Benchmarking combined informatics approaches for virus discovery: Caution is needed when combining viral identification methods

Bridget Hegarty[1*+], James Riddell[2*], Eric Bastien[3], Kathryn Langenfeld[4], Morgan Lindback[3], Jaspreet S. Saini[5], Anthony Wing[3], Jessica Zhang,[6] Melissa Duhaimea[3+]

[1] Department of Civil and Environmental Engineering, Case Western Reserve University, Cleveland, OH
[2] Department of Microbiology, The Ohio State University, Columbus, OH
[3] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI
[4] Department of Civil and Environmental Engineering, Stanford University, Palo Alto, CA
[5] Laboratory for Environmental Biotechnology, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[6] Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI
[*] BH and JR contributed equally to this manuscript
[+] corresponding authors

SI - VSTE - MS
VSTE: Tables for Manuscript and SI

Abstract

Background

Identifying viruses from mixed environmental metagenomic samples has become an essential component of microbiome studies. Many informatics tools have been developed to recover viral sequences from mixed metagenomic datasets. As each possesses different biases and behaviors, it is difficult to know which tool(s) are best suited for a particular study, and can lead to identifying false positives and/or capturing only a small fraction of total viruses in the sample. Studies have combined multiple tool outputs attempting to recover more viruses, but no combined approach has been benchmarked for accuracy.

Results

We benchmarked viral identification tool combinations using mock environmental metagenomic datasets composed of publically available viral, bacterial, archaeal, fungal, and protist sequences. We then applied these tool combinations to different aquatic metagenomes (fresh and saltwater, drinking water, wastewater) and filtering strategies (virus-enriched and bulk metagenome) to evaluate the impact of habitat and viral enrichment on tool and ruleset performance. Two-tool combinations tended to increase the number of viruses recovered without adding significant contamination, but three or more tool combinations significantly increased contamination decreasing accuracy. The highest accuracy achieved on mock bulk metagenomic datasets was 0.75. However, closer inspection revealed a strong viral gene signature in the false positives. Many tool combinations recovered similar sets of viruses.

Conclusion

Combining viral identification tools increases viral recall, but at the expense of increased false positives. Using CheckV and Kaiju as quality filtering steps can help decrease non-viral contamination, but even the most optimal tool combinations do not exceed an MCC of 0.75. This accuracy plateau may be because viral identification tools rely on accurately labeled data for training and validation, but many sequences in NCBI are incorrectly labeled as non-viral entities when they are actually viral, or genes are annotated as viral when they may also be naturally present in non-viral entities. While improved algorithms may lead to more accurate viral identification tools, this should be done in tandem with curating accurately labeled viral gene and sequence databases. For most applications, we recommend the use of our ruleset based on the VirSorter2 and tuning removal rules.

Background

Viruses are an essential component of microbial ecosystems: they influence nutrient cycling and microbial community dynamics[1], account for 20-40% of microbial mortality per day[2], reprogram their hosts' metabolism,[3,4] and horizontally transfer genes between host populations.[5,6] The primary approach used to discover and describe viral diversity is culture-independent metagenomic sequencing. However, viral sequences remain challenging to differentiate from non-viral ones because viruses have no universal marker gene,[7] high mutation rates,[8,9] and relatively small reference databases relative to the magnitude of their diversity.[10] Additionally, current environmental sample collection and sequencing methods recover predominantly short contigs. Short contigs are challenging to classify correctly because they often do not contain enough information to leverage our knowledge of what makes viral sequences distinct.[11,12]

The challenge of identifying viral sequences in metagenomic datasets has driven the development of many viral identification tools over the past decade that aim to differentiate viral sequences from non-viral sequences.[13] Tools differ in the types of viruses they identify, what sequence lengths they are optimized for, and the training data and algorithms underlying them. To be confidently applied to environmental data, viral identification tools must be trained on sequences representative of the microbiota being studied to ensure the tool has seen enough of the sequence space to correctly classify viral sequences. Sequence types commonly found in environmental metagenomes include bacteria, viruses, plasmids, archaea, protists, and fungi. Some tools, such as VirSorter2[12] and VIBRANT,[14] include these diverse sequence types, as well as diversity within each sequence type, expanding the classification accuracy of each tool. Other tools like DeepVirFinder and VirSorter do not include as diverse sequences: DeepVirFinder does not include non-prokaryotic references and VirSorter is only built on bacterial and archaeal virus references. Further, the performance of viral identification tools depends on the interaction between the tool algorithm and the sample type. In a comparison of the viruses recovered by different viral identification tools across 13 environmental samples, differences were found in the number of sequences called viral between environments.[14]

While many viral identifications tools have comparable accuracy, their underlying algorithms are quite different and tend to capture different sets of viruses from the same sample. With so many tools available, it can be difficult to choose the most appropriate one for a given study. Rather than choose one tool, some studies combined the outputs of multiple tools to classify viral sequences to capture a greater portion of the viral signal .[15–18] This approach assumes that combining multiple tools will improve overall accuracy by discovering more viruses without greatly increasing nonviral contamination, but this assumption has not been validated in the literature. In particular, it remains unknown whether or not these multi-tool strategies significantly increase contamination (e.g., by each tool returning non-overlapping false positive viral sequences).

Here, we benchmarked whether multi-tool approaches are capable of distilling a more complete and accurate set of viral sequences. From our analysis, we suggest pipelines applicable for short and long-read sequences, as well as metagenomes and viral-concentrated samples. These pipelines, by returning more viral sequences with less non-viral contamination, will enable new and more accurate insights in microbial ecology.

Methods

Figure 1 provides a brief overview of our methodology. We downloaded viral, bacterial, fungal, plasmid, protist, and archaeal genome sequences from the NCBI reference sequence database, RefSeq, as well as non-RefSeq virus genomes, (https://zenodo.org/record/4297575). These non-RefSeq virus genomes represent a comprehensive validated set of viruses; and were used to train and test VirSorter2. Accession numbers for sequences used in the testing sets are available at https://github.com/DuhaimeLab/VSTE/blob/main/TestingPipeline/IntermediaryFiles/viral_tools_combined.tsv. From these sequences we constructed five mock environmental metagenomes representative of real-world environmental metagenomes to act as our testing datasets. We then assessed 29 available viral identification tools and selected a subset (six tools) for comparative analyses (Figure 1). Using these tools, we defined score cutoffs for viral sequences and compared the accuracy, precision, and recall of each combination.

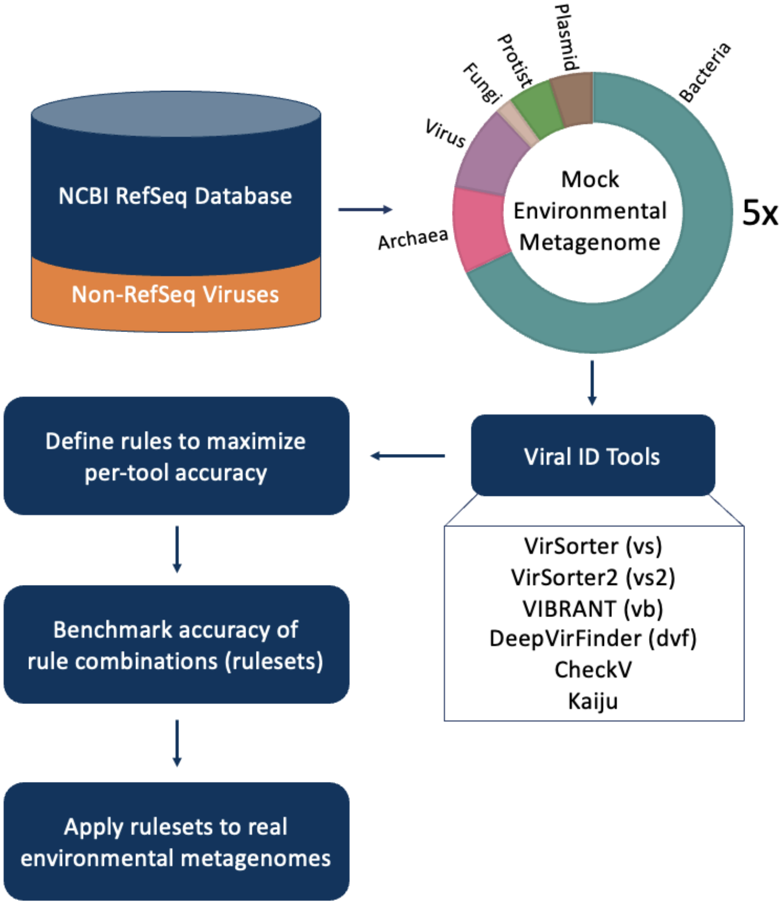

NCBI RefSeq Database
Non-RefSeq Viruses
Plasmid
Protist
Fungi
Virus
Archaea
Bacteria
Mock Environmental Metagenome
5x
Viral ID Tools
VirSorter (vs)
VirSorter2 (vs2)
VIBRANT (vb)
DeepVirFinder (dvf)
CheckV
Kaiju
Define rules to maximize per-tool accuracy
Benchmark accuracy of rule combinations (rulesets)
Apply rulesets to real environmental metagenomes

Figure 1. Workflow Overview. Sequences > 3 kb were randomly downloaded from NCBI and a curated Non-RefSeq viral genomes database to generate five mock environmental metagenomes. They were run through six viral identification tools, where score cutoffs were defined based on each tool's outputs to maximize their accuracy. Accuracy was then assessed for each tool combination. Tool combinations were then used to classify sequences from six real-world aquatic metagenomes: three bulk metagenomes and three virus-enriched metagenomes.

Selection of viral identification tools

29 viral identification tools[12,14,19,19–44] were found through literature search and assessed to determine their suitability for inclusion in this study (Table S1). As new viral identification tools are actively being developed, only tools published before June 2022 were included in this study. Tools were included if they met the following criteria: the tool 1) identifies viruses that infect prokaryotes (i.e., bacteria and archaea), 2) is suitable for application to multiple environments, 3) is designed to target viral contigs of lengths greater than 3 kilobases, 4) can classify millions of contigs within a few days on high performance computing clusters (i.e., not only available on a web server), 5) developers actively respond to user issues, 7) performs well in previous comparative studies of viral identification tools,[12,14,45] and 8) is not specific to prophages.

Four of the viral identification tools met the above criteria: DeepVirFinder,[21] VIBRANT,[14] VirSorter,[41] and VirSorter2.[12] While not strictly viral identification tools, two other tools were used to tune and improve the quality of the viral predictions in our test sets: Kaiju[46] and CheckV.[20] All six will be referred to as "viral identification tools" or simply "tools" in this manuscript. More information about the chosen tools can be found in Table 1.

Table 1 - Overview of viral identification tools selected for inclusion in this study.

Tool name (version)	Tool Description	Algorithmic Approach	Why we included the tool
CheckV[20]	CheckV is an automated pipeline that identifies closed viral genomes, estimates the completeness of genome fragments, and removes host regions from proviruses.	HMM virus and host marker genes, virus-host boundary prediction, AAI based estimation of genome completeness	Not a viral identification tool, but provides useful benchmarking information for refining predictions from other tools.
DeepVirFinder[21]	DeepVirFinder uses a multi-layered deep learning algorithm trained on a positive set of viral sequences from viral RefSeq data and a negative set of prokaryotic ones.	K-mer based deep learning convolutional neural network	Recent and increasingly commonly used tool based on a neural net.
VIBRANT[14]	VIBRANT is a hybrid tool that uses both machine learning and protein similarity to classify viruses as either high, medium, low quality, or non-viral.	Neural network of protein annotations of HMMs	Recent and commonly used tool, provides useful gene annotation information
VirSorter[41]	VirSorter uses probabilistic models with reference and non-reference dependencies, as well as detecting hallmark viral genes.	Probabilistic modeling using HMMs	Commonly used, high quality predictions
VirSorter2[12]	VirSorter2 uses a neural network classifier built on top of the existing VirSorter infrastructure of reference based viral identification.	A multi-classifier combining a Random Forest model and expert knowledge of viral features	Recent and increasingly commonly used tool that better covers viral diversity than most other tools.
Kaiju[46]	Kaiju is a taxonomic classifier that compares metagenomic sequences to NCBI reference databases at the protein level and assigns a near or exact taxon match if one is found.	A taxonomic classifier that uses protein-level classifications to assign reads to taxon from NCBI databases.	Not a viral identification tool, but extremely fast method of taxonomic identification based on NCBI releases

Testing sets were run through each of the viral identification tools (CheckV (v0.9.0), DeepVirFinder (v1.0), Kaiju (v1.9.0), VIBRANT (v1.2.1), VirSorter (v1.0.6), and VirSorter2 (v2.2.3)) using the University of Michigan Great Lakes Supercomputing Cluster. The Kaiju nr_euk database (updated from NCBI 05-23-2022) was used for Kaiju taxonomic classification. All tools were run with default parameters except for specifying a 3000-base contig length cutoff.

Creation of sequence test set

To capture the variability due to test data selection, nine testing sets were created by randomly sampling genomes with replacement from NCBI and VirSorter2 curated databases to create datasets that mimicked metagenomic environmental data. As the variability between testing sets was captured with five (Figure S1), that many were used for subsequent analyses. The testing sets were composed of approximately 68% bacteria, 10% archaea, 10% virus (not proviruses), 5% plasmid, 5% protist, and 2% fungi sequences, totalling ~8k sequences. The proportion of sequences was chosen to be representative of metagenomic data, which are dominated by bacterial sequences.[11,47] The non-viral portion was randomly sampled from 5.3M bacteria, 55.1k archaea, 6.6k plasmid, 69.6k fungi, and 216.5k protist sequences from the NCBI database (accessed Nov 2019 for bacteria and archaea, April 2022 for others). The viral portion of the testing set was generated by random sampling from 13.8k viral sequences from the NCBI database and 370.153k viral sequences from the Non-RefSeq viral genomes database. As DeepVirFinder requires sequences to be less than 2.1 Mb, a custom python script was written to trim the testing set sequences to meet this length cutoff. Further, only sequences longer than 3000 bp were used.

Design of viral identification rulesets

Viral identification tools generate scores that indicate how confident they are that a given sequence is of viral origin, but users are often faced with the dilemma of setting their own score cutoff to decide what to call "viral." To aid in the process of choosing rules and cutoffs for predicting viral contigs, we designed six rules (Figure 2). Each rule includes at least two subrules that use outputs from the six selected tools. These rules were designed through three processes:

1) *Evaluating existing recommendations for tool cutoffs and application*: the recommended cutoffs for distinguishing viral and non-viral sequences in each tool's protocol were used as an initial set of rules.[12,15,16,48]

2) *Curation and evaluation of biological features:* some viral identification tools output biological features for each contig, e.g., VirSorter2 reports the number of viral hallmark genes identified, CheckV reports the completeness of a sequence and relative percentage of viral versus non-viral genes. These biological features were used to create classification criteria (described below in Figure 2) to distinguish viral and non-viral sequences.

The developers' recommendations for calling a sequence viral were used as a starting point; these cutoffs were then adjusted and expanded to maximize the number of true viral contigs being classified as viral and minimize the number of non-viral contigs being classified as viral in the mock environmental microbial communities. An important part of this process was defining two sets of "tuning rules: (1) a "tuning removal" rule that decreases the viral score based on distinctly non-viral sequence features and (2) a "tuning addition" rule that increases the viral score based on distinctly viral sequence features.

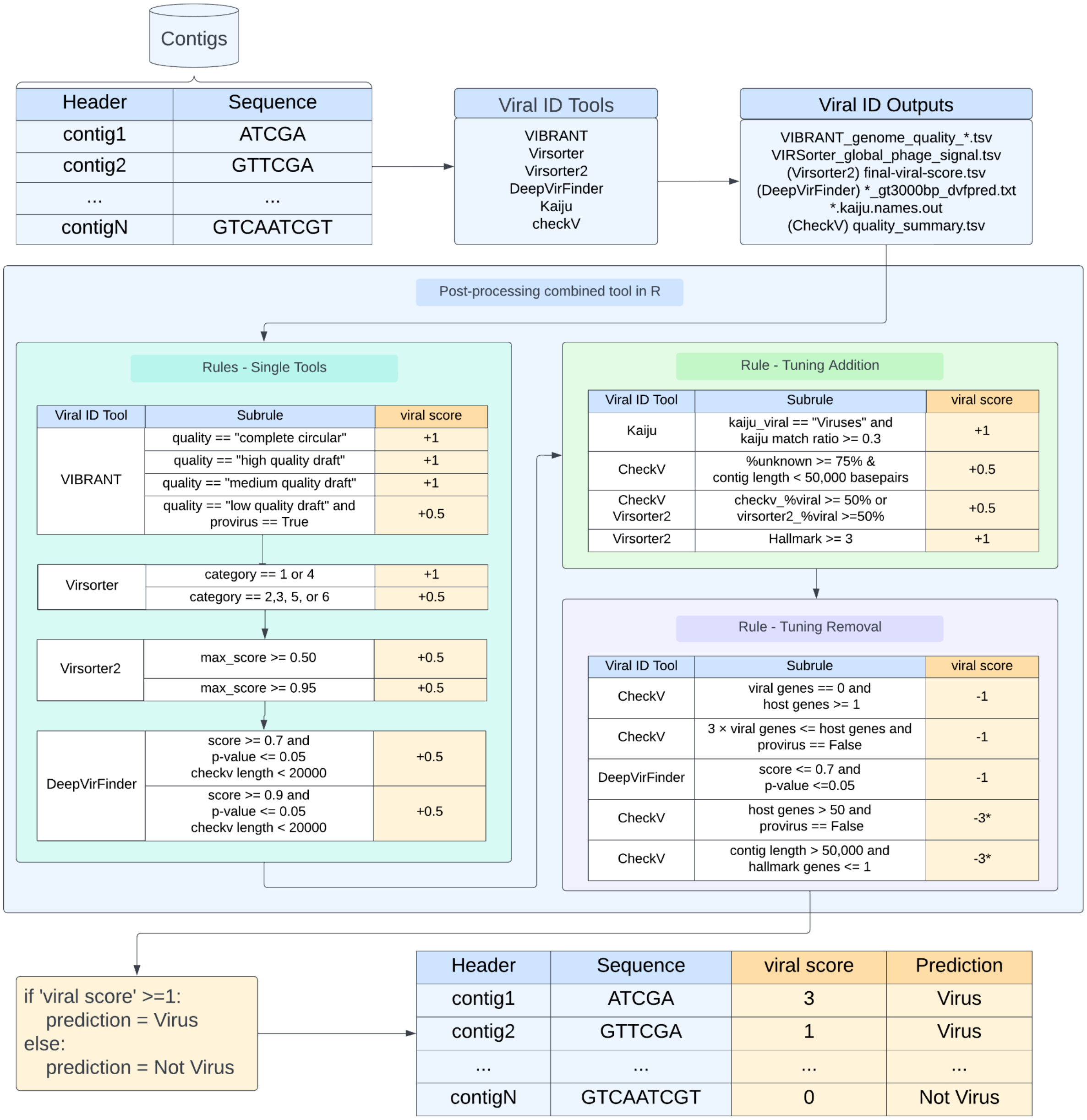
Ultimately, 63 combinations of these six rules (rulesets) were evaluated by comparing the viral score of each sequence to the classification assigned by the database. From these values, precision (the number of true viruses in our test set called viral divided by all contigs called viral), recall (the number of true viruses in our test set called viral divided by all true viruses), and Matthews Correlation Coefficient

(MCC, considers relative proportion of false positives, false negatives, true positives, and true negatives)[49] were calculated (supplemental equations).

Viral Identification Rulesets

Figure 2 provides the details of our viral identification process, which we provide an overview of here. For each putative viral contig or genome, the outputs of the viral identification tools (i.e., VirSorter, VirSorter2, VIBRANT, DeepVirFinder) are given a weighted value that is added to an overall *viral score*. After these single-tool viral identification rules are applied, the "tuning addition" and "tuning removal" rules, which contain subrules derived from different tools, further screen the contigs by looking at viral genes, host genes, and taxonomy.Contigs with a final *viral score* greater than or equal to 1 were considered viral, and scores less than 1 non-viral.

The script to identify viral contigs from these tool outputs is freely available at https://github.com/DuhaimeLab/VSTE, along with scripts to run the full testing pipeline, example runs, and classifier outputs on the environmental samples presented in this study.



Header	Sequence
contig1	ATCGA
contig2	GTTCGA
...	...
contigN	GTCAATCGT

Viral ID Tools

VIBRANT
Virsorter
Virsorter2
DeepVirFinder
Kaiju
checkV

Viral ID Outputs

VIBRANT_genome_quality_*.tsv
VIRSorter_global_phage_signal.tsv
(Virsorter2) final-viral-score.tsv
(DeepVirFinder) *_gt3000bp_dvfpred.txt
*.kaiju.names.out
(CheckV) quality_summary.tsv

Rules - Single Tools

Viral ID Tool	Subrule	viral score
VIBRANT	quality == "complete circular"	+1
	quality == "high quality draft"	+1
	quality == "medium quality draft"	+1
	quality == "low quality draft" and provirus == True	+0.5

Viral ID Tool	Subrule	viral score
Virsorter	category == 1 or 4	+1
	category == 2,3, 5, or 6	+0.5

Viral ID Tool	Subrule	viral score
Virsorter2	max_score >= 0.50	+0.5
	max_score >= 0.95	+0.5

Viral ID Tool	Subrule	viral score
DeepVirFinder	score >= 0.7 and p-value <= 0.05 checkv length < 20000	+0.5
	score >= 0.9 and p-value <= 0.05 checkv length < 20000	+0.5

Rule - Tuning Addition

Viral ID Tool	Subrule	viral score
Kaiju	kaiju_viral == "Viruses" and kaiju match ratio >= 0.3	+1
CheckV	%unknown >= 75% & contig length < 50,000 basepairs	+0.5
CheckV Virsorter2	checkv_%viral >= 50% or virsorter2_%viral >=50%	+0.5
Virsorter2	Hallmark >= 3	+1

Rule - Tuning Removal

Viral ID Tool	Subrule	viral score
CheckV	viral genes == 0 and host genes >= 1	-1
CheckV	3 × viral genes <= host genes and provirus == False	-1
DeepVirFinder	score <= 0.7 and p-value <=0.05	-1
CheckV	host genes > 50 and provirus == False	-3*
CheckV	contig length > 50,000 and hallmark genes <= 1	-3*

```
if 'viral score' >=1:
    prediction = Virus
else:
    prediction = Not Virus
```

Header	Sequence	viral score	Prediction
contig1	ATCGA	3	Virus
contig2	GTTCGA	1	Virus
...	...	...	...
contigN	GTCAATCGT	0	Not Virus

Figure 2. Diagram of approach details. Contigs are first processed by each viral identification tool. Next, the tool outputs are programmatically evaluated to generate a *viral score* according to the diagram flow. The last tool in the pipeline, CheckV, serves as a quality control step; it is used to remove contigs that present significantly more host-like than virus-like biometrics.

Application of new pipeline to environmental metagenomes

All tool combinations were used to identify viruses from five previously published environmental datasets representing different aquatic environments and size fractions (Table 2). Three environments (drinking water, global ocean water, and eutrophic lake water) contained metagenomic assemblies (>2 µm), and three environments (wastewater, eutrophic lake water, oligotrophic lake water) contained virome assemblies (<0.2 and <0.45 µm), meaning samples were enriched for viruses by filtering through a small pore to remove most cellular organisms before DNA extraction.

The default tool settings were used except for the oligotrophic lake and wastewater virus-enriched samples (Table 2), while the `-virome flag was used for VIBRANT and VirSorter to reduce sensitivity. All tools were run with a 3 kb cutoff to remove small contigs.

Results + Discussion

Viral identification tools use algorithms based on knowledge of viral sequences and machine learning to separate viral and non-viral sequences. In this study, we test the hypothesis that combining viral identification tools with different underlying algorithms will improve accuracy. Performance of 63 combinations of six rules (rulesets) were evaluated using five mock metagenomes of known composition (Figure 3, S1). The six rules are as follows: four single-tool rules derived from four viral identification tools (i.e., VIBRANT, VirSorter, VirSorter2, DeepVirFinder) and two additional tuning rules: tuning removal (Kaiju, CheckV, VirSorter2, VirSorter, and VIBRANT) and tuning addition (Kaiju, CheckV, and VirSorter2) (Figure 2). Our tuning rules build on those implemented elsewhere and are critical for reducing false positives in our predictions.[48] In addition to testing on mock environmental metagenomes, we tested the 63 viral identification rulesets on environmental datasets and compared the proportion of viral sequences predicted for each environment. Based on our benchmarking we recommend using VirSorter2 with our tuning removal rule to identify viruses from most metagenomic datasets. In the rest of this section, we will compare the performance of our rulesets and elaborate on their strengths and weaknesses.

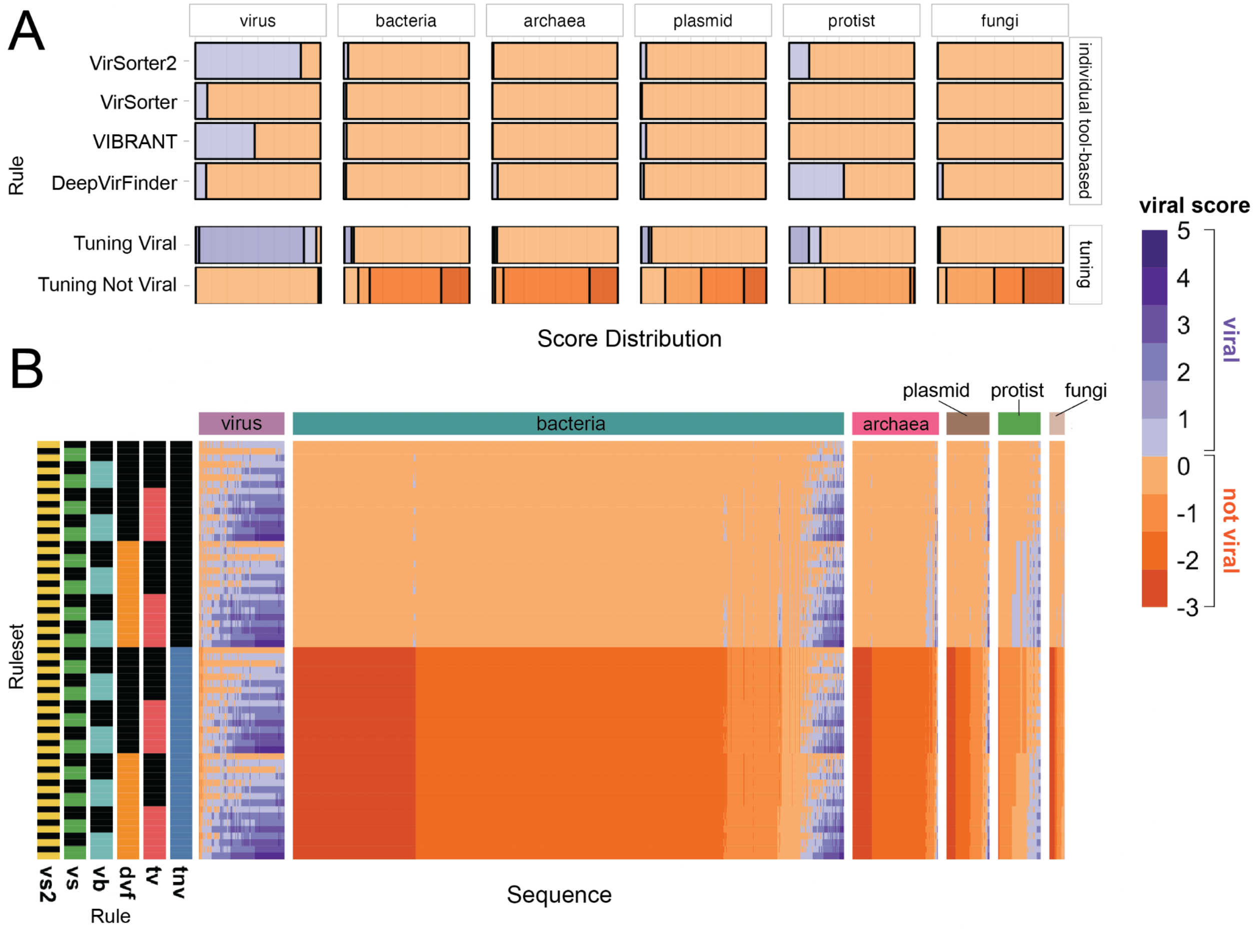


Figure 3. Comparison of different rulesets. (A) Viral score distribution of single-rule rulesets faceted by database taxonomic assignment. (B) Viral scores of all sequences across six mock metagenomes classified by each ruleset. Columns are colored based on the database taxonomic assignment (purple: virus, teal: bacteria, fuschia: archaea, brown: plasmid, green: protist, and tan: fungi) and rows based on the use of each ruleset (yellow: VirSorter2 (vs2), green: VirSorter (vs), light blue: VIBRANT (vb), orange: DeepVirFinder (dvf), pink: tuning addition (tv), and blue: tuning removal (tnv). Both A and B are colored based on the score from the model (score ≥ 1 is classified as viral (purple); score < 1 not viral (orange)).

More tools are better… to a point

Across the 63 rule combinations (rulesets), MCC, our metric for overall performance, ranges from 0.05 (DeepVirFinder) to 0.77 (VirSorter2 + Tuning Removal). With the exception of VirSorter2 (MCC=0.75), viral identification tools on their own either missed most of the viruses in the benchmarking dataset or misclassified such a large number of non-viruses that the viral signal was heavily contaminated (Figure 3A). Of the single-rule rulesets, VIBRANT performs second best (MCC of 0.55), followed by VirSorter (MCC of 0.16) and DeepVirFinder (MCC of 0.05), indicative of lower precision and recall. This may be surprising to users of these tools, given that previous reports have reported accuracies for these tools to be greater than 0.9. However, those studies used a testing set composed of 50% or more viruses compared to our 10% viral sequences and/or did not include taxonomically diverse sequences compared to our taxonomically diverse training data. In their validation, Ren et al.[21] demonstrates how the viral proportion can have a strong effect on their performance metric (AUROC). Given the lower scores and the true taxonomic distribution of environmental metagenome sequences, users likely need to be more conservative (higher classification score cutoffs) in viral calling and assume lower accuracy than previous studies report.

On average, combining rules increased MCC (Figure 4). This is driven by a statistically significant increase in recall for multi-rule rulesets compared to single-rule rulesets and for three or more rule rulesets compared to two-rule rulesets (Table S2). While average precision is constant as rules are combined, the precision of higher precision rules decreases as they are combined with lower precision rules (Table S3). Accuracy is maximized by the VirSorter2 and Tuning Removal rules, and is not improved by adding more rules.

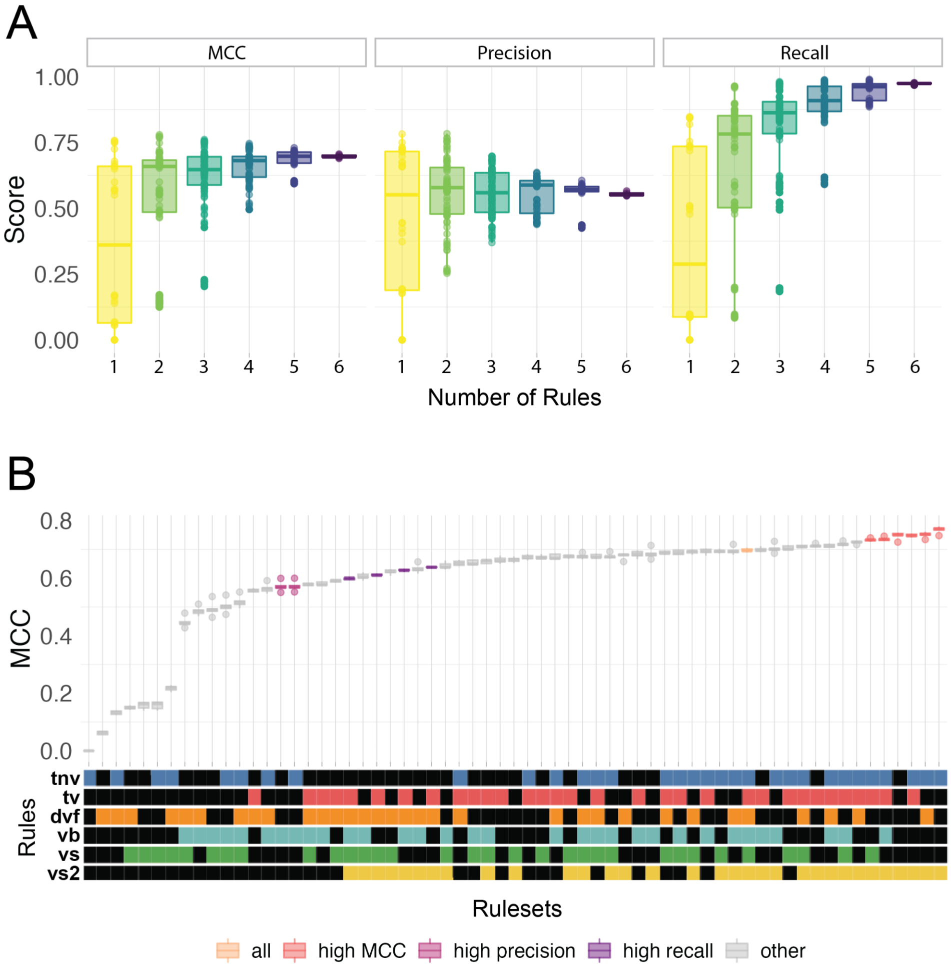


Figure 4. Performance of the 63 rulesets. (A) - Box and whisker plots of the scores representing variation in MCC, precision, and recall of different rulesets based on the number of rules used in prediction. **(B)** Ruleset accuracy (MCC) ordered by increasing MCC and colored based on the ruleset's type (light orange: all rules ("all"), salmon: equivalently high MCC rulesets ("high MCC"), pink: high precision rulesets ("high precision"), purple: high recall rulesets ("high recall"), and grey: all other rulesets ("other").

The VirSorter2-based and tuning removal rules are the most critical tools for accurate virus identification in our testing. VirSorter2's rule was in all of the top ten highest MCC rulesets and tuning removal was in eight. For comparison, none of the other single-tool rules (i.e., VirSorter2, DeepVirFinder, VIBRANT) were in more than three of the top ten highest MCC rulesets independently of the other three. Statistically equivalent MCCs ($p_{adj} > 0.05$) can be achieved by the top six rulesets ("high MCC" rulesets in Figure 4B). In the same way, statistically equivalent high precision (three rulesets) and high recall (four rulesets) rulesets were defined (Figure S2).

As the high MCC examples illustrate, many of the rulesets have a high degree of overlap with others (Figure 5). While only 4% of the combinations identified more than 90% of the same viruses as another combination, 68% of the combinations were more than 50% identical to another combination. While rulesets with VirSorter, DeepVirFinder, and VIBRANT all have more sequences in common as the number of rules in the compared sets increases, this trend is much less pronounced for VirSorter2 (Figure S3), suggesting that VirSorter, DeepVirFinder, and VIBRANT have more tool-specific viral predictions, as compared to VirSorter2. The two tools with the highest MCC in isolation, VIBRANT and VirSorter2, have many sequences in common, but very few are gained by their union in isolation (Figure S4). Of the six single-rule rulesets, VirSorter2 returns the most viruses. Each rule (other than tuning removal) combined with the VirSorter2-based rule adds more viruses with the most being contributed by VirSorter (124% compared to VirSorter2's single-rule set; Table S4). Some of this increase is based on the combined lower quality predictions of two tools. For example, if VirSorter and VirSorter2 both give a viral score of +0.5, then it is called viral by the combined prediction of the two tools (combined viral score = 1). We hypothesized that recall would increase when lower quality predictions from multiple tools were combined. However, we found that there was no effect on MCC (p-value=0.19) or recall (p-value=0.18), but a slightly significant decrease in precision (0.54 vs 0.59, p-value=$2.5*10^{-5}$) when comparing all of the rulesets with and without the +0.5 effect (Figure S5). This pattern is likely due to the rulesets being uncertain about some similar sequences that are unlikely to be viral.

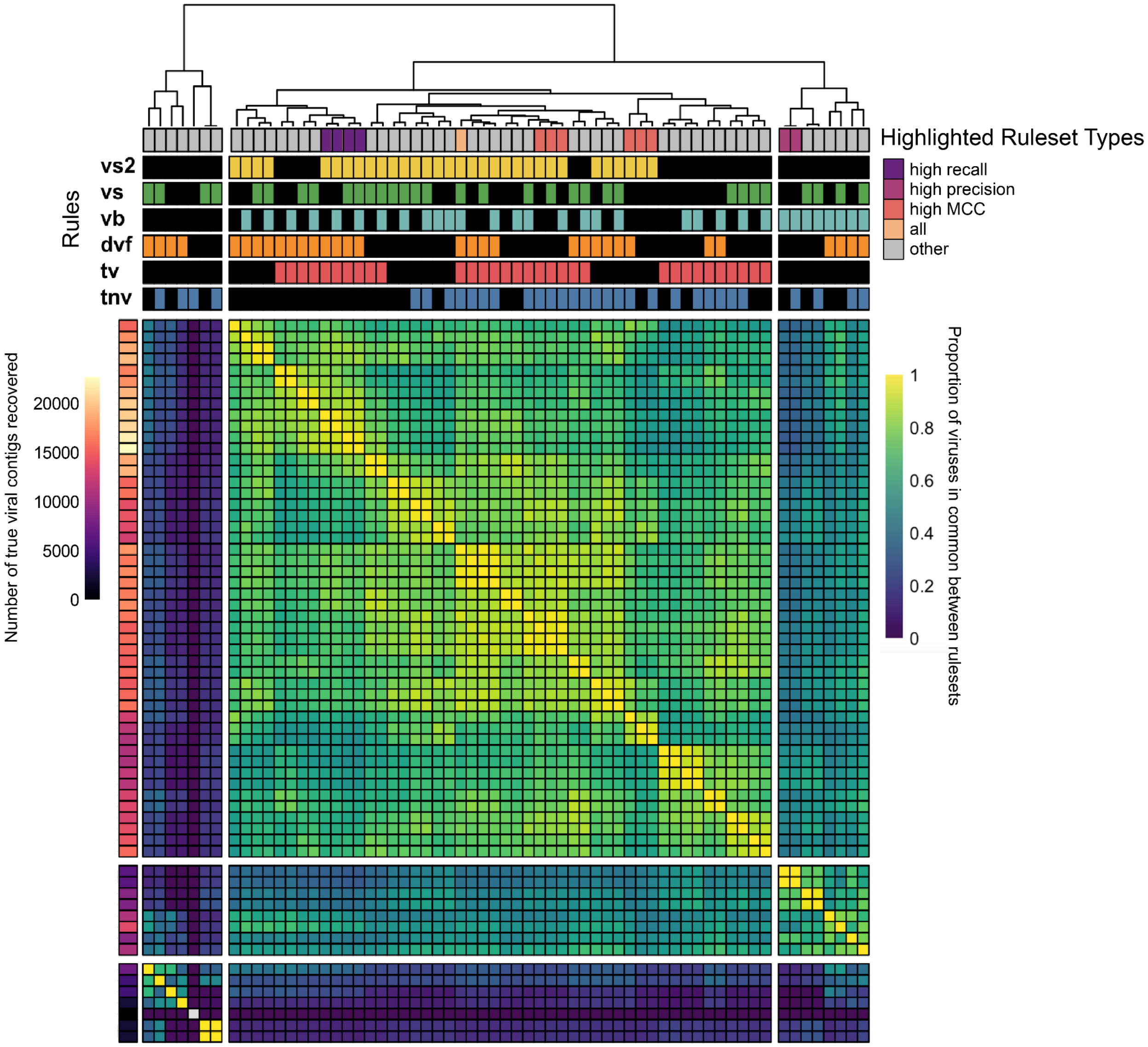


Figure 5 - Proportion of viruses in common between rulesets. Heatmap values calculated by dividing the intersection by the union of the viruses found by both rulesets. The scalebar to the right of the heatmap represents the proportion in common between the tools. The bar to the left of the heatmap represents the total number of viruses identified by each tool combination (scalebar to its left), and the ones above the heatmap whether a tool combination used VirSorter2 (VirSorter2, yellow), VirSorter (vs, green), VIBRANT (vb, light blue), DeepVirFinder (dvf, orange), tuning addition (tv, pink), or tuning removal (tnv, indigo), as well as the ruleset type (high recall, black; high precision, purple; high MCC, pink; all, tan; other, grey).

Distinguishing between viral and non-viral features

To leverage expert knowledge of the differences between viral and non-viral sequences, we designed tuning addition and removal rules (see Figure S7 for subrules). In this way, we demonstrate the value of automating the refinement of viral identification tool predictions (something that is currently a laborious manual process). In general, tuning addition improves MCC, recall, and proportion viral, while tuning removal improves MCC and precision (Figure S8). Reflecting the importance of the tuning rulesets on accuracy, 7 of the 10 highest MCC rulesets have both (Figure 4B). Very few viral sequences are flagged by multiple of the tuning removal subrules (only 0.6% of the true viral sequences). Conversely, few non-viral contigs are flagged as viral by multiple of the tuning addition subrules (2% of the true not viral sequences). Further, the majority of the non-viral sequences (bacterial, protist, archaeal, fungal, or plasmid) were removed by multiple of the tuning removal subrules (76% of the true non-viral sequences). Overall, the tuning subrules increase our prediction accuracy beyond that of the individual tools.

The tuning removal rules increase our confidence that false positives will be removed, even if some of the viral identification rules mislabel them as viral. Based on our benchmarking in this paper, we would urge caution when using some of the new automated snakemake-based pipelines for sequence identification that combine the output of multiple tools. We found that having rules to remove sequences that were called viral by individual (or combinations of) tools was an essential component of the accuracy of our high MCC rulesets.

Even with the tuning rules, we were unable to simultaneously improve both precision and recall beyond approximately 0.75 (Figure S9). Improvements in recall past this point led to sacrifices in precision. This is because, to recover more viruses, it becomes necessary to rely more on genes of unknown provenance. These may include non-viral genes, particularly eukaryotes, which were not represented in DeepVirFinder's, VirSorter's or CheckV's reference datasets (Figure S10). Further, many true viral features overlap with nonviral features (Figure S10) due to our imperfect knowledge of what distinguishes viruses and non-viruses (and the overlap between viral and host sequences). In the effort to capture all the true viruses, nonviruses with these features are misclassified as viral. This challenge is particularly acute when trying to accurately classify both short sequences[12,21] and viral types underrepresented in our testing data (i.e., the accuracy of the "high MCC" ruleset is highest for dsDNAphages compared to other viral sequence types) (Figure S11).

Since the highest MCC we could obtain was 0.77, we were curious if there were any patterns in the types of sequences being misclassified. To our surprise, the "false positive" sequences labeled as bacteria by the NCBI database, but classified as viral by our high MCC ruleset, looked more "viral" than the viruses themselves; that is, the proportion of the sequence labeled as viral orthologous gene (VOG) by VIBRANT was higher in the misclassified bacteria than the viruses (p-value $< 2.2*10^{-16}$) (Figure 7). The false positive plasmids have a similar proportion of viral genes to the viruses (p-value = 0.63). For all three sequence types, the proportion of the sequence that is VOG increases with the number of VOGs in that sequence (data not shown). For bacteria, there is a small set of short sequences with a highly viral signature (sequence predominantly VOGs).

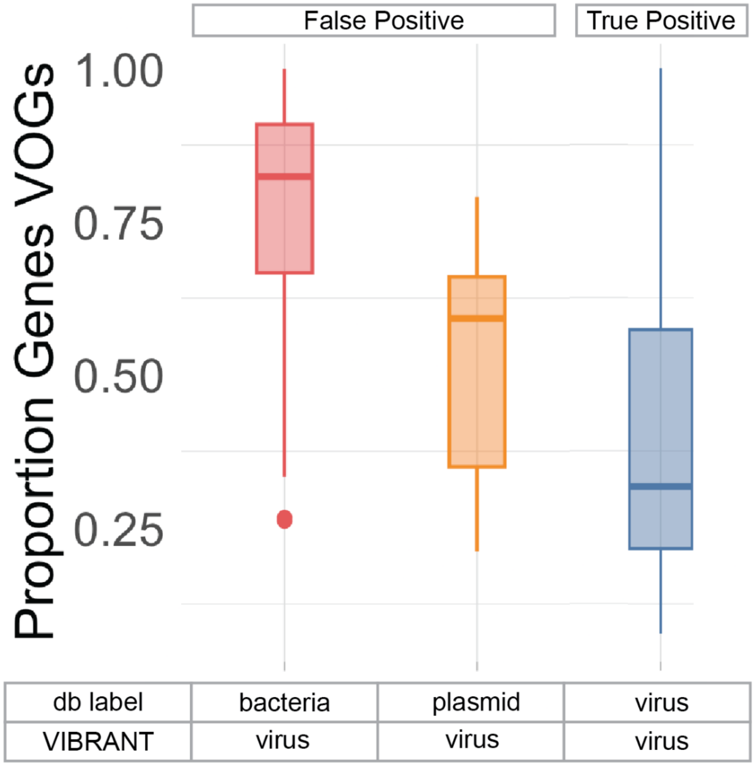


Figure 7 - Box and whisker plot of the proportion of genes on a contig with a VOG annotation based on VIBRANT broken down by sequence type (for the high MCC rules). Only sequences classified as viral by VIBRANT are included in this figure (which did not classify any archaea, fungi, or protists as viral).

Given the high proportion of viral genes on the misclassified sequences and the known problem of phage genomes being mislabeled as bacteria, plasmids, and satellite chromosomes in NCBI, we investigated these sequences further. We first checked to determine whether the bacterial false positive sequences were ΦX 174 or *E. coli*, as ΦX contamination is a known problem across NCBI database sequences due to its use in Illumina libraries.[50] Only approximately 20 ΦX sequences were taxonomically identified by Kaiju. However, approximately 1% of the viruses, 9% of the bacteria, and 5% of the plasmids were identified as *E. coli* by Kaiju. The majority of these were called not viral by our "high MCC" ruleset; and likely represent true *E. coli* bacteria sequences and thus do not explain our high degree of false positives.

Instead, these false positives represent two types of mislabeled sequences: (1) free viruses infecting the pure culture (Figure 8A and 8B) and (2) prophages found in their genomes (Figure 8C). Altogether, this further supports the known problem of phage sequences not having been removed before being deposited on NCBI. Simply put, if researchers are not looking for viruses, they will not find any.

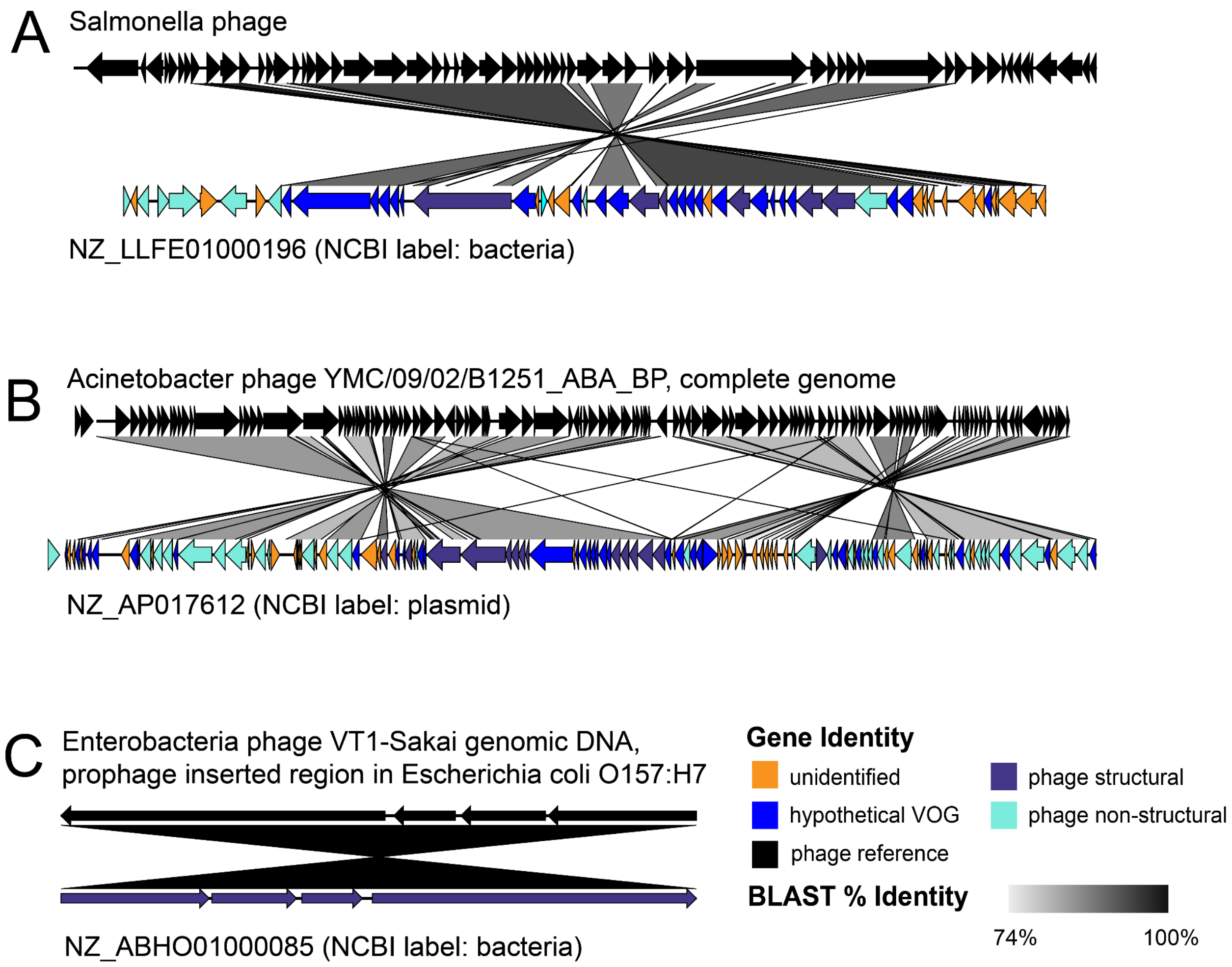

A
Salmonella phage
NZ_LLFE01000196 (NCBI label: bacteria)
B
Acinetobacter phage YMC/09/02/B1251_ABA_BP, complete genome
NZ_AP017612 (NCBI label: plasmid)
C
Enterobacteria phage VT1-Sakai genomic DNA, prophage inserted region in Escherichia coli O157:H7
NZ_ABHO01000085 (NCBI label: bacteria)
Gene Identity
unidentified
phage structural
hypothetical VOG
phage non-structural
phage reference
BLAST % Identity
74%
100%

Figure 8 - Visualizing three representative false positive testing set sequences and their top NCBI Blast hit. (A) NZ_LLFE01000196 (NCBI label: bacteria) versus "Acinetobacter phage YMC/09/02/B1251_ABA_BP, complete genome". (B) NZ_AP017612 (NCBI label: plasmid) versus Salmonella phage SSU5. (C) NZ_ABHO01000085 (NCBI label: bacteria) versus "Enterobacteria phage VT1-Sakai genomic DNA, prophage inserted region in *Escherichia coli* O157:H7". All genes of the testing set sequences are labeled by their gene identity (orange: unidentified, blue: hypothetical VOG, purple: phage structural, and bright blue: phage non-structural. Greyscale bar represent BLAST percent identity score.

Overall, we found that a significant proportion of non-viral sequences are being called viral due to inaccurate labeling of the testing data. Given this inaccurate labeling, we consider our performance statistics to be conservative, and that the viral identification tools and rulesets tested are likely more accurate than we report. Future studies may validate this claim with more accurate and diverse testing sets. Mislabeled sequences pose a serious challenge for viral identification tool development and benchmarking, which all rely, at least partially, on accurately labeled sequences to develop their models. To overcome this challenge, we propose requiring a viral screening step before NCBI sequence uploading. As knowledge of known viral versus host genes was an essential feature for distinguishing between sequences in our training data (and the most important feature for the machine learning classifier), expanding our knowledge of known viral genes is one of the best ways we can improve such a screening step (and viral identification generally).

Discovering novel viruses and identifying their genes is a known problem.[51] Virus sequence libraries are still relatively small compared to bacteria and eukarya sequence libraries despite viruses being at least two magnitudes more numerous and possessing greater sequence diversity.[52,53] Virus sequence libraries are also biased towards a few highly researched domains (e.g., oceans, human pathogens)[54] and methods to extract and identify viruses have led to taxonomic biases.[55] For example, the rules outlined in the "Viral sequence identification SOP with VirSorter2" developed by Guo et al.[48] were created based on manual curation of bulk soil and marine metagenomes.[12] By addressing these challenges, improved viral identification and more accurate ecological conclusions will become possible.

Environmental comparisons

We applied all of the rulesets to publicly available mixed metagenomic assemblies from Lake Erie water, Lake Michigan water, drinking water from the United Kingdom and Netherlands, wastewater from the Ann Arbor Wastewater Treatment Plant, and Tara global oceans (Table 2, Figure S12). To compare the performance of each ruleset, we analyzed the proportion of viruses recovered. A higher proportion of viral sequences were identified in the virus-enriched samples (~45%) compared to the non-enriched metagenomes (7-19%).While the proportion of viral sequences is dependent on the environment and sample-to-sequencing methods, the numbers for non-enriched samples are inline with those previously reported and expectations based on the relative abundance and genome sizes of viruses[15,56].

Table 2 - Simple metadata and proportion viral for the 5 environments tested.

Environment	Location	Size Fractions	Proportion viral by the best "High MCC" and "High Precision" ruleset	Proportion viral by the best "High Recall" ruleset	Proportion viral by the best "all" ruleset	Accession Number
oligotrophic lake water	Lake Michigan	<0.22μm	0.46	0.95	0.81	IMG GOLD 1072636
wastewater	Ann Arbor, Michigan, USA	<0.45μm	0.46	0.91	0.77	NCBI: PRJNA853368
eutrophic lake water	Lake Erie	>0.22μm, >100μm, 53-100μm, 3-53μm	0.12	0.32	0.23	unpublished
		<0.22μm	0.44	0.97	0.83	unpublished
drinking water	United Kingdom and Netherlands	≥0.2μm	0.16	0.69	0.34	NCBI: PRJNA533545
marine water	Tara Global Oceans	0.2–3 μm	0.07	0.22	0.12	EBI: PRJEB402

Across the rulesets, the trends in proportion of viruses recovered mimics the behavior of the testing data (Figure 9 and S13). The "high recall" rulesets return the most viruses, while the "high precision" rulesets returned many fewer (Figure 9 and Table 2). For the virus-enriched samples, nearly all contigs were called viral by the "high recall" rulesets (Figure 9, S12).

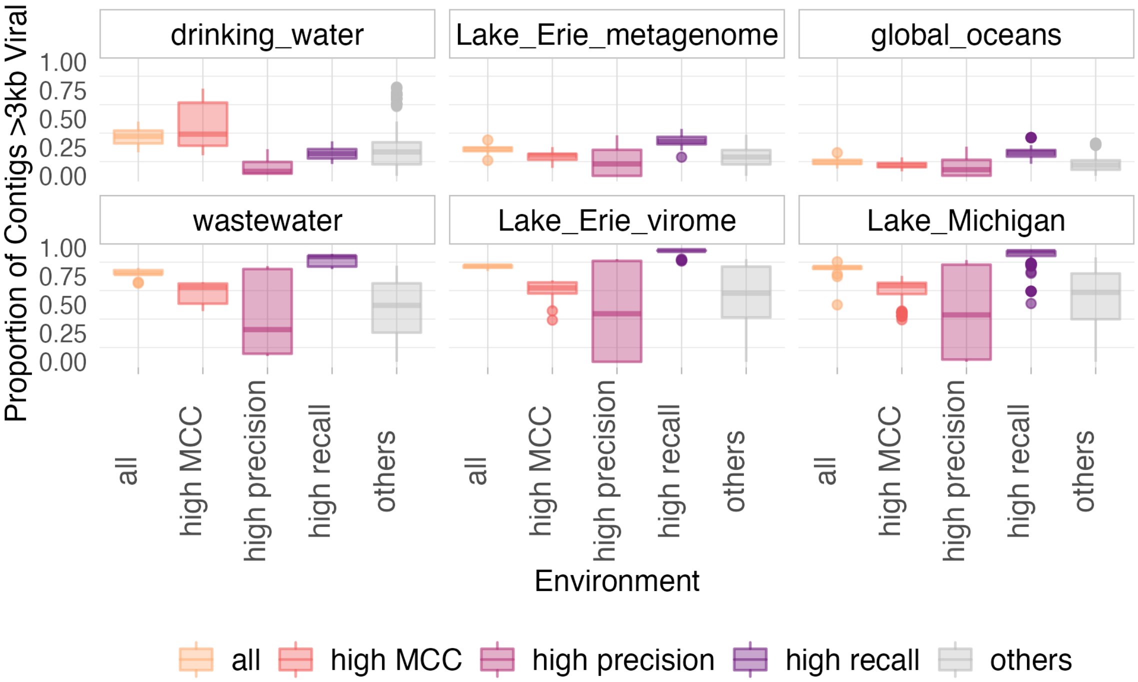


Figure 9 - Proportion of viruses predicted by each tool combination across five environmental datasets. Rulesets are grouped based on the accuracy type on the testing set. Vs2+tnv and vs2 were in both the high MCC and high precision groups.

Implications and Future Work

In silico prediction of viral sequences is a critical first step to any viral metagenomic study. There is a rapidly increasing number of *in silico* viral prediction tools available, many of which have been benchmarked against each other.[12,14,20] However, previous studies have not systematically compared the effect of using a combination of these tools. As downstream analyses and conclusions of viral ecology are predicated on accurate viral prediction, choosing the best tool outputs is paramount. Through the above benchmarking, we demonstrated that specific two and three-rule rulesets provide the highest precision with only minor sacrifices in recall (and not combining all rules).

Our recommendations based on this study vary depending on research question and experimental design. For a typical study investigating viral diversity and functional potential from a mixed metagenome, we recommend our "high MCC" ruleset (VirSorter2 with tuning removal). If eukaryotes were filtered out of the sample (< 3 µm fraction), the tuning additional rule may increase recall. Lastly, if viruses were enriched for (e.g., through filtering out bacteria or other experimental methods), the tuning removal ruleset may be eliminated. We urge caution in this case, as the tuning removal rules assist with removing non-viral contamination. Based on the accuracy of the single-rule sets on the short fragments, we do not recommend researchers to use any of these tools in isolation on mixed metagenomes. For researchers seeking to minimize the number of tools they use, we recommend using VirSorter2 or VIBRANT with some sort of tuning based on CheckV. From our analyses here, the benefit in removing non-viral contamination justifies the extra processing. VirSorter2 has a comparable MCC to multitool rules (though its high recall comes with the caveat of more false positives). Although VIBRANT recovers fewer viruses, its predictions are very precise. We do not recommend using VirSorter (poor recall) or DeepVirFinder (poor precision) in isolation.

One limitation of this work is that available testing data includes sequences that were part of the tools' original training and testing data. All tools included in this study were trained in part using NCBI sequences that overlap with our testing data; however, from the performance statistics this does not appear to have led to overfitting. Additionally, as the most comprehensive set of non-NCBI viruses available, the VirSorter2 non-RefSeq genome set was included in our testing data to expand the diversity of viruses our rulesets could be challenged against. While it was also used to train and test VirSorter2, the performance of VIBRANT was comparable to VirSorter2 in the VirSorter2 paper on the preponderance of that data[12] and these sequences represent a small portion of our testing data. As additional curated viral sets are published,[57] our rules can be tested against those providing further information about the limitations and scope of our rules.

Future work will also allow the effects of combining viral identification tools on a virome's community composition to be more systematically tested across habitats. While the proportion of viruses recovered for these aquatic environments are comparable to previous studies[56], testing against metagenomic datasets from more diverse environments could further validate if the high MCC rulesets can recover the expected proportion of viruses from additional environments.

We focused on tools that were developed primarily for bacteriophage identification. We did not evaluate tools that were specifically for human pathogens or eukaryotic viruses more broadly. Additionally, utilizing new tools for plasmid and eukaryotic sequence identification[58,59] may improve viral identification tool pipelines. However, assessing these tools was beyond the scope of our analyses here.

By rigorously comparing multi-tool approaches, we assess a strategy that is commonly employed; and conclude that it should only be used with great caution. With the rapid rate of new tools being introduced, this paper offers a blueprint for considering which sequence features best delineate viruses from each non-virus sequence type respectively. Ultimately, increasing the proportion of high-confidence viruses identified from mixed metagenomic datasets will enable more accurate ecological analyses by decreasing contamination of the viral signal, particularly from eukaryotic sequences.

Conclusion

Accurately predicting novel viral sequences is challenging due to similarity in features between viruses and their hosts. In this paper, we tested the benefits of combining multiple tools for viral identification. We found that the highest accuracy resulted from two and three rule rulesets; and caution against simply combining together viral identification tools. For most applications, we recommend a combination of VirSorter2 and tuning rules based on features of viral and not viral sequences. In summary, we were able to improve viral identification through particular combinations of rules based on multiple viral identification tools and classifiers.

References

(1) Guidi, L.; Chaffron, S.; Bittner, L.; Eveillard, D.; Larhlimi, A.; Roux, S.; Darzi, Y.; Audic, S.; Berline, L.; Brum, J. R.; Coelho, L. P.; Espinoza, J. C. I.; Malviya, S.; Sunagawa, S.; Dimier, C.; Kandels-Lewis, S.; Picheral, M.; Poulain, J.; Searson, S.; Stemmann, L.; Not, F.; Hingamp, P.; Speich, S.; Follows, M.; Karp-Boss, L.; Boss, E.; Ogata, H.; Pesant, S.; Weissenbach, J.; Wincker, P.; Acinas, S. G.; Bork, P.; de Vargas, C.; Iudicone, D.; Sullivan, M. B.; Raes, J.; Karsenti, E.; Bowler, C.; Gorsky, G. Plankton Networks Driving Carbon Export in the Oligotrophic Ocean. *Nature* **2016**, *532* (7600), 465–470. https://doi.org/10.1038/nature16942.

(2) Wilhelm, S. W.; Suttle, C. A. Viruses and Nutrient Cycles in the Sea: Viruses Play Critical Roles in the Structure and Function of Aquatic Food Webs. *BioScience* **1999**, *49* (10), 781–788. https://doi.org/10.2307/1313569.

(3) Howard-Varona, C.; Lindback, M. M.; Bastien, G. E.; Solonenko, N.; Zayed, A. A.; Jang, H.; Andreopoulos, B.; Brewer, H. M.; Glavina del Rio, T.; Adkins, J. N.; Paul, S.; Sullivan, M. B.; Duhaime, M. B. Phage-Specific Metabolic Reprogramming of Virocells. *ISME J.* **2020**, *14* (4), 881–895. https://doi.org/10.1038/s41396-019-0580-z.

(4) Hurwitz, B. L.; U'Ren, J. M. Viral Metabolic Reprogramming in Marine Ecosystems. *Curr. Opin. Microbiol.* **2016**, *31*, 161–168. https://doi.org/10.1016/j.mib.2016.04.002.

(5) Beumer, A.; Robinson, J. B. A Broad-Host-Range, Generalized Transducing Phage (SN-T) Acquires 16S RRNA Genes from Different Genera of Bacteria. *Appl. Environ. Microbiol.* **2005**, *71* (12), 8301–8304. https://doi.org/10.1128/AEM.71.12.8301-8304.2005.

(6) Göller, P. C.; Elsener, T.; Lorgé, D.; Radulovic, N.; Bernardi, V.; Naumann, A.; Amri, N.; Khatchatourova, E.; Coutinho, F. H.; Loessner, M. J.; Gómez-Sanz, E. Multi-Species Host Range of Staphylococcal Phages Isolated from Wastewater. *Nat. Commun.* **2021**, *12* (1), 6965. https://doi.org/10.1038/s41467-021-27037-6.

(7) Sullivan, M. B. Viromes, Not Gene Markers, for Studying Double-Stranded DNA Virus Communities. *J. Virol.* **2015**, *89* (5), 2459–2461. https://doi.org/10.1128/JVI.03289-14.

(8) Drake, J. W. The Distribution of Rates of Spontaneous Mutation over Viruses, Prokaryotes, and Eukaryotes. *Ann. N. Y. Acad. Sci.* **1999**, *870* (1), 100–107. https://doi.org/10.1111/j.1749-6632.1999.tb08870.x.

(9) Peck, K. M.; Lauring, A. S. Complexities of Viral Mutation Rates. *J. Virol.* **2018**, *92* (14), e01031-17. https://doi.org/10.1128/JVI.01031-17.

(10) O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretdin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.; Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D. Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation. *Nucleic Acids Res.* **2016**, *44* (Database issue), D733–D745. https://doi.org/10.1093/nar/gkv1189.

(11) Ponsero, A. J.; Hurwitz, B. L. The Promises and Pitfalls of Machine Learning for Detecting Viruses in Aquatic Metagenomes. *Front. Microbiol.* **2019**, *10.*

(12) Guo, J.; Bolduc, B.; Zayed, A. A.; Varsani, A.; Dominguez-Huerta, G.; Delmont, T. O.; Pratama, A. A.; Gazitúa, M. C.; Vik, D.; Sullivan, M. B.; Roux, S. VirSorter2: A Multi-Classifier, Expert-Guided Approach to Detect Diverse DNA and RNA Viruses. *Microbiome* **2021**, *9* (1), 37. https://doi.org/10.1186/s40168-020-00990-y.

(13) Andrade-Martínez, J. S.; Camelo Valera, L. C.; Chica Cárdenas, L. A.; Forero-Junco, L.; López-Leal, G.; Moreno-Gallego, J. L.; Rangel-Pineros, G.; Reyes, A. Computational Tools for the Analysis of Uncultivated Phage Genomes. *Microbiol. Mol. Biol. Rev.* **2022**, *86* (2), e00004-21. https://doi.org/10.1128/mmbr.00004-21.

(14) 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* (1), 90. https://doi.org/10.1186/s40168-020-00867-0.

(15) Hegarty, B.; Dai, Z.; Raskin, L.; Pinto, A.; Wigginton, K.; Duhaime, M. A Snapshot of the Global Drinking Water Virome: Diversity and Metabolic Potential Vary with Residual Disinfectant Use. *Water Res.* **2022**, *218*, 118484. https://doi.org/10.1016/j.watres.2022.118484.

(16) Gregory, A. C.; Zayed, A. A.; Conceição-Neto, N.; Temperton, B.; Bolduc, B.; Alberti, A.; Ardyna, M.; Arkhipova, K.; Carmichael, M.; Cruaud, C.; Dimier, C.; Domínguez-Huerta, G.; Ferland, J.; Kandels, S.; Liu, Y.; Marec, C.; Pesant, S.; Picheral, M.; Pisarev, S.; Poulain, J.; Tremblay, J.-É.; Vik, D.; Acinas, S. G.; Babin, M.; Bork, P.; Boss, E.; Bowler, C.; Cochrane, G.; de Vargas, C.; Follows, M.; Gorsky, G.; Grimsley, N.; Guidi, L.; Hingamp, P.; Iudicone, D.; Jaillon, O.; Kandels-Lewis, S.; Karp-Boss, L.; Karsenti, E.; Not, F.; Ogata, H.; Pesant, S.; Poulton, N.; Raes, J.; Sardet, C.; Speich, S.; Stemmann, L.; Sullivan, M. B.; Sunagawa, S.; Wincker, P.; Babin, M.; Bowler, C.; Culley, A. I.; de Vargas, C.; Dutilh, B. E.; Iudicone, D.; Karp-Boss, L.; Roux, S.; Sunagawa, S.; Wincker, P.; Sullivan, M. B. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **2019**, *177* (5), 1109-1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

(17) Rocha, U. N. da; Kasmanas, J. C.; Kallies, R.; Saraiva, J. P.; Toscan, R. B.; Štefanič, P.; Bicalho, M. F.; Correa, F. B.; Baştürk, M. N.; Fousekis, E.; Barbosa, L. M. V.; Plewka, J.; Probst, A. J.; Baldrian, P.; Stadler, P. F.; Consortium, C.-T. MuDoGeR: Multi-Domain Genome Recovery from Metagenomes Made Easy. bioRxiv August 11, 2022, p 2022.06.21.496983. https://doi.org/10.1101/2022.06.21.496983.

(18) Vik, D.; Gazitúa, M. C.; Sun, C. L.; Zayed, A. A.; Aldunate, M.; Mulholland, M. R.; Ulloa, O.; Sullivan, M. B. Genome-Resolved Viral Ecology in a Marine Oxygen Minimum Zone. *Environ. Microbiol.* **2021**, *23* (6), 2858–2874. https://doi.org/10.1111/1462-2920.15313.

(19) Tisza, M. J.; Belford, A. K.; Domínguez-Huerta, G.; Bolduc, B.; Buck, C. B. Cenote-Taker 2 Democratizes Virus Discovery and Sequence Annotation. *Virus Evol.* **2021**, *7* (1), veaa100. https://doi.org/10.1093/ve/veaa100.

(20) Nayfach, S.; Camargo, A. P.; Schulz, F.; Eloe-Fadrosh, E.; Roux, S.; Kyrpides, N. C. CheckV Assesses the Quality and Completeness of Metagenome-Assembled Viral Genomes. *Nat. Biotechnol.* **2021**, *39* (5), 578–585. https://doi.org/10.1038/s41587-020-00774-7.

(21) Ren, J.; Song, K.; Deng, C.; Ahlgren, N. A.; Fuhrman, J. A.; Li, Y.; Xie, X.; Poplin, R.; Sun, F. Identifying Viruses from Metagenomic Data Using Deep Learning. *Quant. Biol.* **2020**, *8* (1), 64–77. https://doi.org/10.1007/s40484-019-0187-4.

(22) Czeczko, P.; Greenway, S. C.; de Koning, A. P. J. EzMap: A Simple Pipeline for Reproducible Analysis of the Human Virome. *Bioinforma. Oxf. Engl.* **2017**, *33* (16), 2573–2574. https://doi.org/10.1093/bioinformatics/btx202.

(23) Amgarten, D.; Braga, L. P. P.; da Silva, A. M.; Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front. Genet.* **2018**, *9*.

(24) Antipov, D.; Raiko, M.; Lapidus, A.; Pevzner, P. A. MetaviralSPAdes: Assembly of Viruses from Metagenomic Data. *Bioinformatics* **2020**, *36* (14), 4126–4129. https://doi.org/10.1093/bioinformatics/btaa490.

(25) *MetaPhinder—Identifying Bacteriophage Sequences in Metagenomic Data Sets | PLOS ONE*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0163111 (accessed 2023-05-08).

(26) Deaton, J.; Yu, F. B.; Quake, S. R. Mini-Metagenomics and Nucleotide Composition Aid the Identification and Host Association of Novel Bacteriophage Sequences. *Adv. Biosyst.* **2019**, *3* (11), 1900108. https://doi.org/10.1002/adbi.201900108.

(27) Arndt, D.; Grant, J. R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D. S. PHASTER: A Better, Faster Version of the PHAST Phage Search Tool. *Nucleic Acids Res.* **2016**, *44* (W1), W16–W21. https://doi.org/10.1093/nar/gkw387.

(28) Starikova, E. V.; Tikhonova, P. O.; Prianichnikov, N. A.; Rands, C. M.; Zdobnov, E. M.; Ilina, E. N.; Govorun, V. M. Phigaro: High-Throughput Prophage Sequence Annotation. *Bioinformatics* **2020**, *36* (12), 3882–3884. https://doi.org/10.1093/bioinformatics/btaa250.

(29) Fang, Z.; Tan, J.; Wu, S.; Li, M.; Xu, C.; Xie, Z.; Zhu, H. PPR-Meta: A Tool for Identifying Phages and Plasmids from Metagenomic Fragments Using Deep Learning. *GigaScience* **2019**, *8* (6), giz066. https://doi.org/10.1093/gigascience/giz066.

(30) Liu, F.; Miao, Y.; Liu, Y.; Hou, T. RNN-VirSeeker: A Deep Learning Method for Identification of Short Viral Sequences From Metagenomes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2022**, *19* (3), 1840–1849. https://doi.org/10.1109/TCBB.2020.3044575.

(31) Auslander, N.; Gussow, A. B.; Benler, S.; Wolf, Y. I.; Koonin, E. V. Seeker: Alignment-Free Identification of Bacteriophage Genomes by Deep Learning. *Nucleic Acids Res.* **2020**, *48* (21), e121. https://doi.org/10.1093/nar/gkaa856.

(32) Li, Y.; Wang, H.; Nie, K.; Zhang, C.; Zhang, Y.; Wang, J.; Niu, P.; Ma, X. VIP: An Integrated Pipeline for Metagenomics of Virus Identification and Discovery. *Sci. Rep.* **2016**, *6*, 23774. https://doi.org/10.1038/srep23774.

(33) 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* (9), e0222271. https://doi.org/10.1371/journal.pone.0222271.

(34) Ren, J.; Ahlgren, N. A.; Lu, Y. Y.; Fuhrman, J. A.; Sun, F. VirFinder: A Novel k-Mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data. *Microbiome* **2017**, *5* (1), 69. https://doi.org/10.1186/s40168-017-0283-5.

(35) Wood, D. E.; Lu, J.; Langmead, B. Improved Metagenomic Analysis with Kraken 2. *Genome Biol.* **2019**, *20* (1), 257. https://doi.org/10.1186/s13059-019-1891-0.

(36) Garretto, A.; Hatzopoulos, T.; Putonti, C. VirMine: Automated Detection of Viral Sequences from Complex Metagenomic Samples. *PeerJ* **2019**, *7*, e6695. https://doi.org/10.7717/peerj.6695.

(37) Zheng, T.; Li, J.; Ni, Y.; Kang, K.; Misiakou, M.-A.; Imamovic, L.; Chow, B. K. C.; Rode, A. A.; Bytzer, P.; Sommer, M.; Panagiotou, G. Mining, Analyzing, and Integrating Viral Signals from Metagenomic Data. *Microbiome* **2019**, *7* (1), 42. https://doi.org/10.1186/s40168-019-0657-y.

(38) Abdelkareem, A. O.; Khalil, M. I.; Elaraby, M.; Abbas, H.; Elbehery, A. H. A. VirNet: Deep Attention Model for Viral Reads Identification. In *2018 13th International Conference on Computer Engineering and Systems (ICCES)*; 2018; pp 623–626. https://doi.org/10.1109/ICCES.2018.8639400.

(39) Wommack, K. E.; Bhavsar, J.; Polson, S. W.; Chen, J.; Dumas, M.; Srinivasiah, S.; Furman, M.; Jamindar, S.; Nasko, D. J. VIROME: A Standard Operating Procedure for Analysis of Viral Metagenome Sequences. *Stand. Genomic Sci.* **2012**, *6* (3), 421–433. https://doi.org/10.4056/sigs.2945050.

(40) Rampelli, S.; Soverini, M.; Turroni, S.; Quercia, S.; Biagi, E.; Brigidi, P.; Candela, M. ViromeScan: A New Tool for Metagenomic Viral Community Profiling. *BMC Genomics* **2016**, *17*, 165. https://doi.org/10.1186/s12864-016-2446-3.

(41) Roux, S.; Enault, F.; Hurwitz, B. L.; Sullivan, M. B. VirSorter: Mining Viral Signal from Microbial Genomic Data. *PeerJ* **2015**, *3*, e985. https://doi.org/10.7717/peerj.985.

(42) Miao, Y.; Liu, F.; Hou, T.; Liu, Y. Virtifier: A Deep Learning-Based Identifier for Viral Sequences from Metagenomes. *Bioinformatics* **2022**, *38* (5), 1216–1222. https://doi.org/10.1093/bioinformatics/btab845.

(43) Zhao, G.; Wu, G.; Lim, E. S.; Droit, L.; Krishnamurthy, S.; Barouch, D. H.; Virgin, H. W.; Wang, D.

VirusSeeker, a Computational Pipeline for Virus Discovery and Virome Composition Analysis. *Virology* **2017**, *503*, 21–30. https://doi.org/10.1016/j.virol.2017.01.005.

(44) Glickman, C.; Hendrix, J.; Strong, M. Simulation Study and Comparative Evaluation of Viral Contiguous Sequence Identification Tools. *BMC Bioinformatics* **2021**, *22* (1), 329. https://doi.org/10.1186/s12859-021-04242-0.

(45) Ho, S. F. S.; Wheeler, N. E.; Millard, A. D.; van Schaik, W. Gauge Your Phage: Benchmarking of Bacteriophage Identification Tools in Metagenomic Sequencing Data. *Microbiome* **2023**, *11* (1), 84. https://doi.org/10.1186/s40168-023-01533-x.

(46) Menzel, P.; Ng, K. L.; Krogh, A. Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju. *Nat. Commun.* **2016**, *7* (1), 11257. https://doi.org/10.1038/ncomms11257.

(47) Edwards, R. A.; Rohwer, F. Viral Metagenomics. *Nat. Rev. Microbiol.* **2005**, *3* (6), 504–510. https://doi.org/10.1038/nrmicro1163.

(48) Guo, J.; Vik, D.; Pratama, A. A.; Roux, S.; Sullivan, M. Viral Sequence Identification SOP with VirSorter2. **2021**.

(49) Chicco, D.; Jurman, G. The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics* **2020**, *21*, 6. https://doi.org/10.1186/s12864-019-6413-7.

(50) Mukherjee, S.; Huntemann, M.; Ivanova, N.; Kyrpides, N. C.; Pati, A. Large-Scale Contamination of Microbial Isolate Genomes by Illumina PhiX Control. *Stand. Genomic Sci.* **2015**, *10*, 18. https://doi.org/10.1186/1944-3277-10-18.

(51) Rose, R.; Constantinides, B.; Tapinos, A.; Robertson, D. L.; Prosperi, M. Challenges in the Analysis of Viral Metagenomes. *Virus Evol.* **2016**, *2* (2), vew022. https://doi.org/10.1093/ve/vew022.

(52) Jansson, J. K.; Wu, R. Soil Viral Diversity, Ecology and Climate Change. *Nat. Rev. Microbiol.* **2023**, *21* (5), 296–311. https://doi.org/10.1038/s41579-022-00811-z.

(53) Suttle, C. A. Marine Viruses — Major Players in the Global Ecosystem. *Nat. Rev. Microbiol.* **2007**, *5* (10), 801–812. https://doi.org/10.1038/nrmicro1750.

(54) Paez-Espino, D.; Eloe-Fadrosh, E. A.; Pavlopoulos, G. A.; Thomas, A. D.; Huntemann, M.; Mikhailova, N.; Rubin, E.; Ivanova, N. N.; Kyrpides, N. C. Uncovering Earth's Virome. *Nature* **2016**, *536* (7617), 425–430. https://doi.org/10.1038/nature19094.

(55) Langenfeld, K.; Chin, K.; Roy, A.; Wigginton, K.; Duhaime, M. B. Comparison of Ultrafiltration and Iron Chloride Flocculation in the Preparation of Aquatic Viromes from Contrasting Sample Types. *PeerJ* **2021**, *9*, e11111. https://doi.org/10.7717/peerj.11111.

(56) DeLong, E. F.; Preston, C. M.; Mincer, T.; Rich, V.; Hallam, S. J.; Frigaard, N.-U.; Martinez, A.; Sullivan, M. B.; Edwards, R.; Brito, B. R.; Chisholm, S. W.; Karl, D. M. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science* **2006**, *311* (5760), 496–503. https://doi.org/10.1126/science.1120250.

(57) Elbehery, A. H. A.; Deng, L. Insights into the Global Freshwater Virome. *Front. Microbiol.* **2022**, *13*.

(58) Camargo, A. P.; Roux, S.; Schulz, F.; Babinski, M.; Xu, Y.; Hu, B.; Chain, P. S. G.; Nayfach, S.; Kyrpides, N. C. You Can Move, but You Can't Hide: Identification of Mobile Genetic Elements with GeNomad. bioRxiv March 6, 2023, p 2023.03.05.531206. https://doi.org/10.1101/2023.03.05.531206.

(59) Gabrielli, M.; Dai, Z.; Delafont, V.; Timmers, P. H. A.; van der Wielen, P. W. J. J.; Antonelli, M.; Pinto, A. J. Identifying Eukaryotes and Factors Influencing Their Biogeography in Drinking Water Metagenomes. *Environ. Sci. Technol.* **2023**, *57* (9), 3645–3660. https://doi.org/10.1021/acs.est.2c09010.