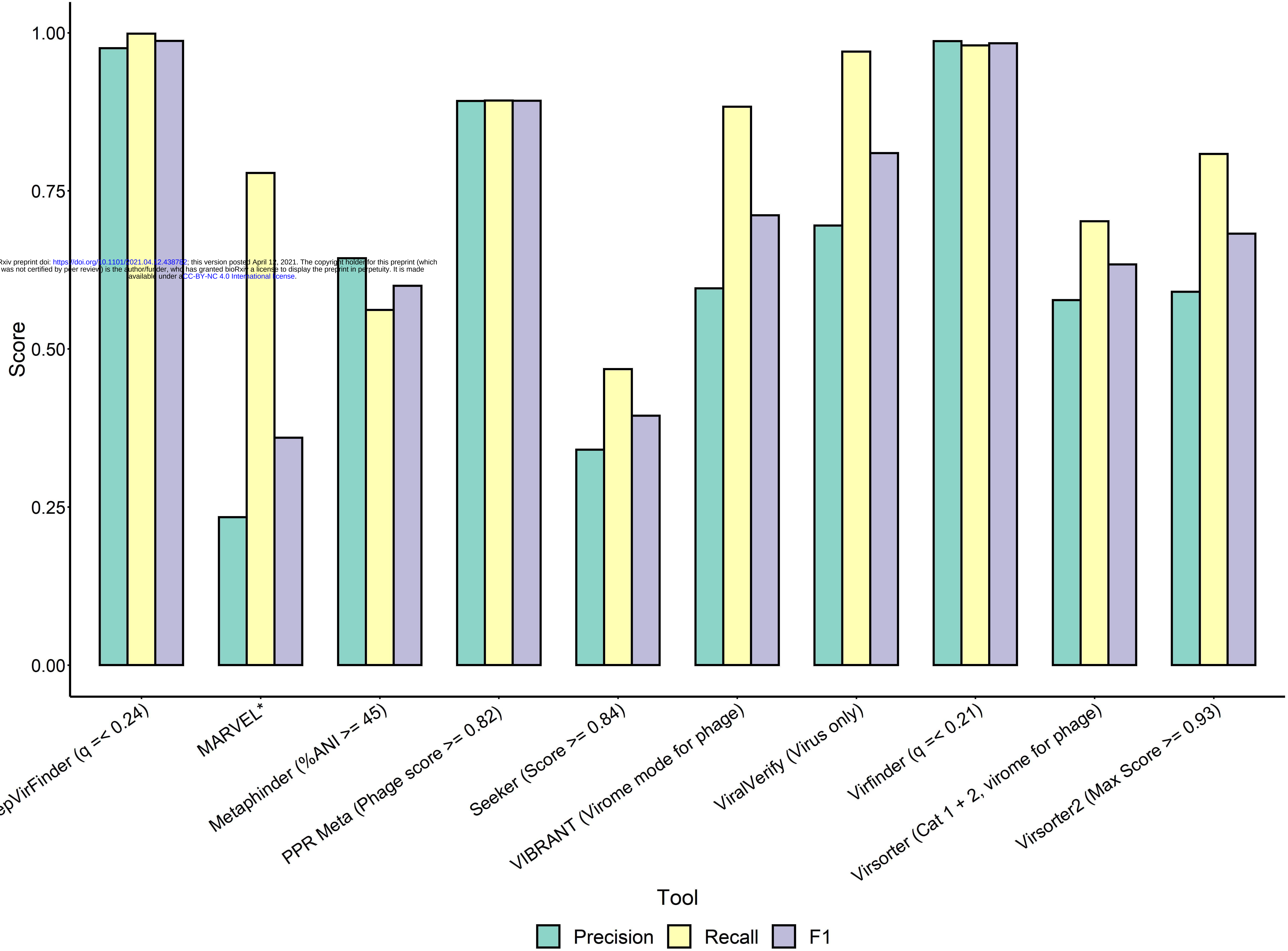
66. Tampuu A, Bzhalava Z, Dillner J, Vicente R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. PLoS ONE. 2019;14. doi:10.1371/journal.pone.0222271.

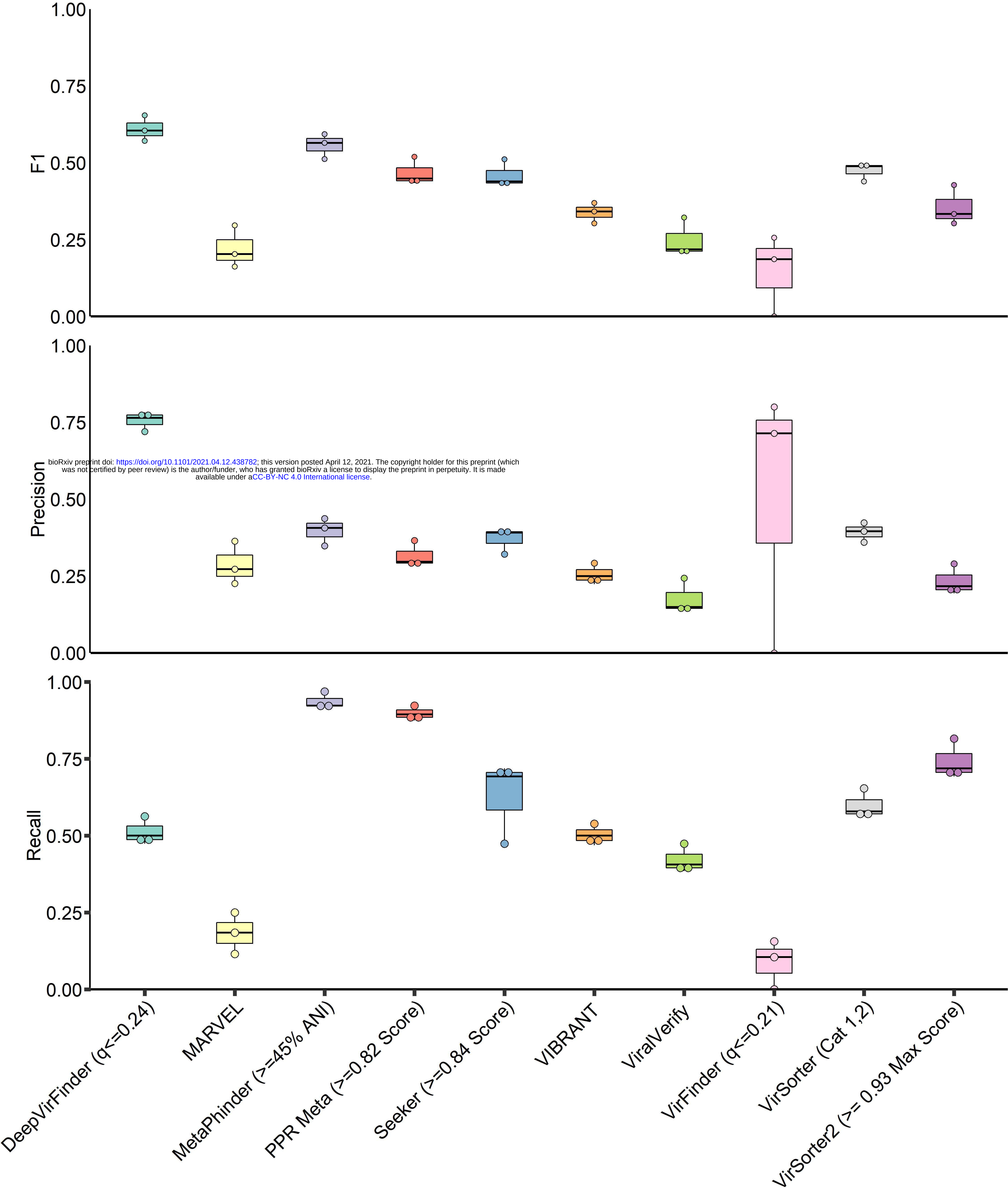
67. Garretto A, Hatzopoulos T, Putonti C. virMine: automated detection of viral sequences from complex metagenomic samples. PeerJ. 2019;7:e6695.

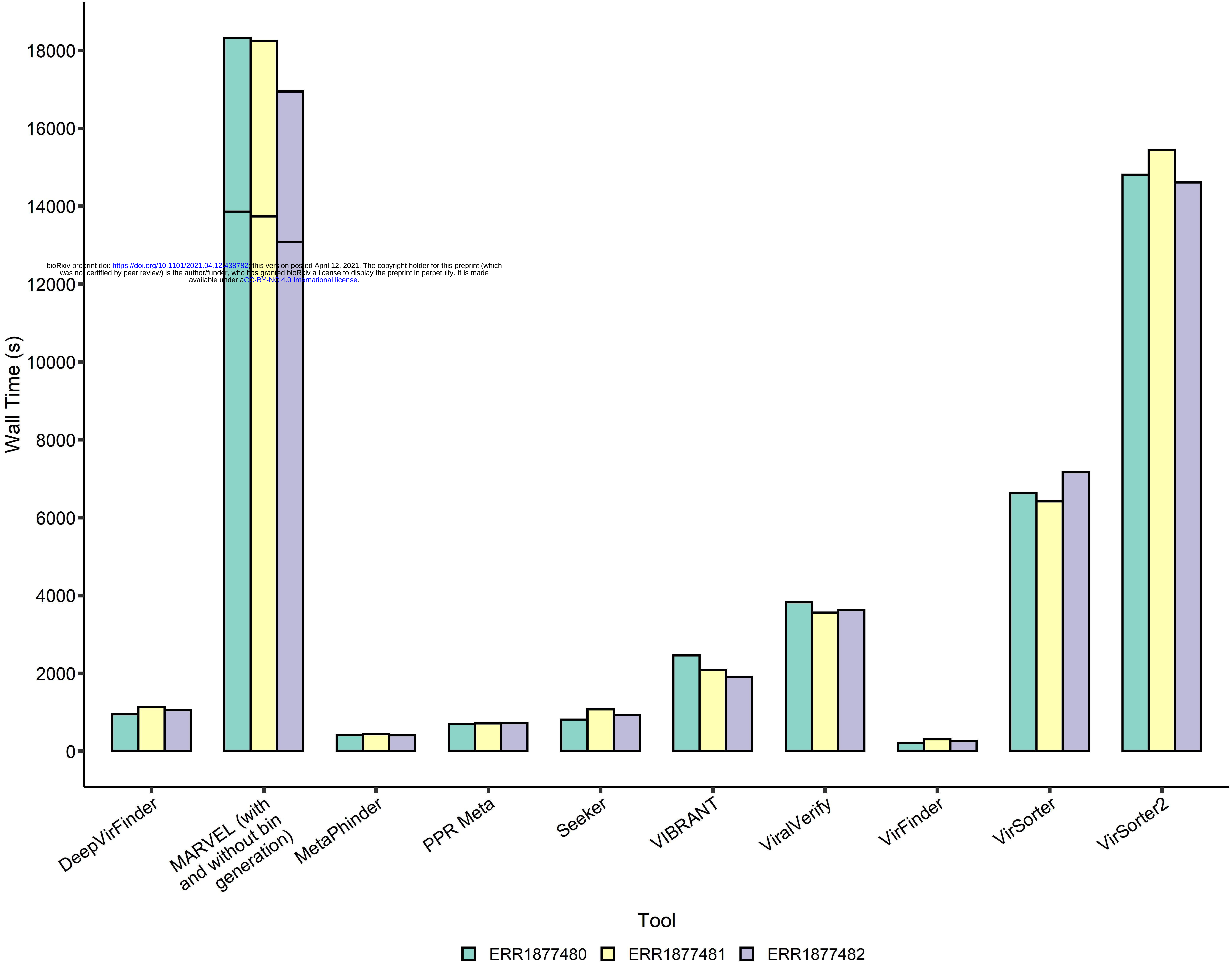
68. Zheng T, Li J, Ni Y, Kang K, Misiakou M-A, Imamovic L, et al. Mining, analyzing, and integrating viral signals from metagenomic data. Microbiome. 2019;7. doi:10.1186/s40168-019-0657-y.

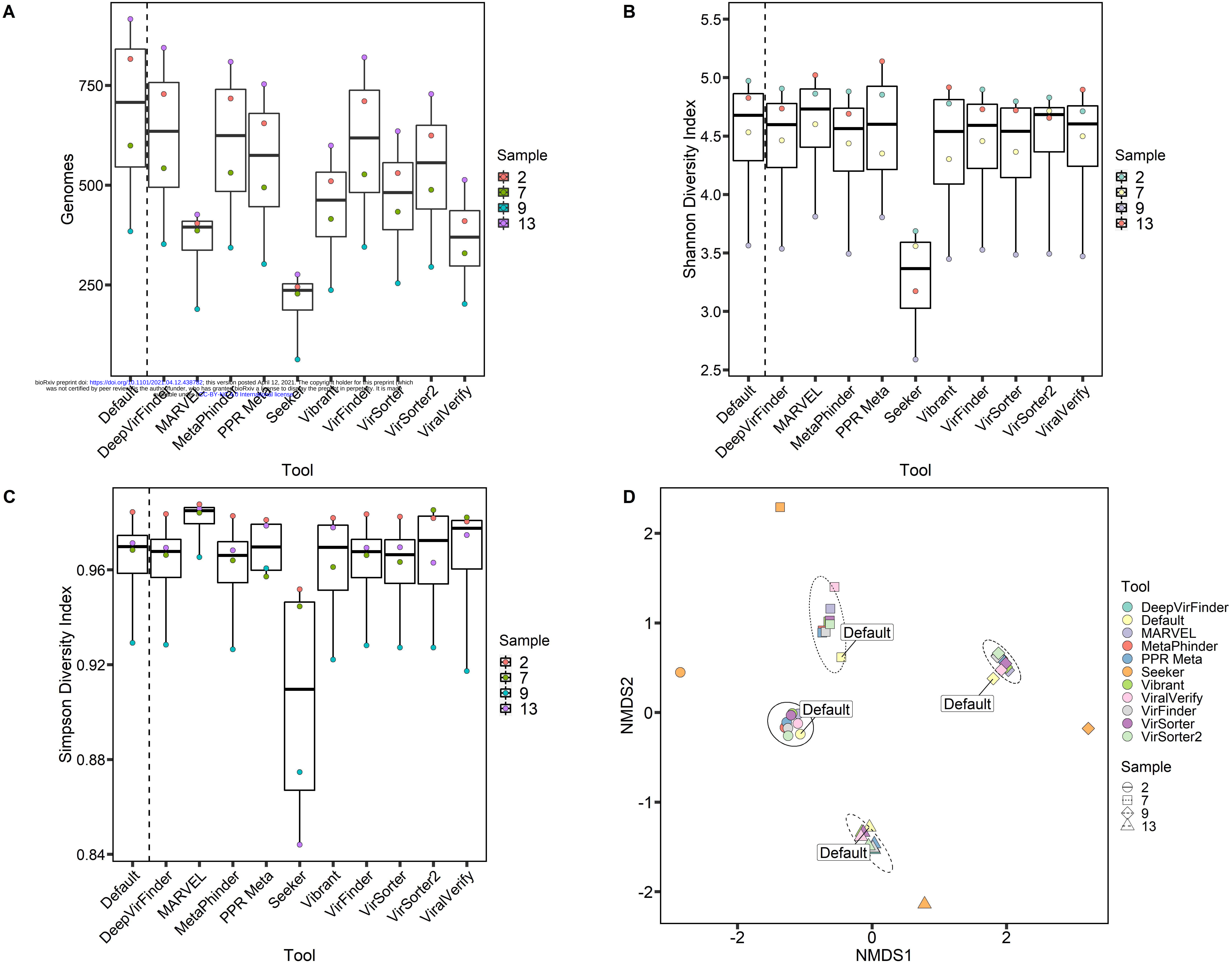
69. Abdelkareem AO, Khalil MI, Elaraby M, Abbas H, Elbehery AHA. VirNet: Deep attention model for viral reads identification. 2018 13th Int Conf Comput Eng Syst ICCES. 2018;:623–6.

70. 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.


1.00
0.75
0.50
0.25
0.00
Score
DeepVirFinder (q =< 0.24)
MARVEL*
Metaphinder (%ANI >= 45)
PPR Meta (Phage score >= 0.82)
Seeker (Score >= 0.84)
VIBRANT (Virome mode for phage)
ViralVerify (Virus only)
Virfinder (q =< 0.21)
Virsorter (Cat 1 + 2, virome for phage)
Virsorter2 (Max Score >= 0.93)
Tool
Precision
Recall
F1


1.00
0.75
0.50
0.25
0.00
F1
1.00
0.75
0.50
0.25
0.00
Precision
1.00
0.75
0.50
0.25
0.00
Recall
DeepVirFinder (q<=0.24)
MARVEL
MetaPhinder (>=45% ANI)
PPR Meta (>=0.82 Score)
Seeker (>=0.84 Score)
VIBRANT
ViralVerify
VirFinder (q<=0.21)
VirSorter (Cat 1,2)
VirSorter2 (>= 0.93 Max Score)


Wall Time (s)
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
DeepVirFinder
MARVEL (with and without bin generation)
MetaPhinder
PPR Meta
Seeker
VIBRANT
ViralVerify
VirFinder
VirSorter
VirSorter2
Tool
ERR1877480
ERR1877481
ERR1877482


A
Genomes
750
500
250
Tool
Default
DeepVirFinder
MARVEL
MetaPhinder
PPR Meta
Seeker
Vibrant
VirFinder
VirSorter
VirSorter2
ViralVerify
Sample
2
7
9
13
B
Shannon Diversity Index
5.5
5.0
4.5
4.0
3.5
3.0
2.5
Tool
Default
DeepVirFinder
MARVEL
MetaPhinder
PPR Meta
Seeker
Vibrant
VirFinder
VirSorter
VirSorter2
ViralVerify
Sample
2
7
9
13
C
Simpson Diversity Index
0.96
0.92
0.88
0.84
Tool
Default
DeepVirFinder
MARVEL
MetaPhinder
PPR Meta
Seeker
Vibrant
VirFinder
VirSorter
VirSorter2
ViralVerify
Sample
2
7
9
13
D
NMDS2
2
1
0
-1
-2
NMDS1
-2
0
2
Default
Default
Default
Default
Tool
DeepVirFinder
Default
MARVEL
MetaPhinder
PPR Meta
Seeker
Vibrant
ViralVerify
VirFinder
VirSorter
VirSorter2
Sample
2
7
9
13