Benchmarking combined informatics approaches for virus discovery: Caution is needed when combining *in silico* identification methods

Authors

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

Abstract

Background

The identification of viruses from environmental metagenomic samples using informatics tools has offered critical insights in microbiome studies. However, it remains difficult for researchers to know, for their specific study, which tool(s) and settings are best suited to minimize false positives and maximize capture of true viruses. Studies are increasingly combining multiple tool outputs attempting to recover more viruses, but no combined approach has been benchmarked for accuracy. Here, we benchmarked 63 viral identification 'rulesets' against mock metagenomes composed of publically available viral, bacterial, archaeal, fungal, and protist sequences. These rulesets are based on combinations of four single-tool rules and two multi-tool tuning rules. We applied these rulesets to various aquatic metagenomes (fresh and saltwater, drinking water, wastewater) and filtering strategies (targeting the virus-enriched and cell-enriched fractions of microbial communities) to evaluate the impact of habitat and viral enrichment on individual and combined tool performance.

Results

We found that combining rules increased viral recall, but at the expense of increased false positives. Of the 63 rulesets, 6 had MCCs (Matthew's Correlation Coefficient, our accuracy measure) that were statistically equivalent (padj ≥ 0.05) to the highest MCC ruleset (MCC=0.77). These rulesets all included VirSorter2, 5 included our "tuning removal" rule, and none used more than 4 of our 6 rules. DeepVirFinder, VIBRANT, and VirSorter were each found once in the "high MCC" rulesets, but never in combination with each other. Our validation suggests that the MCC plateau is caused by inaccurate classification of the data that viral identification tools rely on for training and validation. In the aquatic metagenomes, our "highest MCC" ruleset identified a higher proportion of viral sequences in the virus-enriched samples (44-46%) than the non-enriched, cellular metagenomes (7-19%).

Conclusion

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 the ruleset that uses VirSorter2 and our data-driven tuning removal rules. By providing a rigorous overview of the behavior of *in silico* viral identification strategies, our findings guide the use of existing viral identification tools and offer a blueprint for feature engineering of new tools that will lead to higher-confidence viral discovery in microbiome studies.

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 sequences 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 representative 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, a number of studies have combined the outputs of multiple tools to classify viral sequences to capture a greater portion of the viral signal.[15–21] 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 rigorously evaluated. 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 recommend pipelines specific to short and long-read sequences, as well as cellular metagenomes and virus-enriched samples. By returning more viral sequences with less non-viral contamination, these pipelines will enable new and more accurate insights into the ways viruses impact microbial ecosystem functions, with far-reaching implications for human and environmental health.

Methods

Creation of sequence testing set

To create a testing set for benchmarking multi-tool pipelines, we downloaded viral, bacterial, fungal, plasmid, protist, and archaeal genome sequences from the NCBI reference sequence database (RefSeq), as well as a unique set of non-RefSeq virus genomes compiled by the VirSorter2 tool developers, herein 'VirSorter2 database'[22] (Figure 1A; SI Table 1). These non-RefSeq virus genomes represent a comprehensive validated set of viruses. We originally created nine testing sets by randomly sampling sequences with replacement from these sources to create datasets that mimicked metagenomic environmental data. As the variability between testing sets plateaued at five datasets (Figure S1), five datasets were used for subsequent analyses (Figure 1A). The testing sets were designed to contain approximately 68% bacteria, 10% archaea, 10% virus (not proviruses), 5% plasmid, 5% protist, and 2% fungi sequences, totalling ~8k sequences (Figure 1B). The proportion of sequences was chosen to be representative of cellular metagenomic data, which are dominated by bacterial sequences, and reportedly contain ~10% viral sequences.[15,23] 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 VirSorter2 database. As DeepVirFinder requires sequences to be less than 2,100 kb, a custom python script was written to trim the testing set sequences to meet this length cutoff. We did not expect trimming to impact results since the largest phage genome reported is smaller than this length cutoff (735 kb).[24] Further, only sequences longer than 3000 bp were used because previous reports demonstrate tool accuracy significantly decreases below 3000 bp.[12]

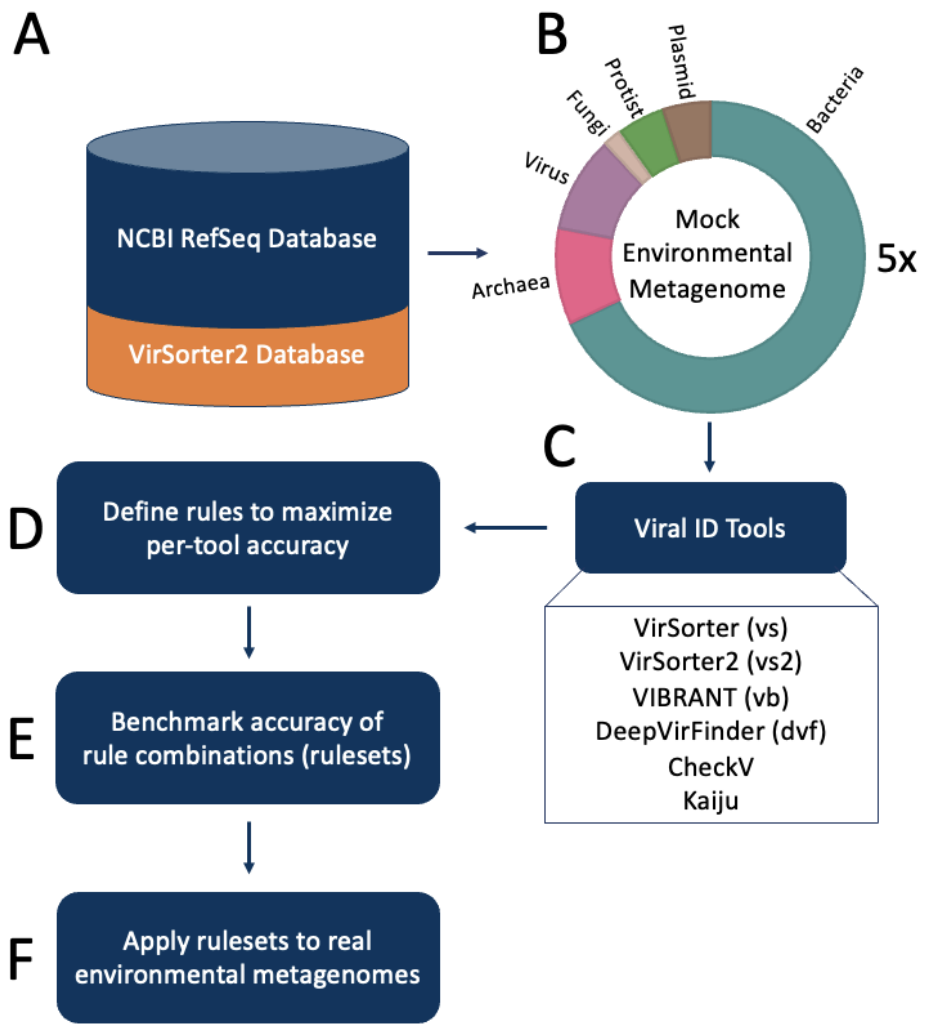

Figure 1. *Overview of study workflow. (A) A set of sequences > 3 kb were randomly downloaded from NCBI and a curated Non-RefSeq viral genomes database ('VirSorter2 database') to (B) generate five mock environmental metagenomes. (C) Mock metagenomes were run through six viral identification tools; (D) where score cutoffs were defined based on each tool's outputs to maximize their accuracy. (E) Accuracy was then assessed for each tool combination to guide the development of defined "rulesets." (F) Rulesets were then used to classify sequences from six real-world aquatic metagenomes: three cell-enriched metagenomes and three virus-enriched metagenomes.*

Selection of viral identification tools

29 viral identification tools[12,14,25,25–50] were found through literature search and assessed to determine their suitability for inclusion in this study (Table S1). 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 (e.g., not only for the human gut), 3) is designed to target viral sequences of lengths greater than 3 kb, 4) can classify millions of sequences 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,[12,14,51] 8) is not specific to prophages, and 9) was published before June 2022.

Four of the 29 viral identification tools met the above criteria: DeepVirFinder,[27] VIBRANT,[14] VirSorter,[47] and VirSorter2.[12] While not designed strictly as viral identification tools, Kaiju[52] and CheckV[26] were used to tune the viral predictions in our test sets All six are referred to as "viral identification tools" or simply "tools" in this manuscript (Table 1, Figure 1C).

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[26]	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[27]	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[47]	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[52]	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

The 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. The '-virome' flag was not used for the eutrophic lake water virome assemblies, but the eutrophic lake water virome still was most similar to the other virus-enriched samples (Figure 7), and other studies showed the number of contigs recovered were

very similar[53]. All tools were run with default parameters except for specifying a 3 kb contig length cutoff.

Design of viral identification rules

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 which sequences to call "viral." To aid in the process of choosing rules and cutoffs for predicting viral sequences, we designed six rules (Figure 1D). Each rule includes at least two *subrules* that use outputs from the six selected tools. These subrules were designed through two processes:

1) *Evaluation of 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.[12,15,16,54]

2) *Curation and evaluation of biological features:* some viral identification tools generate information describing biological features for each sequence, 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, VIBRANT identifies virus orthologous genes (VOGs). These biological features were used to create classification criteria (Figure 2) to distinguish viral and non-viral sequences.

The developers' recommendations for calling a sequence viral served as a baseline for assigning a given sequence a *'viral score'* that captured the relative likelihood that a given sequence was viral. The cutoffs were then adjusted to maximize the number of true viral sequences being classified as viral (true positives) and minimize the number of non-viral sequences being classified as viral (false positives) in the mock environmental microbial communities. We first defined four single-tool rules (Figure 2A-D) derived from four viral identification tools (i.e., VIBRANT, VirSorter, VirSorter2, DeepVirFinder). Next, we defined two sets of tuning rules derived from outputs of different tools that indicated a strongly viral or non-viral signal (Figure 2E-F). These included: (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.

Each putative viral sequence is assigned a numerical value by each rule; these are combined to give a value that comprises the sequence's *viral score* (Figure 2G-H). Sequences with a final *viral score* ≥ 1 were considered viral, and scores < 1 non-viral (Figure 2G).

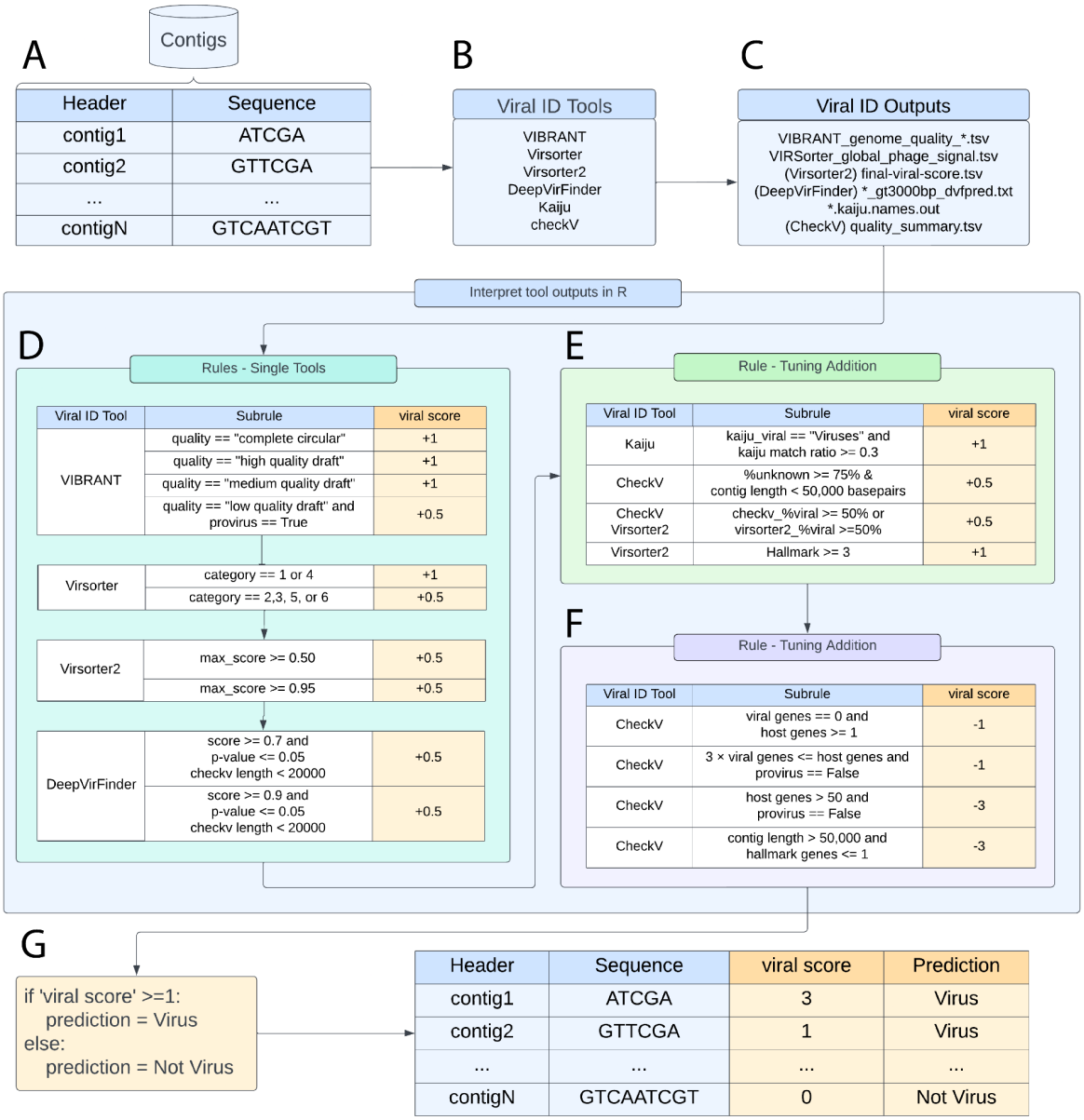

Figure 2. Diagram of approach details. *(A) Sequences are first processed by each (B) viral identification tool. Next, (C) the tool outputs are programmatically post-processed to generate a viral score based on both (D) single-tool rules and the data-driven creation of (E) tuning addition and (F) tuning removal processes. (G) This combined post-processing generates a viral score that indicates whether each sequence input is predicted as "Virus" or "Not Virus." Subrules are scored based on the confidence of the prediction: low-confidence = ±0.5, confident = ±1, highly confident not viral = -3.*

Evaluation of viral identification rulesets

Ultimately, 63 combinations (rulesets) of these six rules were evaluated by comparing the viral score of each sequence to the classification assigned by the database (Figure 1E). 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 the relative proportion of false positives, false negatives, true positives, and true negatives)[55] were calculated.

Application of rulesets to environmental metagenomes

All rulesets were used to identify viruses from five previously published environmental datasets representing different aquatic environments and size fractions (Table 2, Figure 1F). 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), where the `-virome` flag was used for VIBRANT and VirSorter to reduce sensitivity, as recommended by the developers of those tools given that a greater fraction of total sequences are expected to be viral in virus-enriched samples. All tools were run with a 3 kb cutoff to remove small sequences.

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

Environment	Location	Size Fractions	Number of sequences > 3kb	Accession Number
oligotrophic lake water	Lake Michigan	<0.22µm	253,673	IMG GOLD: 1072636
wastewater	Ann Arbor, Michigan, USA	<0.45µm	411,557	NCBI: PRJNA853368
eutrophic lake water	Lake Erie	<0.22µm	18,672	NCBI: PRJNA988094
		>0.22µm	15,181	NCBI: PRJNA988094
drinking water	United Kingdom and Netherlands	≥0.2µm	133,391	NCBI: PRJNA533545
marine water	Tara Global Oceans (global)	0.2–3 µm	1,864,711	NCBI: PRJEB1787

Code availability

The scripts to run the full testing pipeline presented above, including those to identify viral contigs from these tool outputs, example runs, and classifier outputs on the environmental metagenomic samples, are freely available at https://github.com/DuhaimeLab/VSTE.

Results and 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 (rulesets) of the six rules 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; Figure 2A-D) and two additional tuning rules: tuning removal (which removes predictions using Kaiju, CheckV, VirSorter2, VirSorter, and VIBRANT outputs) and tuning addition (which adds predictions using Kaiju, CheckV, and VirSorter2 outputs; Figure 2E-F). In this section, we compare the performance of our rulesets, elaborate on their strengths and weaknesses, and provide recommendations for use.

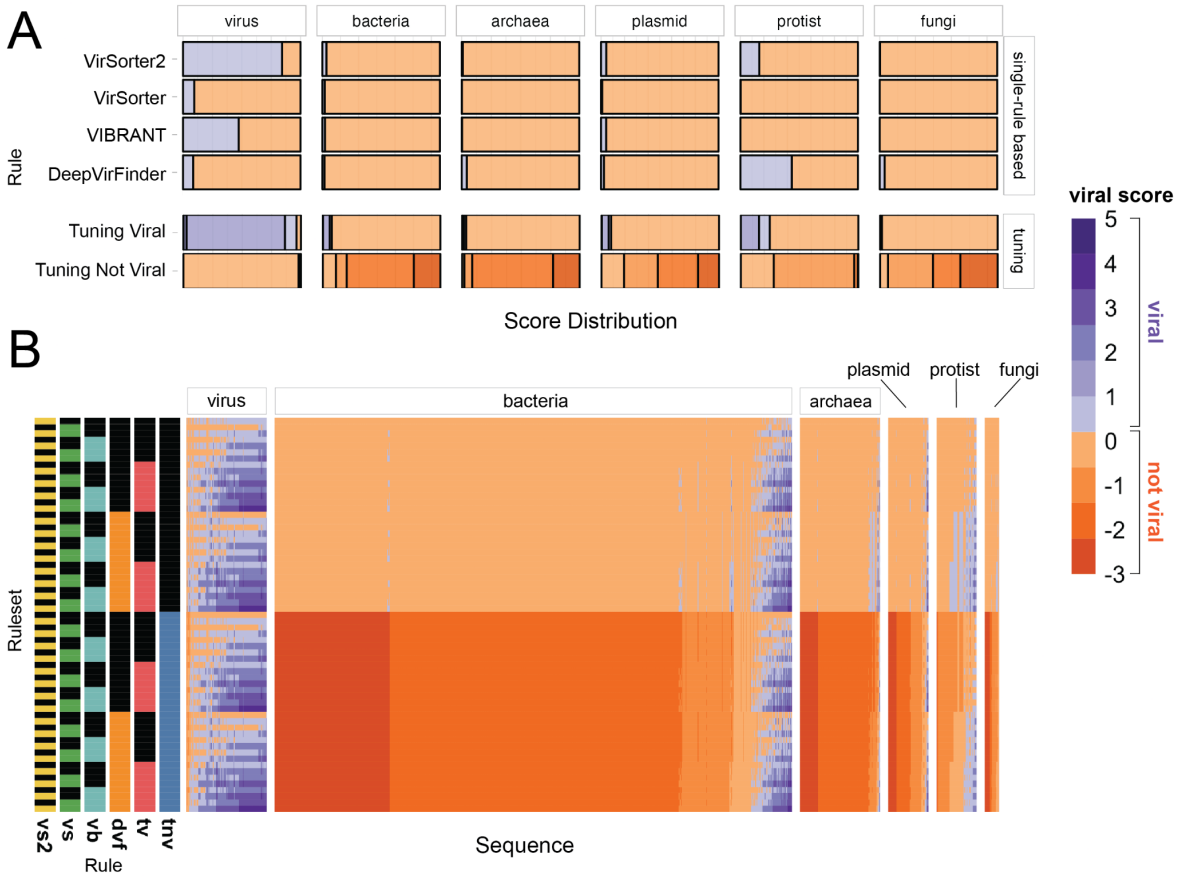

Figure 3. Comparison of different rulesets. *(A) Distribution of viral scores assigned to mock metagenome sequences for our 6 rules: 4 single-tool and 2 tuning rules. (B) Viral scores of all sequences across six mock metagenomes classified by each ruleset. Ruleset rows are colored based on whether or not each rule was used to attain the viral score result for a given sequence (yellow: VirSorter2 (vs2), green: VirSorter (vs), light blue: VIBRANT (vb), orange: DeepVirFinder (dvf), pink: tuning addition (tv), blue: tuning removal (tnv), black (rule was not used). Viral scores in 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). All sequences are grouped by their assigned taxonomy.*

More tools are better…to a point

Across the 63 rulesets, MCC, our metric for overall performance ('accuracy', herein), ranges from 0.05 (DeepVirFinder) to 0.77 (VirSorter2 + Tuning Removal). With the exception of VirSorter2 (MCC=0.75), single-rule rulesets (i.e., viral identification tools run 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). Previous studies have reported higher accuracy for these tools, but 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 the taxonomically diverse training data used here.[14,27,47,54] Using testing datasets representative of environmental metagenomic datasets is important. We found that our MCCs increased when using more typical 50% viral, 50% non-viral testing sets (max MCC=0.91); and, in their validation of DeepVirFinder, Ren et al.[27] demonstrated that the relative proportion of viral to non-viral sequences in a dataset can have a strong effect on AUROC (area under receiver operating characteristic; a performance metric). Given our results and the observed taxonomic distribution of environmental cellular metagenomic sequences,[15,23] users likely need to be more conservative (higher classification score cutoffs) in viral calling and assume lower accuracy than previous studies have reported.

Combining rules generally increased the average 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 (Fig. 4A), the precision of higher precision rules (precision > 0.7; the VIBRANT and VirSorter2 rules) decreases as they are combined with lower precision rules (precision < 0.5; the DeepVirFinder and VirSorter rules). Accuracy is maximized by the VirSorter2 and tuning removal ruleset, and is not improved by adding more rules (Figure 4B).

The VirSorter2-based and tuning removal rules are the most critical for accurate virus identification in our testing. When rulesets were ranked by increasing MCC, top performing "high MCC" rulesets were identified as those that did not demonstrate a statistically significant decrease relative to the highest MCC ruleset ($p_{adj} > 0.05$; Fig. 4B). VirSorter2's rule ('vs2') was in all six of these high MCC rulesets (Fig. 4B) and tuning removal ('tnv') was in five of them. For comparison, the other single-tool rules (i.e., VirSorter, DeepVirFinder, and VIBRANT) were each only in one of the high MCC rulesets, where they always co-occurred with VirSorter2 and the tuning removal rule (Fig. 4B). None of the high MCC rulesets have more than four rules. In the same way, "high precision" (three rulesets; Fig S2A) and "high recall" (four rulesets; Fig S2B) rulesets were defined.

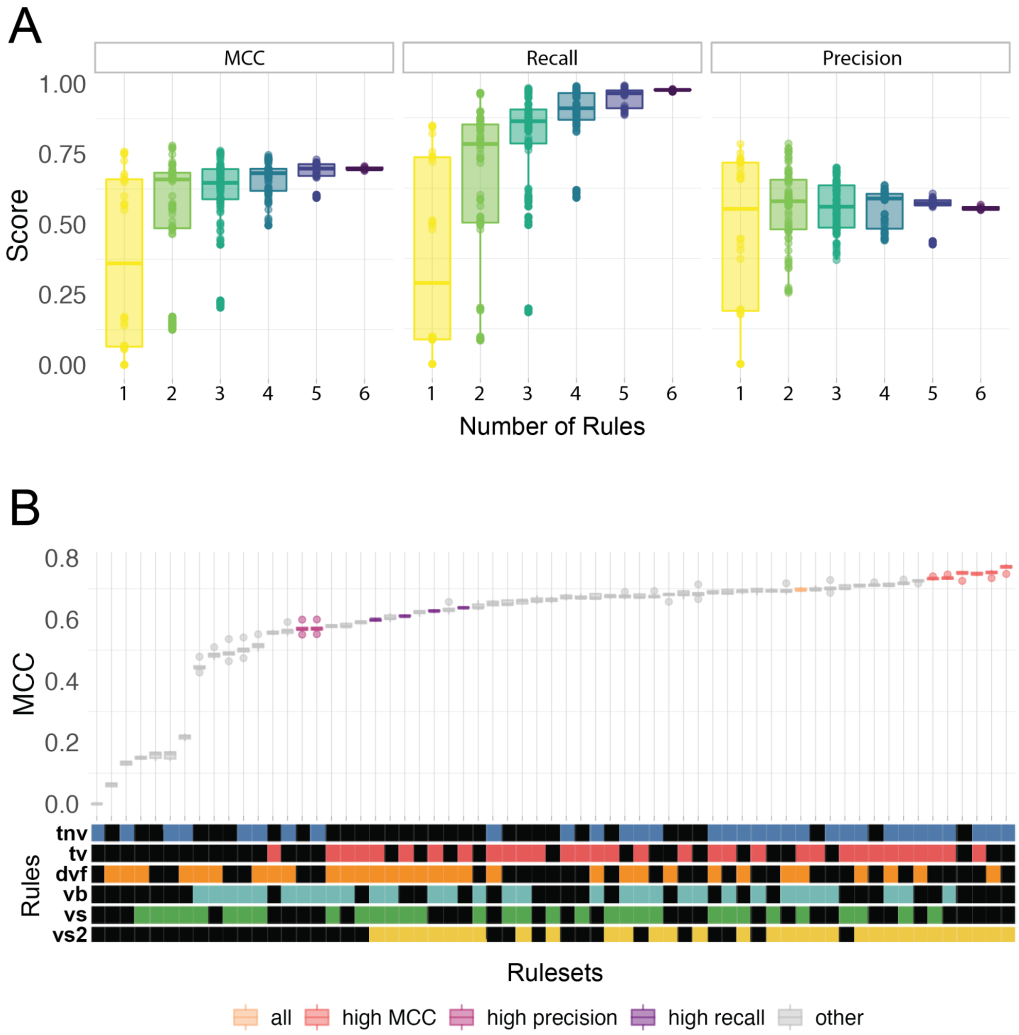

Figure 4. Performance of the 63 rulesets. *(A) - Box and whisker plots of the performance scores representing variation in MCC, precision, and recall of different rulesets based on the number of rules used for prediction. (B) Ruleset accuracy (MCC) ordered by increasing MCC and colored based on the ruleset's type according to statistically equivalent ($p_{adj} \geq 0.05$) rulesets: light orange: all rules ("all"), salmon: statistically significantly high MCC rulesets ("high MCC"), pink: statistically significantly high precision rulesets ("high precision", Figure S2A), purple: statistically high recall rulesets ("high recall", Figure S2B), and grey: all other rulesets ("other"). For A and B, the middle line represents the group mean; boxes above and below the middle line represent the top and bottom quartiles, respectively; whiskers above and below the boxes represent 1.5 times the interquartile range (roughly the 95% confidence interval), outliers are represented by circles beyond the whiskers. The boxplots are overlaid with points that represent the group data.*

There is a high degree of overlap in the viruses predicted by the different rulesets (Figure 5). 68% of the comparisons between rulesets (each ruleset compared against every other ruleset) were more than 50% identical to each other; and 30 rulesets were at least 90% identical to at least one other ruleset (representing 4% of the total comparisons between rulesets). 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). This suggests that our observed increase in recall through combining rules is being driven by a subset of rules that give more tool-specific viral predictions (i.e., VirSorter, DeepVirFinder, and VIBRANT), leading to the question: "how confident should users be of their tool-specific predictions?").

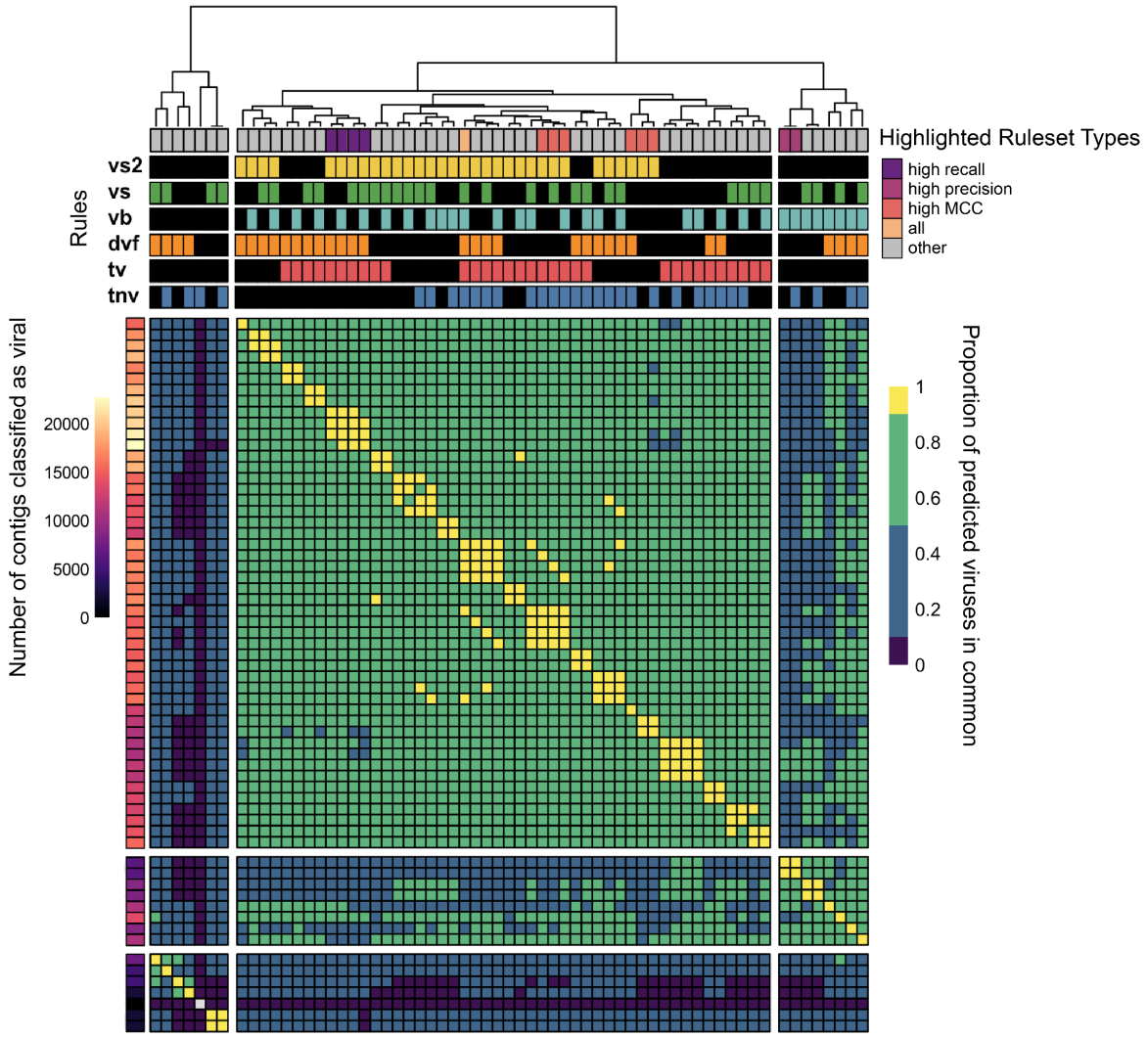

Figure 5 - Proportion of viruses in common between rulesets. *Heatmap values calculated by dividing the intersection (called viral by both rulesets) by the union (called viral by at least one) of the viruses found by both rulesets, which represents the proportion in common between the tools (scale bar on right: dark purple: 0-0.1, blue: 0.1-0.5, green: 0.5-0.9, yellow: 0.9-1). The bar to the left of the heatmap represents the total number of viruses identified by each tool combination (scalebar to its left). The bars above the heatmap indicate the tool(s) used in the rulesets: VirSorter2 (vs2, 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).*

One assumption made during *in silico* virus identification by previous studies is that sequences with low-confidence predictions by multiple tools are more likely to be viral viral (i.e., called viral by multiple +0.5 subrules). For example, if one tool predicts a sequence as viral with low confidence, it may be disregarded. But if multiple tools each provide low confidence predictions for a given sequence, many studies have presumed the sequence is more likely to be viral, thereby combining low quality predictions to arrive at the set of predicted viruses.[15,16,18] However, we found this was not a safe strategy. Rulesets using multiple low-quality prediction subrules in the single-tool rules did not significantly increase MCC (p-value=0.19) or recall (p-value=0.18); and in fact slightly decreased precision (0.54 vs 0.59, p-value=$2.5*10^{-5}$) when compared to rulesets that did not use low-quality prediction subrules (Figure S). This pattern is likely due to the rulesets being similarly uncertain about sequences that are unlikely to be viral; thus introducing a significant amount of non-viral contamination. The additive uncertainty did not create certainty. We recommend being cautious of sequences classified as viral by multiple low-quality predictions and should be manually inspected before calling them viral.

As metagenomes frequently have a high proportion of short sequence fragments and the correct identification of short fragments with only a few genes is particularly challenging,[26] we tested the rulesets on 3-5 kb fragments from our original testing sets. Like the >3 kb testing sets, many rulesets performed similarly for the 3-5 kb fragments (Figure S6). Unlike in the >3 kb testing sets where the VirSorter2 single-rule ruleset was in the high MCC set, no single-rule ruleset was in the high MCC rulesets for the short fragments (Fig. S6, 4B). Also in contrast with the >3 kb testing sets, the six-rule ruleset was identified as a high MCC ruleset (Fig. S6, 4B) and DeepVirFinder is in 8 of the 12 high MCC rulesets. In part due to our newly defined viral tuning rules, the accuracy of the viral predictions reported here for these short fragments is greater than previously published.[12] This increase in accuracy suggests that datasets where short fragments abound may particularly benefit from our tuning rules. Further, while researchers with only 3-5 kb fragments may consider using more tools, equally accurate predictions can still be achieved from the VirSorter2 and tuning removal ruleset.

Tuning rules increase confidence of viral predictions

To leverage expert knowledge of the differences between viral and non-viral sequences, we designed tuning addition and removal rules (Fig. 2E-F; see Figure S6 for subrule performance). These rules were designed based on specific outputs from multiple tools that were extremely precise in identifying viral and non-viral sequences. In general, tuning addition improved MCC and recall, while tuning removal improved MCC and precision (Figure S7). Reflecting the importance of the tuning rulesets on accuracy, 7 of the 10 highest MCC rulesets have both tuning addition and removal rules (Figure 4B). The tuning *removal* rule was able to identify 89% of the testing set's non-viral sequences and only (mis)identified 2% of testing set's viral sequences as non-viral. The tuning *addition* rule accurately identified 74% of the viral sequences and only misidentified 4% of non-viral sequences in the testing set as viral. Overall, the tuning rules increased our prediction accuracy beyond that of the rulesets composed of the single-tool rules. As such, we demonstrated the value of automating the refinement of viral identification tool predictions, a task that, if done at all, is currently a laborious manual process.

Even with the tuning rules, we could not improve both precision and recall beyond 0.77 (Figure S8). Building a more accurate classifier means overcoming barriers such as imperfect gene reference datasets,

overlap between viral and host sequences, and underrepresentation of viral types. This is because in order to recover more viruses, it becomes necessary to rely more on genes of unknown provenance. These may include non-viral genes, particularly of eukaryotes, which were not represented in the reference datasets of DeepVirFinder, VirSorter or CheckV (Figure S9-16). Further, many true viral features overlap with nonviral features (Figure S9-16) due to our imperfect knowledge of what distinguishes viruses and non-viruses (and homologous sequences shared by both viruses and cellular organisms), leading nonviruses with virus-like features to be misclassified as viruses. This challenge is particularly acute when trying to accurately classify both short sequences[12,27] and viral types underrepresented in our testing data (e.g., the accuracy of the "high MCC" ruleset is highest for dsDNAphages compared to other viral sequence types) (Figure S17).

Mislabeled sequences within databases hinder tool accuracy and validation

In an effort to improve upon the maximum MCC of 0.77, we looked for patterns in the types of sequences being misclassified that could aid future tool design. 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. Specifically, the proportion of the sequence's genes identified as viral orthologous genes (VOGs) was higher in the misclassified bacteria than the known viruses (p-value < $2.2*10^{-16}$; Figure 6A). The plasmids misclassified as viral have a similar proportion of viral genes as the viruses (p-value = 0.63). For all three sequence types, the proportion of sequences represented by VOGs increases with the number of VOGs in that sequence (Figure 6B). If these highly viral sequences are actually not viruses, viral identification tools can be improved by removing these genes from viral gene databases. On the contrary, if these sequences are actually viruses, viral identification tools can be improved by relabeling these sequences in sequence databases because tools rely on accurate database classification for training and testing.

Manual inspection of a subset of these misclassified sequences revealed that they are viral sequences (Figure 6C, S18). These false positives represented two types of mislabeled sequences: (1) viruses (either extracellular virions or intracellular extrachromosomal viral genomes) co-sequenced when a host isolate was sequenced (Figure 6C and S18A) and (2) prophages integrated into their hosts's genome (Figure S18B). We also screened for ΦX contamination, as it is a known problem for NCBI database sequences using Illumina library preparation due to its use in Illumina libraries.[56] Only 19 ΦX sequences were taxonomically identified by Kaiju; and, thus do not explain our high degree of false positives. These findings support the known problem of phage sequences not being removed before being deposited on NCBI,[56] and lead to phages being misclassified as bacteria, archaea, plasmids, and cellular and satellite chromosomes. Mislabeled sequences in public databases make it difficult to produce accurate viral identification tools because developers rely on accurately labeled datasets to train and test their classifiers. We recommend that before uploading sequences to public databases that non-viral sequences are screened for viral contamination.

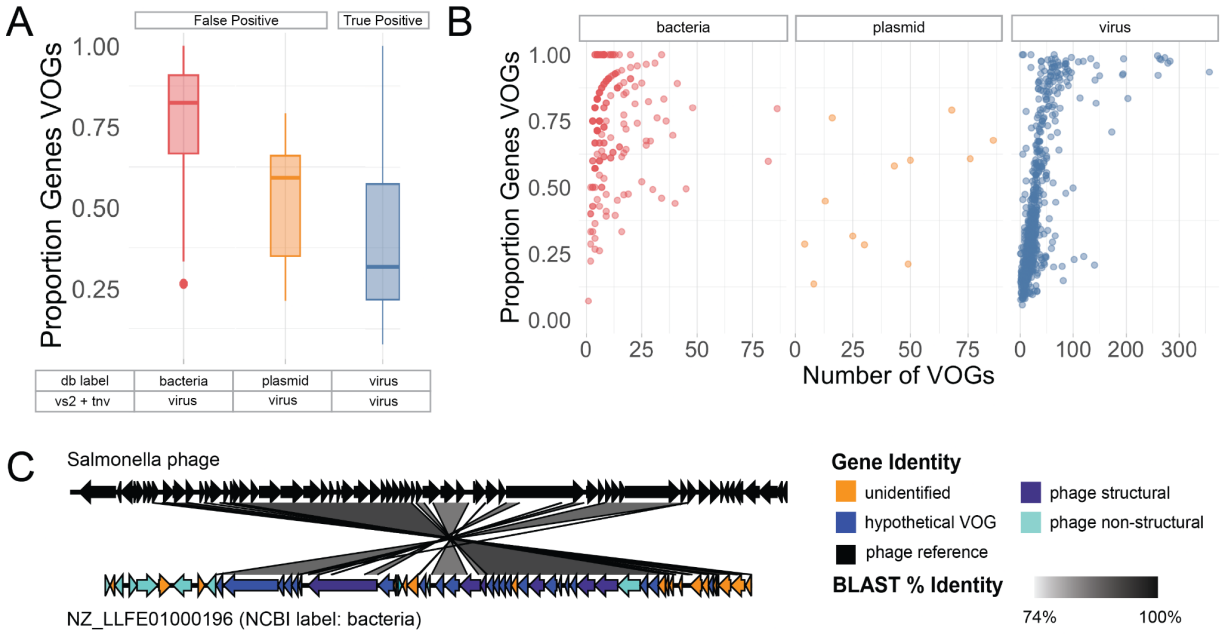

Figure 6 - Mislabeled sequences. *(A) Box and whisker plot of the proportion of genes on a sequence with a VOG annotation broken down by sequence type (for the high MCC rules). (B) Proportion of a sequence's genes that are called VOGs by VIBRANT versus the number of genes called VOGs faceted by sequence type. Because VIBRANT is the only tool that provides VOG annotation, only sequences classified as viral by VIBRANT are included in panels A and B (which only included bacteria, viruses, and plasmids). (C) Sequence synteny plots indicating sequence similarity between a representative bacterial "false positive": NZ_LLFE01000196 (NCBI label: bacteria) versus Salmonella phage SSU5. 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.*

Sample preparation and viral identification tool choice affects viral sequence recovery

To compare our rulesets across environments, we evaluated the proportion of sequences classified as viral for five publicly available aquatic metagenomic assemblies (Table 2, Figure S19). We found that viral sequence recovery varied greatly based on sample preparation (e.g., virus-enriched or not) and viral identification tool(s) used (Figure 7). The highest MCC ruleset identified a higher proportion of viral sequences in the virus-enriched samples (44-46%) compared to the non-enriched metagenomes (7-19%). The trends in proportion of viruses recovered across rulesets for the environmental datasets mimics the behavior of the testing data: the "high recall" rulesets classified the most sequences as viral (with presumably the greatest non-viral contamination, given the results on our testing sets), while the "high precision" rulesets labeled the fewest sequences as viral (with presumably the least non-viral contamination, given the results on the testing sets) (Figure 7, and Table S4). Previous metagenomic studies of biomass collected on a 0.22 µm filter have found viral sequences to be less than 16% of the total metagenome,[15,23] which is inline with our results. Further, these ranges in cellular-enriched samples

(>0.22 µm fraction) is consistent with the ~10% that one would expect, given that viral genomes tend to be 1-2 orders of magnitude smaller than their microbial hosts and are on average an order of magnitude more abundant.[15]

For the virus-enriched samples, nearly all sequences were called viral by the "high recall" rulesets (Figure 7, S18). It is likely that some of these sequences are false positives because the tuning removal rule reduces the number of sequences called viral by 14.9% (1 SD = 2.1%) when comparing the all vs highest recall rulesets; and our benchmarking demonstrated that the tuning removal rule effectively removed the contaminating non-viral sequences without removing true viruses. Further, one of the few studies to report the proportion of viral sequences found that 30-60% of the sequences with known taxonomy were similar to at least one viral sequence.[57] Even if all unknown sequences were viruses from this study, the proportion of viruses would not exceed 92%, which is still lower than the proportions we found in two of our three virus-enriched samples and further suggests the importance of tuning removal even for viral-enriched metagenomes. We present this "high recall" example to caution readers against using the "high recall" rulesets on virus-enriched metagenomes. It may be tempting to assume virus-enrichment removes nearly all non-viral sequence contamination, but even for virus-enriched fragments we instead recommend using the tuning removal rule unless the number of sequences is small enough to be manually inspected.

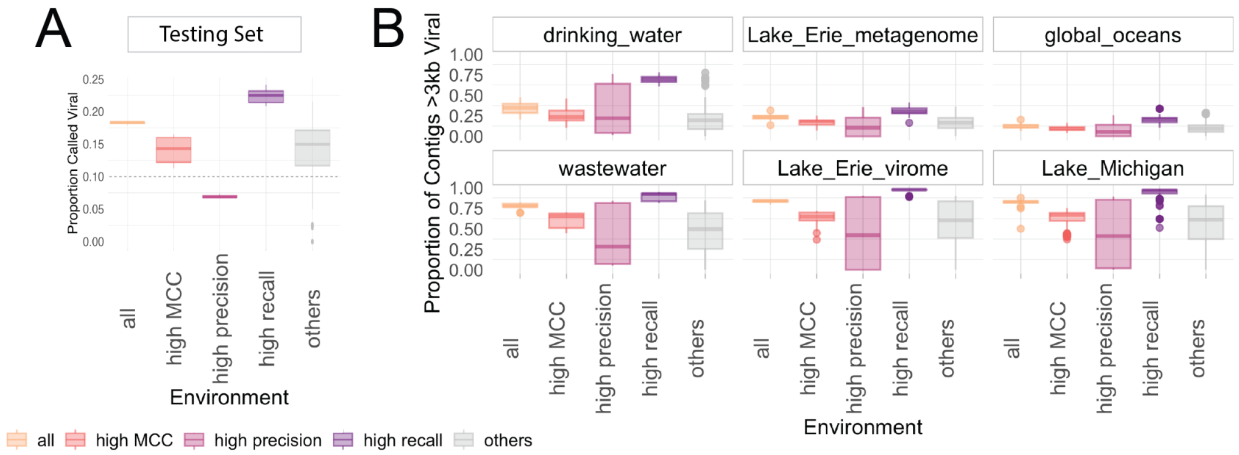

Figure 7 - Proportion of viruses predicted by each tool combination across five environmental datasets. *(A) environmental. (B) testing sets. Rulesets are grouped based on the accuracy type on the testing set shown in the highlighted rulesets from Figure 5.*

Recommendations and Future Work

In silico prediction of viral sequences is a critical first step to any metagenomic study that aims to resolve viral ecology and virus-microbe interactions. As downstream analyses and conclusions are predicated on accurate viral prediction, it is paramount to choose the most suitable tool(s) (and how to interpret its output(s)) among the rapidly increasing number of *in silico* viral prediction tools available. Through the above benchmarking, we demonstrated that specific two-rule rulesets can provide the highest precision with only minor sacrifices in recall, whereas the worse performance of combining all tools may lead to

erroneous biological conclusions. For this reason, we urge caution when using recent automated pipelines for sequence identification that combine the output of multiple tools.[17,19–21]

Our recommendations based on this study vary depending on research question and experimental design (Figure S20). 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 the majority of eukaryotes were filtered out of the sample (e.g., <3 µm fraction was sequenced), the tuning additional rule may increase recall. Based on the poor accuracy of the single-rule sets on the short sequences (3-5 kb testing set, Figure S5), we do not recommend researchers to use any of the tools in isolation. For researchers seeking to minimize the number of tools they use, we recommend using VirSorter2 or VIBRANT with our CheckV-based tuning subrules. VirSorter2 has a comparable MCC to multitool rules (Figure 4), though its high recall comes at the cost of more false positives (Figure S2). Although VIBRANT recovers fewer viruses, its predictions are very precise (Figure 4, Figure S2).

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, and some were trained using the VirSorter2 non-RefSeq genome set. These datasets also happen to harbor a large number of mislabeled sequences, making it difficult to assess the accuracy of the benchmarking results. As additional curated viral sets are published,[58] other researchers can test our rules against the new datasets providing further information about the limitations and scope of our rules.

We focused on tools that were developed primarily for bacteriophage identification, as these are most commonly used by microbial ecologists and microbiome researchers. We did not evaluate tools that were specifically for prophages, human viral pathogens, eukaryotic viruses, or archaeal viruses more broadly (Table S1). Future integration of new tools for plasmid and eukaryotic sequence identification[56,59] is likely to improve viral identification tool pipelines.

Conclusion

With the rapid development of new viral identification tools, this paper offers a blueprint for intentional, data-driven validation of tool combinations. We found that the highest accuracy resulted from rulesets with four or fewer rules. For most applications, we recommend a combination of VirSorter2 and tuning rules based on features of viral and not viral sequences; and caution against simply combining viral identification tools expecting higher quality virus sets. By increasing the proportion of high-confidence viruses identified from mixed metagenomic datasets through intentional, data-driven combination of tools, this study enables more accurate ecological analyses by decreasing contamination of the viral signal, particularly from eukaryotic sequences.

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) Zhou, Z.; Martin, C.; Kosmopoulos, J. C.; Anantharaman, K. ViWrap: A Modular Pipeline to Identify, Bin, Classify, and Predict Viral–Host Relationships for Viruses from Metagenomes. *iMeta n/a* (n/a), e118. https://doi.org/10.1002/imt2.118.
(20) Pandolfo, M.; Telatin, A.; Lazzari, G.; Adriaenssens, E. M.; Vitulo, N. MetaPhage: An Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data. *mSystems* **2022**, *7* (5), e00741-22. https://doi.org/10.1128/msystems.00741-22.
(21) Ru, J.; Khan Mirzaei, M.; Xue, J.; Peng, X.; Deng, L. ViroProfiler: A Containerized Bioinformatics Pipeline for Viral Metagenomic Data Analysis. *Gut Microbes* **2023**, *15* (1), 2192522. https://doi.org/10.1080/19490976.2023.2192522.
(22) Guo, J. VirSorter 2 Database, 2022. https://doi.org/10.5281/zenodo.4297575.
(23) 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.
(24) Al-Shayeb, B.; Sachdeva, R.; Chen, L.-X.; Ward, F.; Munk, P.; Devoto, A.; Castelle, C. J.; Olm, M. R.; Bouma-Gregson, K.; Amano, Y.; He, C.; Méheust, R.; Brooks, B.; Thomas, A.; Lavy, A.; Matheus-Carnevali, P.; Sun, C.; Goltsman, D. S. A.; Borton, M. A.; Sharrar, A.; Jaffe, A. L.; Nelson, T. C.; Kantor, R.; Keren, R.; Lane, K. R.; Farag, I. F.; Lei, S.; Finstad, K.; Amundson, R.; Anantharaman, K.; Zhou, J.; Probst, A. J.; Power, M. E.; Tringe, S. G.; Li, W.-J.; Wrighton, K.; Harrison, S.; Morowitz, M.; Relman, D. A.; Doudna, J. A.; Lehours, A.-C.; Warren, L.; Cate, J. H. D.; Santini, J. M.; Banfield, J. F. Clades of Huge Phages from across Earth's Ecosystems. *Nature* **2020**, *578* (7795), 425–431. https://doi.org/10.1038/s41586-020-2007-4.
(25) 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.
(26) 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.
(27) 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.
(28) 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.
(29) 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*.
(30) 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.
(31) *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).
(32) 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.
(33) 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.
(34) 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.
(35) 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.
(36) 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.
(37) 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.
(38) 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.
(39) 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.
(40) 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.
(41) 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.
(42) 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.
(43) 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.
(44) 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.
(45) 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.

(46) 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.

(47) 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.

(48) 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.

(49) 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.

(50) 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.

(51) 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.

(52) 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.

(53) Marquet, M.; Hölzer, M.; Pletz, M. W.; Viehweger, A.; Makarewicz, O.; Ehricht, R.; Brandt, C. What the Phage: A Scalable Workflow for the Identification and Analysis of Phage Sequences. *GigaScience* **2022**, *11*, giac110. https://doi.org/10.1093/gigascience/giac110.

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

(55) 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.

(56) 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.

(57) Roux, S.; Enault, F.; Robin, A.; Ravet, V.; Personnic, S.; Theil, S.; Colombet, J.; Sime-Ngando, T.; Debroas, D. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PloS One* **2012**, *7* (3), e33641. https://doi.org/10.1371/journal.pone.0033641.

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

(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.