Expanding standards in viromics: *in silico* evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation

Akbar Adjie Pratama¹, ², Ben Bolduc¹, ², Ahmed A. Zayed¹, ², Zhi-Ping Zhong¹, ², Jiarong Guo¹, ², Dean Vik¹, ², Maria Consuelo Gazitua³, James Wainaina¹, ², Simon Roux⁴ Corresp., Matthew B. Sullivan⁵ Corresp.

¹ Department of Microbiology, Ohio State University, Columbus, OH, United States
² Center of Microbiome Science, Ohio State University, Columbus, OH, United States
³ Viromica Consulting, Santiago, Chile
⁴ Lawrence Berkeley National Lab, Berkeley, CA, United States
⁵ Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, Ohio, United States

Corresponding Authors: Simon Roux, Matthew B. Sullivan
Email address: sroux@lbl.gov, sullivan.948@osu.edu

**Background:** Viruses influence global patterns of microbial diversity and nutrient cycles. Though viral metagenomics (viromics), specifically targeting dsDNA viruses, has been critical for revealing viral roles across diverse ecosystems, its analyses differ in many ways from those used for microbes. To date, viromics benchmarking has covered read pre-processing, assembly, relative abundance, read mapping thresholds and diversity estimation, but other steps would benefit from benchmarking and standardization. Here we use *in silico*-generated datasets and an extensive literature survey to evaluate and highlight how dataset composition (i.e. viromes vs bulk metagenomes) and assembly fragmentation impact (i) viral contig identification tool, (ii) virus taxonomic classification, and (iii) identification and curation of auxiliary metabolic genes (AMGs).

**Results:** The *in silico* benchmarking of five commonly used virus identification tools show that gene-content-based tools consistently performed well for long (≥3 kbp) contigs, while k-mer- and blast-based tools were uniquely able to detect viruses from short (<3 kbp) contigs. Notably, however, the performance increase of k-mer- and blast-based tools for short contigs was obtained at the cost of increased false positives (sometimes up to ~5% for virome and ~75% bulk samples), particularly when eukaryotic or mobile genetic element sequences were included in the test datasets. For viral classification, variously sized genome fragments were assessed using gene-sharing network analytics to quantify drop-offs in taxonomic assignments, which revealed correct assignations ranging from ~95% (whole genomes) down to ~80% (3 kbp sized genome fragments). A similar trend was also observed for other viral classification tools such as VPF-class, ViPTree and VIRIDIC, suggesting that caution is warranted when classifying short genome fragments and not full genomes. Finally, we highlight how fragmented assemblies can lead to erroneous identification of AMGs and outline a best-practices workflow to curate candidate AMGs in viral genomes assembled from metagenomes.

**Conclusion:** Together, these benchmarking experiments and annotation guidelines should aid researchers seeking to best detect, classify, and characterize the myriad viruses ‘hidden’ in diverse sequence datasets.

Keywords: Benchmarks, standard operating procedure, viruses, viromics, ecology
Expanding standards in viromics: *in silico* evaluation of dsDNA viral genome identification, classification, and auxiliary metabolic gene curation

Akbar Adjie Pratama\textsuperscript{1,2}, Ben Bolduc\textsuperscript{1,2}, Ahmed A. Zayed\textsuperscript{1,2}, Zhi-Ping Zhong\textsuperscript{1,2}, Jiarong Guo\textsuperscript{1,2}, Dean Vik\textsuperscript{1,2}, M. Consuelo Gazitúa\textsuperscript{3}, James Wainaina\textsuperscript{1,2}, Simon Roux\textsuperscript{4*} & Matthew B. Sullivan\textsuperscript{1,2,5*}

\textsuperscript{1} Department of Microbiology, Ohio State University, Columbus, OH, United States of America.
\textsuperscript{2} Center of Microbiome Science, Ohio State University, Columbus, OH, United States of America.
\textsuperscript{3} Viromica Consulting, Santiago, Chile.
\textsuperscript{4} Lawrence Berkeley National Laboratory, Berkeley, CA, United States of America.
\textsuperscript{5} Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH, United States.

*Correspondence:
Simon Roux,
JGI – Joint genome institute, Lawrence Berkeley National Laboratory
sroux@lbl.gov.

Matthew B. Sullivan,
496 W 12th Ave, Riffe Building 947, Columbus, OH 43210
sullivan.948@osu.edu.
Abstract

Background:
Viruses influence global patterns of microbial diversity and nutrient cycles. Though viral metagenomics (viromics), specifically targeting dsDNA viruses, has been critical for revealing viral roles across diverse ecosystems, its analyses differ in many ways from those used for microbes. To date, viromics benchmarking has covered read pre-processing, assembly, relative abundance, read mapping thresholds and diversity estimation, but other steps would benefit from benchmarking and standardization. Here we use in silico-generated datasets and an extensive literature survey to evaluate and highlight how dataset composition (i.e. viromes vs bulk metagenomes) and assembly fragmentation impact (i) viral contig identification tool, (ii) virus taxonomic classification, and (iii) identification and curation of auxiliary metabolic genes (AMGs).

Results:
The in silico benchmarking of five commonly used virus identification tools show that gene-content-based tools consistently performed well for long (≥3 kbp) contigs, while k-mer- and blast-based tools were uniquely able to detect viruses from short (<3 kbp) contigs. Notably, however, the performance increase of k-mer- and blast-based tools for short contigs was obtained at the cost of increased false positives (sometimes up to ~5% for virome and ~75% bulk samples), particularly when eukaryotic or mobile genetic element sequences were included in the test datasets. For viral classification, variously sized genome fragments were assessed using gene-sharing network analytics to quantify drop-offs in taxonomic assignments, which revealed correct assignations ranging from ~95% (whole genomes) down to ~80% (3 kbp sized genome fragments). A similar trend was also observed for other viral classification tools such as VPF-class, ViPTree and VIRIDIC, suggesting that caution is warranted when classifying short genome fragments and not full genomes. Finally, we highlight how fragmented assemblies can lead to erroneous identification of AMGs and outline a best-practices workflow to curate candidate AMGs in viral genomes assembled from metagenomes.

Conclusion:
Together, these benchmarking experiments and annotation guidelines should aid researchers seeking to best detect, classify, and characterize the myriad viruses ‘hidden’ in diverse sequence datasets.

Keywords: Benchmarks, standard operating procedure, viruses, viromics, ecology

Introduction
Viruses that infect microbes play significant roles across diverse ecosystems. For example, in marine systems, viruses are now broadly recognized to modulate biogeochemical cycles via lysis (e.g. heterotrophic prokaryotes lysis) (Fuhrman, 1999; Wilhelm & Suttle, 1999), alter evolutionary trajectory of core metabolisms via horizontal gene transfer (e.g. Sullivan et al., 2006), and impact the downward flux of carbon that helps the oceans buffer us (humans) against climate change (Guidi et al., 2016; Lara et al., 2017; Laber et al., 2018; Blanc-Mathieu et al., 2019).

Viromics (viral metagenomics) has helped further our understanding of marine viral genomic diversity, and ecosystem roles (Mizuno et al., 2013; Anantharaman et al., 2014; Coutinho et al., 2017; Nishimura et al., 2017a; Ahlgren et al., 2019; Haro-Moreno, Rodriguez-Valera & López-Pérez, 2019; Ignacio-espinoza, Ahlgren & Fuhrman, 2019; Luo et al., 2020). Ecologically, we now have global ocean catalogs approaching 200K dsDNA viruses that have been used to provide ecological maps of community structure and drivers (Mizuno et al., 2013; Brum et al., 2015; Roux et al., 2016; Coutinho et al., 2017; Gregory et al., 2019), and to formally (Gregory et al., 2019) and empirically (Gregory et al., 2019; Haro-Moreno, Rodriguez-Valera & López-Pérez, 2019) demonstrate that these viral populations represent species. Biogeochemically, viral roles in biogeochemistry now appear more nuanced as viruses impact biogeochemical cycling not only by lysing their microbial hosts as has been studied for decades (Fuhrman, 1999; Wilhelm & Suttle, 1999), but also by reprogramming cellular biogeochemical outputs either broadly through viral take-over and infection (the ‘virocell’) or more pointedly by expressing ‘auxiliary metabolic genes’ (AMGs) during infection that alter specific metabolisms of the cell (Breitbart et al., 2007; Lindell et al., 2007; Rosenwasser et al., 2016; Howard-Varona et al., 2020). While AMGs were initially discovered in cultures [e.g. photosynthesis genes (Mann et al., 2003)], viromics has drastically expanded upon these to now also include dozens of AMGs for...
functions across central carbon metabolism, sugar metabolism, lipid–fatty acid metabolism, signaling, motility, anti-oxidation, photosystem I, energy metabolism, iron–sulfur, sulfur, DNA replication initiation, DNA repair, and nitrogen cycling (Clokie et al., 2006; Sharon et al., 2007; Dinsdale et al., 2008; Millard et al., 2009; Wommack et al., 2015; Hurwitz, Brum & Sullivan, 2015; Roux et al., 2016; Breitbart et al., 2018; Ahlgren et al., 2019; Gazitúa et al., 2020; Kieft et al., 2020; Mara et al., 2020).

Beyond the oceans, viromics is also providing novel biological insights in e.g. humans (Lim et al., 2015; Norman et al., 2015; Reyes et al., 2015; Aiemjoy et al., 2019; Clooney et al., 2019; Fernandes et al., 2019; Gregory et al., 2020b), soils (Zablocki, Adriaenssens & Cowan, 2015; Trubl et al., 2018; Jin et al., 2019; Li et al., 2019; Santos-Medellín et al., 2020), and extreme environments (Adriaenssens et al., 2015; Scola et al., 2017; Bäckström et al., 2019; Zhong et al., 2020). Together these studies provide baseline ecological understanding of viral diversity and functions across diverse ecosystems.

Critically, however, viromics remains an emerging science frontier with methods and standards very much in flux. To date, standardization efforts have included (i) establishing quantitative data generation methods (Yilmaz, Allgaier & Hugenholtz, 2010; Duhaime et al., 2012; Hurwitz et al., 2013; Solonenko & Sullivan, 2013; Conceição-Neto et al., 2015; Roux et al., 2017), and (ii) analytical benchmarks for read pre-processing, metagenomics assembly, and thresholds for relative abundance, read mapping and diversity estimation (Brum et al., 2015; Gregory et al., 2016; Roux et al., 2017). Further, though not from viral particle derived metagenomes (viromes), related efforts have also been made to provide recommendations for how best to analyze viruses in bulk metagenomic samples (Paez-Espino et al., 2016, 2017; Dutilh et al., 2017; Emerson et al., 2018).

Here we contribute to this growing set of community-driven benchmarks and guidelines. Specifically, we use in silico datasets that mimic viromes (specifically of dsDNA viruses) and bulk metagenomes with varied inference from non-virus ‘distractor’ sequences to evaluate (i) options for viral identification, (ii) genomic fragment sizes for viral classification via gene-sharing networks, as well as (iii) provide guidelines for best practices for the evaluation of candidate AMGs.

**Materials and Methods**
Dataset

Datasets used in this study included genomes from: (i) NCBI virus RefSeq v.203 (released December 2020); to avoid including the same genomes used in any of the viral identification tools and vConTACT v2, we chose only complete genomes released after May 2020 (1,213 genomes, see Table S1), (ii) Bacteria RefSeq v.203 (174,973,817 genomic scaffolds), (iii) archaea RefSeq v.203 (2,116,989 genomic scaffolds), (iv) NCBI plasmids v.203 (1,339,171 genomes), and (iv) Human GRCh38 as the eukaryotic dataset. All datasets were downloaded from NCBI RefSeq, last accessed in December 2020 (the links are listed in the “availability of data and materials” section below). In addition, we also added ~142 dsDNA cyanophage genomes to include a set of closely related genomes, as can sometimes be obtained from viromics experiments (Table S1) (Gregory et al., 2016).

Dataset simulation

In silico simulations were adapted to benchmark the viromics pipelines for (i) virus identification and (ii) virus classification. The overall framework of dataset simulation strategies is shown in Figure 1. The simulation created four randomized subsampled datasets that were further fragmented to mimic fragmented assemblies of viromes and bulk metagenomes for viral contig identification and classification. An in-house script was used to (non-overlapping) fragment viral, eukaryotic, prokaryotic, and plasmid sequences into non-overlapping fragments of different lengths, i.e. $L = 500$ bp, 1 kbp, 3 kbp, 5 kb, 10 kbp, and 20 kbp. Non-overlapping fragments from each sequence category (viral, prokaryotic, eukaryotic, plasmid) were then combined to reflect mock communities’ composition (see below). These mixed datasets were used to benchmark viral contigs identification tools (Figure 1), while benchmarking of virus classification was performed only on fragmented sequences from viral RefSeq (Figure 1).

Mock communities

The four mock communities (with four replicates for each dataset) were randomly constructed to include different virus, prokaryotic, eukaryotic and plasmid sequences in ratios (Figure 1) that varied to represent communities enriched in viral genomes (Roux et al., 2015), i.e.
“virome_1 (up to 20,021 sequences; ratio, 10:1:1:0.01)” and “virome_2 (up to 20,021 sequences; ratio, 10:1:0.01:1)” or cellular genomes, i.e. “bulk_1 (up to 270,271 sequences; ratio, 1:10:0.01:10)” and “bulk_2 (up to 22,035 sequences; ratio, 1:10:1:1)” (Figure 1). To further investigate the potential source of errors in viral contigs identification, we also fragmented datasets consisting only of archaea, plasmid and eukaryotes (human; Figure 1).

Viral contig identification

The tools used for viral identification included VirSorter (Roux et al., 2015), MetaPhinder (Jurtz et al., 2016), MARVEL (Amgarten et al., 2018), DeepVirFinder (Ren et al., 2019), and VIBRANT (Kieft, Zhou & Anantharaman, 2020). Different cutoffs were applied for each of the tools, as follows, (i) we used two different versions of VirSorter, v1.0.5 and v1.10. VirSorter v1.05 used the viromedb database, while VirSorter v1.10 included the same viromedb database, as well as the Xfam database (Emerson et al., 2018; Gregory et al., 2019). For VirSorter (version 1.0.5 and 1.1.0; with "--db 2 --virome --diamond"), different cutoffs were used and compared: either all VirSorter predictions were considered as viruses (categories 1–6), or, only predictions of categories 1, 2, 4, and 5 was considered as viruses. For DeepVirFinder (version 1.0), we used three score cutoffs: ≥0.7, ≥0.9 and ≥0.95 and p-values ≤ 0.05. For MARVEL (version 0.2), two score cutoffs were used: ≥70% and ≥90%. Finally, for VIBRANT we used two different versions, i.e. version 1.1.0 and version 1.2.0; with “-virome” and no “-virome” setting), and MetaPhinder, default settings were used. The performance metrics to evaluate the efficiency of each tool were:

\[
MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN) \times (TN + FP)(TN + FN)}}
\]

(1)

Where MMC is Matthews’s correlation coefficient, TP is true positive, TN is true negative, FP is false positive, and FN is false negative. MCC values range between -1 to 1, with 1 indicating perfect efficiency (Chicco & Jurman, 2020).

\[
Recall = \frac{TP}{TP + FP}
\]

(2)

Where TP is true positive, and FP is false positive.

False - discovery rate = \[
\frac{FP}{FP + TN}
\]
Where FP is false positive, and TN is true negative.

\[
Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}
\]

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

\[
F1 = \frac{2TP}{2TP + FP + FN} \tag{4}
\]

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

\[
PVV = \frac{TP}{TP + FP} \tag{5}
\]

Where PVV is positive predictive value, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

\[
Specificity = \frac{TN}{TN + FP} \tag{6}
\]

Where TN is true negative, and FP is false positive.

**Statistical analysis**

A Wilcoxon test was used to compare the overall performance of viral identification, on the basis of fragment length (with 20 kbp as a reference group), including MCC, recall, false discovery, accuracy, F1, PVV, and specificity. The analysis was done using the R program (https://www.r-project.org/).

**Viral classification**

To evaluate the impact of fragmented assembly on a gene-sharing network-based viral classification, we leveraged vConTACT v2 (Jang et al., 2019) and used fragmented viral RefSeq genomes of different lengths (i.e. 500 bp, 1 kbp, 3 kbp, 5 kbp, 10 kbp, and 20 kbp) with default parameters. Furthermore, we also applied vConTACT v2 to complete genomes as a control dataset. It is worth noting that vConTACT v2 (originally) uses RefSeq v.85 (now has been updated to RefSeq v.99) as a reference database and manually validated ICTV taxonomies (ICTV Master Species List v1.3- February 2018) (Jang et al., 2019). The metrics used were those of Jang et al., (2019) including: (i) accuracy (Acc), (ii) clustering-wise separation (Sep), (iii) the
positive predictive value (PPV), (iv) clustering-wise sensitivity (Sn), (v) cluster-wise separation (Sep_c), and (vi) complex (ICTV taxonomy)-wise separation (Sep_co). The formulas are available in Jang et al., (2019).

In addition to vConTACT v2, we also evaluated the impact of fragmented assembly on viral classification using VPF-class (protein family based) (Pons et al., 2021), VipTree (genome-wide similarities-based) (Nishimura et al., 2017b), and VIRIDIC (BLASTN-based) (Moraru, Varsani & Kropinski, 2020). To evaluate the result, for VPF-class, we used taxonomic assignation of fragments with confidence score (CS) of ≥0.2 and membership ratio (MR) of ≥0.2, that have been reported to result in 100% of accuracy (Pons et al., 2021). For VIRIDIC and ViPTree, since no taxonomic assignation is automatically generated, we used the similarity and distance matrices provided by these tools to evaluate their performance on fragmented genomes, by comparing the similarity/distances obtained from genome fragments to the ones obtained from complete genomes (Nishimura et al., 2017b; Moraru, Varsani & Kropinski, 2020).

AMG curation analysis

Recommendations and best practices for AMG curation were based on a survey of the recent AMG literature, including especially Roux et al., (2016), Enault et al., (2017), Breitbart et al., (2018), Kieft et al., (2020). To illustrate the major challenges in the AMG identification process, we used DRAMv (Shaffer et al., 2020) to identify candidate AMGs in virus genomes from Emerson et al., 2018; Mara et al., 2020). The following parameters were used: AMGs score of 1-3 and AMG flag of -M and -F. To verify the functional annotation of the candidate AMGs, we manually checked the genomic context of the viral contigs, i.e. the annotation of the neighboring genes (especially the presence of viral hallmark and viral-like genes), and the position of AMG with respect to the contig’s edge. Next, we then manually looked for the presence of promoter/terminator regions using BPROM (Linear discriminant function (LDF) >2.75; (Richardson & Watson, 2013)), and ARNold (default setting; (Macke et al., 2001)). Conserved regions and active sites in the protein sequences were analyzed using PROSITE (Sigrist et al., 2013) and HHPred (Zimmermann et al., 2018) using the PROSITE collection of motifs (ftp://ftp.expasy.org/databases/prosite/prosite.dat), and PDB_mmCIF70_14_Oct (default) databases, respectively. For protein structural similarity, we used Phyre2 (confidence >90% and 70% coverage; (Kelly et al., 2015)), and predicted quaternary structures using SWISS-MODEL.
with a Global Model Quality Estimation (GMQE) score above 0.5 (Waterhouse et al., 2018).

Eventually, we selected one representative example for different typical cases of either genuine AMGs or false-positive detections, which are visualized using genome maps drawn with EasyFig (Sullivan, Petty & Beatson, 2011).

Results and Discussion

Establishment of mock communities for in silico testing

We first benchmarked and compared strategies for identification of viruses across different types of metagenomes. Researchers have identified viruses from virus-enriched metagenomes (viromes), as well as bulk and/or cellular metagenomes that are typically dominated by prokaryotic or eukaryotic sequences, all with variable representation of other mobile elements (e.g. plasmids and transposons). We thus established mock community datasets that included viral, prokaryotic, eukaryotic and plasmid sequences in varied ratios to represent a diversity of datasets likely to be encountered in environmental samples (Figure 1).

Briefly, two mock communities represented viromes and two represented bulk metagenomes, with ratios of virus: prokaryote: eukaryote: plasmid sequences as follows: “virome_1” ratio = 10:1:1:0.01, “virome_2” ratio = 10:1:0.01:1, “bulk_1” ratio = 1:10:0.01:10 and “bulk_2” ratio = 1:10:1:1 (see Methods and Materials for details, Figure 1). Clearly benchmarking are needed for other viral types since our focus here was dsDNA viruses. It is also worth noting that to better mimic viral populations in natural system, we complemented RefSeq genomes by specifically adding closely related genomes to the datasets from the only such deeply sequenced ‘reference’ dataset available (cyanophages (Gregory et al., 2016), see Materials and Methods). To reflect the fragmented assembly typically obtained from short-read metagenomes, we extracted random subsets of varying length (500 bp – 20 kbp) from these genomes, which were then combined at different ratios. Importantly, for viral RefSeq dataset, we only consider those newly submitted viral genome beyond May 2020, this to avoid including genomes that were used in training of any of the tools benchmarked here.
Comparison of viral identification tools

Several bioinformatic analysis tools have been developed to identify viruses from metagenomes (Table 1), using three major approaches: (i) similarity to known viruses, (ii) gene content/features, and (iii) k-mer frequency (i.e. nucleotide composition). Here, we first compared the performance of the most commonly used viral identification tools: VirSorter (Roux et al., 2015), MetaPhinder (Jurtz et al., 2016), MARVEL (Amgarten et al., 2018), DeepVirFinder (Ren et al., 2019), and VIBRANT (Kieft, Zhou & Anantharaman, 2020) against our suite of mock communities. We attempted to include two additional tools PHASTER (Arndt et al., 2017), and VirMiner (Zheng et al., 2019) -- but these did not scale and were eventually not included in the test results presented here. A range of parameters and cutoffs (see Methods and Materials for details) were used to assess the performance of each tool across different fragment sizes (ranging 500 bp – 20 kbp). Tool performance was evaluated using the following metrics: (i) “efficiency”, assessed using Matthews correlation coefficient, an overall statistic for assessing the recall and false-discovery, which this measure (MCC) offers a more informative and truthful evaluation than accuracy and F1 score (Chicco & Jurman, 2020), (ii) recall, (iii) false-discovery rate, (iv) accuracy, (v) F1, (vi) PVV, and (vi) specificity (see the formulas in Materials and Methods).

Overall, we found that viral contigs were better identified (increased efficiency, MCC) as fragment sizes increased, and this was true for all tools evaluated (Figure 2 and Figure S1-4, Wilcoxon test, p-value ≤0.0001). However, tools based on gene content, i.e. VIBRANT, MARVEL, and VirSorter (v1.05 and v1.10) decreased sharply in efficiency (MCC) with input sequences ≤3 kbp and particularly ≤1 kbp (Fig. 2A-C), whereas this decrease was less pronounced for DeepVirFinder (k-mer based) and MetaPhinder (BLASTN based) at these smaller size ranges (MCC values ~0.20-0.625; Figure 2A). Notably, the trade-off of this efficiency was a higher false-discovery that reached as much as ~5% for virome and ~80% for bulk samples in our mock communities as compared to <1% when longer fragments were used (Figure 2C).

We next explored how permissive versus conservative parameter cutoffs impacted viral identification based on permissive and conservative cutoffs recommended for each tool (Roux et al., 2015; Jurtz et al., 2016; Amgarten et al., 2018; Ren et al., 2019; Kieft, Zhou &
Anantharaman, 2020) (see Materials and Methods, and Figure 2). As expected, ‘conservative’
thresholds led to lower recall and lower false-discovery than ‘permissive’ for all tools (Figure 2).
This illustrates the trade-off that researchers are faced with maximizing viral identification
(eespecially for fragment sizes ≤ 3 kbp) using ‘permissive’ cutoffs and/or tools not based on gene
content will almost always be associated with a higher rate of false-discovery. Ultimately, the
initial research question of the study has to be considered to make the decision of which type of
cutoffs to use.

Finally, we evaluated whether false-positive detections were associated with specific types of
non-viral sequences, including other mobile genetic elements and “novel” microbial genomes.
To this end, we generated datasets composed only of archaea, plasmid, or eukaryotic sequences,
and measured false-discovery rates across the viral identification tools (Figure S3). It is
important to note however that, to our knowledge, there is currently no “clean” plasmid database
that is not also containing phages/viruses’ genome. Therefore, our benchmark is based on a
cleaning based on "complete" plasmid/phages, and primarily looking at how genome
fragmentation impacts the delineation of plasmid vs phage. Most tools showed an especially high
false-discovery rate for plasmid and/or eukaryotic sequences, including VIBRANT v.1.2.0 when
using the virome flag (highest in eukaryote up to >90% false-discovery, while other version of
VIBRANT is less affected), MetaPhinder (highest in plasmid up to >40% false-discovery),
MARVEL (up to ~20% false discovery for plasmid dataset), and VirSorter when using the
virome flag (up to ~24% false-discovery in eukaryote datasets) (Figure S3). This suggests the
data used to train these tools may have under-represented eukaryotic and/or plasmid sequence
and highlights the importance of including diverse non-viral sequences in a balanced training set
when establishing machine-learning based viral contig detection tools, as previously highlighted
(Ponsero & Hurwitz, 2019; Kieft, Zhou & Anantharaman, 2020). Overall, two tools stand out in
terms of maintaining the lowest false-discovery across the datasets: gene-content based VirSorter
(conservative cutoff) and MARVEL (score ≥90%).

Together these comparisons suggest that viral identification efficiency increases with
fragment length, and almost all tools are able to identify true viral contigs of 10 kbp or longer. At
length >3 kbp, “gene-content based tools” are able to maximize viral recall and minimize false
discovery at either permissive or conservative cutoffs, with VirSorter and MARVEL performing
best under conditions where “distractor genomes” (e.g. eukaryote, DPANN-archaea or plasmids)
are expected to be prevalent. For researchers specifically aiming to identify short (<3 kbp) viral
genome fragments, \(k\)-mer based tools (DeepVirFinder) and BLAST-based tool (MetaPhinder)
would be the preferred choices, although while being aware of the potential high rate of false-
positive detections, especially in samples where distractor genomes are expected to be prevalent.

**Virus classification using fragmented data in gene-sharing networks**

Once contigs from metagenomic assemblies are identified as viral, the next challenge a
researcher faces is to determine what kind of virus they represent. Gene-sharing network
analytics have emerged as a means to semi-automate such classification, and taxonomic
assignations for whole genomes are robust even when the network includes varying amounts of
fragmented genomes (Jang et al., 2019), but no studies have evaluated the taxonomic
assignations of fragmented genomes themselves. Because viral genomes assembled from
metagenomes are often partial, we sought to better understand how gene-sharing network
approaches would perform for metagenome-derived viral sequences at various fragment lengths.

To answer this question, we first established a dataset of known genomes and then
fragmented it to five fragment sizes that are commonly obtained from virome assemblies (Roux
et al., 2017). Next, we evaluated the accuracy of taxonomic assignments for the variously sized
genome fragments against those from complete genomes. Our results showed the percentage of
sequences accurately assignable to specific viral taxa increased with fragment length.
Specifically, the percentage of sequences clustered in a vConTACT v2 gene-sharing network
increased from 61% to > 80% from 3kb to fragment to complete genomes (Figure 3A). This
difficulty in robustly integrating short genome or genome fragments in a gene-sharing network is
further illustrated by the network topology itself, which shows a much higher fragmentation of
the network for 3kb fragment compared to complete genomes, accompanied by an inflated
number of “new VCs” and a higher number of unclustered sequence (whether outlier,
overlapping, or singleton, Figure S5). In addition to this lower rate of clustered sequences, short
fragments also displayed a reduced percentage of sequences assigned to the correct genus
(Figure S6) and overall lower performance across all vConTACT v2 metrics tested (Figure S7).
This is consistent with the original vConTACT v2 benchmark which also noted that accurate
classification was challenging to achieve for short complete genomes, i.e., genomes ≤ 5kb (Jang
Hence, short fragments (<10kb) may not be informative enough in terms of gene content to be robustly placed in a gene-sharing network and may artificially form ‘new’ virus clusters.

Currently, beyond vContact2, most viral classification tools such as VIRIDIC and VipTree have also been optimized to classify full viral genomes (Nishimura et al., 2017b; Moraru, Varsani & Kropinski, 2020). We thus sought to evaluate whether this decrease in performance with short fragments was a specificity of gene-sharing networks or was also observed for other taxonomic classification approaches. To test this, we performed similar comparisons of taxonomic assignment for varying genome fragment lengths using other viral classification tools including VipTree (genome-wide similarities-based), VIRIDIC (BLASTN-based), and VPF-class (protein family based). The general results show that the performance of these tools also increased with fragment size (Figure 3B, Figure S6, Figure S8). For VPF-class, the percentage of sequence with a taxonomic assignation increased from ~46% for 3 kbp fragments to ~82% for 20 kbp (Figure 3B), while the percentage of sequences assigned to the correct genus also increased with sequence length (Figure S8B). For ViPTree and VIRIDIC, an increase in performance was also observed from 3 kbp through 20 kbp (Figure S8). Together these results suggest genome fragmentation negatively impact virus taxonomic classification for all common approaches, with only longer genome fragments (≥10 kbp) providing sufficient information for an accurate and meaningful taxonomy assignment.

Auxiliary metabolic gene or not, that is the question

As sequencing technology and assembly algorithms improve, the increasing genomic context of uncultivated viruses provides an invaluable window into our ability to identify novel virus-encoded auxiliary metabolic genes, or AMGs. Problematically, however, until complete virus genomes are available, robustly identifying metabolically interesting genes in assembled (viruses) sequences from metagenomes remains a challenge for the field (e.g. see re-analyses of past ‘AMG’ studies in (Roux et al., 2013; Enault et al., 2017)). There are two major challenges in AMG analysis. First, even the most highly purified virus particle metagenome includes some degree of cellular genomic fragments (Roux et al., 2013; Zolfo et al., 2019). Thus, it is critical to demonstrate that any candidate AMG is indeed virus-encoded and not derived from cellular ‘contamination’, which requires adequate genomic context. Second, standard sequence analysis


cannot always determine whether a candidate AMG is involved in a metabolic pathway or
instead associated with ‘primary’ viral functions such as genome replication or host lysis. Based
upon previous (Clokie et al., 2006; Sharon et al., 2007; Dinsdale et al., 2008; Millard et al., 2009;
Wommack et al., 2015; Hurwitz, Brum & Sullivan, 2015; Roux et al., 2016; Breitbart et al.,
2018; Roitman et al., 2018; Ahlgren et al., 2019; Gazitúa et al., 2020; Kieft et al., 2020; Mara et
al., 2020), we propose guidelines to systematize the evaluation of candidate AMGs including: (i)

virus identification and quality assessment, (ii) AMG identification, genomic context assessment,
and functional annotation, and (iii) further investigation of putative AMGs function (Figure 4).

**Virus identification and quality assessment**

For AMG studies, we recommend using a combination of tools with strict quality thresholds
to identify high-confidence virus sequences (Figure 4, “Viral contigs identification”). For
example, high-confidence sequences might be those identified by Virsorter (cat 1,2) and
VirFinder/DeepVirFinder (score \( \geq 0.9 \), \( p \)-value \( <0.05 \)). For length of the contig, while we have in
the past used viral contigs \( \geq 1.5 \) kbp for AMG detection (Hurwitz, Hallam & Sullivan, 2013;
Hurwitz, Brum & Sullivan, 2015; Roux et al., 2016), improved sequencing and assembly
capabilities offer the opportunity to be less permissive since smaller contigs increase the risk of
false positives. Currently, we recommend increasing the minimum size threshold for AMGs
work to \( \geq 10 \) kbp, or those that are circular (and thus interpreted to be complete genomes).
Complementary to virus identification tools, we recommend using ViromeQC (Zolfo et al.,
2019) to evaluate virome contamination at the dataset level, and CheckV (Nayfach et al., 2020)
to identify and remove host contamination based on gene content for individual sequences.
Finally, for cases where integrated prophages are likely assembled in a contig including both a
host and a viral region, we recommend using prophage-specialized tools such as PHASTER
(Arndt et al., 2017) for more refined prophage/provirus identification and boundary demarcation.

*You are confident you have a virus sequence, but does it include any candidate AMGs?*

Next, candidate AMGs must be identified within the selected high-confidence viral contigs
(Figure 4 “Identification of candidate AMGs”). The key step in this process is to correctly
interpret results from a functional annotation pipeline to distinguish genes involved in host metabolism from genes involved in the viral replication cycle, often based on existing ontologies or manually defined keywords (Breitbart et al., 2018). To further refine this candidate AMG identification, it has been proposed that metabolic genes associated with a KEGG metabolic pathway would constitute “Class I” AMGs (i.e. highest confidence) while metabolic genes not directly included in a metabolic pathway (e.g. transport function) would represent “Class II” AMGs (lower confidence; (Hurwitz, Brum & Sullivan, 2015)). Importantly, depending on the definition one uses for ‘host metabolism’ vs ‘core viral functions’, some genes currently described in the literature as AMGs might not be legitimate AMGs, including some nucleotide-related genes (Kieft, Zhou & Anantharaman, 2020) or glycosyl transferases and glycoside hydrolases that are often used for surface attachment and entry (Shaffer et al., 2020). We thus recommend researchers to use the utmost caution when analyzing genes for which a true role and function remains uncertain and avoid systematically calling these simply “AMGs” without further qualifiers or caveats.

While prior AMG identification has often been done using manual inspection of the contigs functional annotation, there is opportunity now to advance towards a more systematic and semi-scalable approach to identify AMGs, with two new automated tools recently released. DRAM (Distilled and Refined Annotation of Metabolism), which is optimized for microbial annotation, but includes a ‘DRAM-v’ module for viruses, leverages expert-curated AMG databases for functional annotation and a two-component scoring system to assess the likelihood of a gene being encoded on a virus genome (Shaffer et al., 2020). Meanwhile VIBRANT, which is built for virus identification but also performs functional annotation, automatically curate KEGG-based annotations to highlight candidate AMGs associated to KEGG ‘metabolic pathways’ and ‘sulfur relay system’ categories (Kieft, Zhou & Anantharaman, 2020). Both tools thus provide a quick and automated way to obtain a list of candidate AMGs which nevertheless must be further analyzed to (i) verify that the candidate AMG is indeed encoded by a virus, and (ii) verify that the candidate AMG is indeed involved in a cellular metabolic pathway.

How do you recognize a candidate AMG that may not actually be virus-encoded?
Although automated annotation tools such as DRAM-v and VIBRANT are helpful in speeding up the identification of candidate AMG, any detailed ecological or evolutionary analysis of an AMG requires additional manual inspection of both genomic context and predicted functions. Here, we illustrate examples of typical “mistakes” made by automated tools (Figure 4 “Genomic context assessment of candidate AMGs”).

First, two examples of sequences likely to be genuinely viral, either closely related to a known phage (contig_1, ‘likely viral’) or not (contig_2, ‘possible viral’) are presented in Figure 4. These sequences are mostly composed of viral or unknown genes, with little to no “cellular-like” gene outside of the single candidate AMG. Next to these however, are four examples of AMGs predicted yet unlikely to be viral (‘unlikely viral’ candidates). Contig_3 represents a sequence ~120 kbp with dense, short genes, and no viral/viral-like genes. This sequence is likely to be a cellular genomic region, possibly a mobile genetic element, that could easily be mistaken for a phage by automated tools. Next in contig_4, the candidate AMG is surrounded by genes that reveal little evidence of belonging to a viral genome, but where VirSorter (categories 1 and 2) and/or VirFinder (score ≥0.9 and p-value <0.05) suggest the contig overall is of viral origin. Conservatively, these genes AMGs should not be considered further due to the low contextual evidence of the region immediately surrounding the candidate AMG to be of viral origin. Finally, in contig_5, the candidate AMG is at the edge of the viral contig along with a tRNA and a phage integrase. This example likely represents the miscall of a prophage boundary, and the AMG-containing region is likely a small fraction of the host genome, where metabolic genes are much more common (Table S2). Overall, further examining the specific genomic context around each candidate AMG is highly recommended in order to identify false-positive detections, i.e. non-viral sequences wrongly considered as viral by automated tools. This is especially critical in AMG analysis because these non-viral regions, while overall rare among the entire set of sequences predicted as viral, will typically have a much higher probability of including genes annotated as metabolic, i.e. candidate “AMGs”. Hence, even a small number of contaminating sequences can substantially impact downstream AMG analyses.

How to recognize a true metabolic AMG?
As for their viral origin (see above), the predicted function of candidate AMGs will typically need to be refined beyond the results of automated functional annotation pipelines. While the ideal proof of function is through biochemical assay of the AMGs to support the *in silico* inferred function, this is laborious and time-consuming lab work, such that only a handful of AMGs known to date has been experimentally validated -- *psbA* (Lindell et al., 2005; Clokie et al., 2006), *pebS* (Dammeyer et al., 2008), and glycoside hydrolase (Emerson et al., 2018). To provide scalable *in silico* evaluation of putative AMGs and guide future experimental validation, we recommend the following analyses (Figure 4 “AMG functional analysis”).

First, deeper functional analyses should be conducted to assess, where possible, whether the AMG contains known conserved residues and active sites, as well as whether structural predictions are consistent with the sequence-based functional prediction (Figure 4). The analysis of protein conserved regions and active sites can be done manually via inspection of sequence alignments, as well as semi-automatedly where possible using, e.g. PROSITE (Sigrist et al., 2013) and HHPred (Zimmermann et al., 2018). For protein structural predictions there are several available tools including Phyre² (Kelly et al., 2015), SWISS-MODEL (Waterhouse et al., 2018), and I-TASSER (Yang & Zhang, 2015). Protein structure is known to be more conserved than primary protein sequence, thus enabling the annotation of more divergent proteins, as well as supporting other functional annotation pipelines (Kelly et al., 2015). Importantly, when interpreting results of predicted structures and structure-based similarity for candidate AMGs, one should verify that the predicted structure is consistent with the predicted biological function, but also consider the relationship between top hits, in which one would expect to have several of the top hits homologous to each other (Roux et al., 2016; Gazitúa et al., 2020). The latest recommended cut-offs for these functional annotation tools are provided in Table 2.

Evolutionary analyses can be used to assess whether selection appears to be acting on the viral gene homolog. For instance, the ratio of non-synonymous (*pN*) to synonymous (*pS*) polymorphisms (*pN/pS*) -- known as *pN/pS* -- can be used to evaluate whether the candidate AMGs is under purifying selection as would be expected for a functional gene (Schloissnig et al., 2013; Roux et al., 2016). Pragmatically, *pN/pS* values can be calculated manually using tools designed
specifically for analyzing micro- and macro-diversity in metagenomes (e.g. MetaPop (Gregory et al., 2020a)).

Your AMG appears viral and predicted to be functional and involved in host cell metabolism, what is its ecological and evolutionary story to tell?

Until this point, the candidate AMGs have gone through a series of meticulous vetting steps resulting in putative AMGs that can be used for downstream analyses such as phylogeny, ecological analysis, and experimental functional assays. We provide recommendations for each as follows (Figure 4 “Additional (optional) analysis of the putative AMGs”).

To assess the evolutionary history of AMGs, phylogenetic analysis is carried out on individual AMGs and their corresponding microbial homologs. Briefly, for each AMG, one first needs to obtain homologs via sequence similarity searches (e.g. BLAST vs an appropriate database), then do multiple sequence alignments (e.g. MAFFT (Katoh et al., 2002)), assess for intragenic recombination (e.g. RDP4 software (Martin et al., 2015)), build phylogenetic trees (e.g. IQ-TREE (Nguyen et al., 2015)), and visualize them (e.g. iTOL, (Letunic & Bork, 2019)). With these data in-hand, each phylogenetic tree can be examined to determine the number of transfer events that have occurred between microbes and viruses, as well as the ‘origin’ of the AMGs within the cellular and viral sequences in the analyses (sensu Sullivan et al., (2010)).

Bona fide AMGs also typically have an ecological story to tell. Currently, the abundance of AMGs is estimated by read mapping against the viral populations that contain those AMGs (Gazitúa et al., 2020). However, a more sophisticated approach, where possible, would be to use the evolutionary inferences and multiple sequence alignments to identify virus-specific ‘signatures’ in the sequences that could be read-mapped to differentiate viral from cellular contributions to the gene, transcript, or protein pool in any given natural community. While such analyses are quite rare, e.g. (Sharon et al., 2007; Tzahor et al., 2009) growing AMG datasets should empower researchers to address this question of the virus ‘AMG’ contributions. Further, as virus-host prediction capabilities improve (Edwards et al., 2016; Villarroel et al., 2016; Galiez et al., 2017; Emerson et al., 2018; Wang et al., 2020), there is opportunity to combine these with AMG predictions to build understanding of ecologically-critical nuances of virus-host interactions.
interactions. Finally, viral AMGs are under very different selective pressures than their host homologues given their viro-centric roles during infection. Will functional validation reveal viral versions that are fundamentally different? On one side, we may expect viral AMGs to have subtle mutations that might impact their enzyme efficiency (e.g. mutations in the PEST domain of PsbA (Sharon et al., 2007)) or substrate preferences (e.g. (Enav et al., 2018)). On the other side, we may expect viruses to encode more efficient proteins with “new” functions. An example here is cyanophage-encoded ‘PebA’, which was thought to be a divergent 15,16-dihydrobiliverdin: ferredoxin oxidoreductase (pebA), but experimentally was shown to combine the capabilities of two host enzymes, PebA and PebB, to directly convert biliverdin IXα to phycoerythrobilin and was thus renamed to PebS, a phycoerythrobilin synthase (Dammeyer et al., 2008).

Together, we hope these guidelines provide best-practice standard operating procedures for scientists to identify and evaluate candidate AMGs, as well as an emerging roadmap for how best to robustly bring this more nuanced and under-studied component of virus-host interactions to light so that viruses can be better incorporated into ecosystem models.

Conclusions

While viromics has proven invaluable for revealing the roles of viruses across diverse ecosystems, the emergent field of viral ecogenomics is in a state of rapid flux, experimentally and analytically. Here, we add to recent best practices efforts by evaluating and providing benchmarking for identifying and classifying viruses from viral-particle-enriched and bulk metagenomes, as well as recommendations for best practices for studying virus-encoded auxiliary metabolic genes. These efforts addressed some critical issues in standard operating procedures for viral ecogenomics researchers. Similar efforts will be needed to establish best practices in studying new emerging types of analysis and data including micro-diversity of virus populations (Gregory et al., 2019), and long-read sequencing (Warwick-Dugdale et al., 2018; Zablocki et al., 2021). Further, technological and analytical opportunities are being developed to better capture ssDNA and RNA viruses, as well as to establish dsDNA viral activity.
(Moniruzzaman et al., 2017; Emerson et al., 2018; Roux et al., 2019; Sommers et al., 2019; Starr et al., 2019; Trubl et al., 2019; Callanan et al., 2020). Finally, though viral discovery is now performed tens to hundreds of thousands of viruses at a time, the ability to link these new viruses to their hosts is still limited. Improved in silico approaches, such as those based on BLAST similarity, $k$-mers (such as WIsH (Galiez et al., 2017), and HostPhinder (Villarroel et al., 2016)), oligonucleotide frequency (VirHostMatcher (Wang et al., 2020)), and CRISPR-Cas (Paez-Espino et al., 2016) have been recently proposed to predict the potential hosts of uncultivated viruses, which still need to be thoroughly tested and benchmarked across a variety of dataset types and sizes. Moreover, predictions from these in silico prediction tools need to be complemented with robustly benchmarked, high-throughput experimental methods, e.g. epicPCR, viral tagging, Hi-C (Deng et al., 2014; Bickhart et al., 2019; Yaffe & Relman, 2020; Sakowski et al., 2021) to validate these predictions.

Acknowledgements:

We would like to thank Drs. Heather Maughan and Chistine Sun for comments on the structure of an early draft of the manuscript, Drs. Ho Bin Jang and Olivier Zablocki, as well as Mohamed M. Mohamed, and Funing Tian for the many constructive discussions.

Abbreviations:

MCC: Matthews’s correlation coefficient,
Sn: clustering-wise sensitivity,
PPV: the positive predictive value,
Acc: accuracy,
$\text{Sep}_{co}$: complex (ICTV taxonomy)-wise separation,$\text{Sep}_{cl}$: cluster-wise separation, and
Sep: clustering-wise separation.

Funding:
Funding was provided by NSF (#OCE1829831, #ABI1758974), the U.S. Department of Energy (#DE-SC0020173 and #248445), and the Gordon and Betty Moore Foundation (#3790) to MBS. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231. An award from the Ohio Supercomputer Center (OSC) to MBS supported computing resources used here.

Grant Disclosures:

- NSF (#OCE1829831, #ABI1758974).
- The U.S. Department of Energy #DE-SC0020173 and #248445, and # DE-AC02-05CH11231.
- The Gordon and Betty Moore Foundation (#3790).
- The Ohio Supercomputer Center (OSC).

Ethics approval and consent to participate

Not applicable

Conflict of interest:

Maria Consuelo Gazitua is a founder and CEO of Viromica Consulting, Chile. All the authors declare that they have no competing interests.

Authors’ Contributions:

- Akbar Adjie Pratama conceived and designed the experiments, prepare the simulated datasets, performed the experiments, analyzed the data, prepared the manuscript, prepared figures and/or tables.
- Ben Bolduc, Ahmed A. Zayed, conceived the study, and provided the scripts, analyzed the data.
- Zhi-Ping Zhong, Jiarong Guo, performed the viral identification analysis.
- Dean Vik, M. Consuelo Gazitúa, James Wainaina, prepared specific sections of the manuscript.
Simon Roux & Matthew B. Sullivan conceived and designed the experiments, prepared the manuscript.

All authors helped write and edit the manuscript and approved the final draft.

Availability of data and materials:

All sequencing data used to simulate datasets is publicly available and was obtained from NCBI: (i) Viral RefSeq v.203 for viral sequences (https://www.ncbi.nlm.nih.gov/nuccore/viruses/), (ii) Bacteria RefSeq v.203 (https://www.ncbi.nlm.nih.gov/nuccore/), (iii) archaea RefSeq v.203 (https://www.ncbi.nlm.nih.gov/nuccore/), (iv) plasmids RefSeq v.203 (https://www.ncbi.nlm.nih.gov/nuccore/) and (v) human GRCh38 for eukaryote (accession number GCA_000001405.15). The 142 cyanophages’ genomes are available in the GenBank repository under the accession numbers KJ019026- KJ019131, KJ019134- KJ019165, JN371768, and KF156338-40.

The scripts used for virus identification, classification, mock communities’ datasets, virus datasets, vConTACT v2 input and AMG input files have been uploaded to https://bitbucket.org/MAVERICLab/standard_viroomics2/

Supplementary information:

Supplementary table 1. Virus genomes used in this study; Supplementary table 2. AMGs annotation table; Supplementary figure 1. Additional metrics for viral identification across datasets; Supplementary figure 2. Results from the viral identification analysis for VIBRANT with no virome flag; Supplementary figure 3. False-discovery rate of viral sequence identification tools when challenged with Archaea, Eukaryote and Plasmid datasets; Supplementary figure 4. Pairwise analyses of the combined viral identification tools efficiency parameters for varying genome sizes; Supplementary figure 5. Gene-sharing network topology and performances; Supplementary figure 6. Percentage of ‘correct’ genus-assignments; Supplementary figure 7. Comparison of the performances of viral gene-sharing network-based classification, vConTACT v2 across fragment size; Supplementary figure 8. Performance evaluation of VipTree and VIRIDIC across fragment sizes.
References:

Adriaenssens EM, Van Zyl L, De Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan DA. 2015. Metagenomic analysis of the viral community in Namib Desert hypoliths. *Environmental Microbiology* 17:480–495. DOI: 10.1111/1462-2920.12528.

Ahlgren NA, Fuchsman CA, Rocap G, Fuhrman JA. 2019. Discovery of several novel, widespread, and ecologically distinct marine Thaumarcheota viruses that encode amoC nitrification genes. *ISME Journal* 13:618–631. DOI: 10.1038/s41396-018-0289-4.

Aiemjoy K, Altan E, Aragie S, Fry DM, Phan TG, Deng X, Chanyalew M, Tadesse Z, Callahan EK, Delwart E, Keenan JD. 2019. Viral species richness and composition in young children with loose or watery stool in Ethiopia. *BMC Infectious Diseases* 19:1–10. DOI: 10.1186/s12879-019-3674-3.

Amgarten D, Braga LPP, da Silva AM, Setubal JC. 2018. MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. *Frontiers in Genetics* 9:1–8. DOI: 10.3389/fgene.2018.00304.

Anantharaman K, Duhaime MB, Breier JA, Wendt K, Brandy M, Toner, Dick GJ. 2014. Abundance of viruses in deep oceanic waters. *Science* 344:757–760. DOI: 10.1126/science.1252229.

Arndt D, Marcu A, Liang Y, Wishart DS. 2017. PHAST, PHASTER and PHASTEST: Tools for finding prophage in bacterial genomes. *Briefings in Bioinformatics* 1:1–8. DOI: 10.1093/bib/bbx121.

Bäckström D, Yutin N, Jørgensen SL, Dharamshi J, Homa F, Zaremba-Niedwiedzka K, Spang A, Wolf YI, Koonin E V., Ettema TJG. 2019. Virus Genomes from Deep Sea Sediments Expand the Ocean Megavirome and Support Independent Origins of Viral Gigantism. *mBio* 10:e02497-18. DOI: 10.1128/MBio.02497-18.

Bickhart DM, Watson M, Koren S, Panke-Buisse K, Cersosimo LM, Press MO, Van Tassell CP, Van Kessel JAS, Haley BJ, Kim SW, Heiner C, Suen G, Bakshy K, Liachko I, Sullivan ST, Myer PR, Ghurye J, Pop M, Weimer PJ, Phillippy AM, Smith TPL. 2019. Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biology* 20:1–18. DOI: 10.1186/s13059-019-1760-x.

Blanc-Mathieu R, Kaneko H, Endo H, Chaffron S, Hernández-Veláquez R, Nguyen CH, Mamitsuka H, Henry N, Vargas C de, Sullivan MB, Suttle CA, Guidi L, Ogata H. 2019. Viruses of the eukaryotic plankton are predicted to increase carbon export efficiency in the global sunlit ocean. *bioRxiv*. DOI: 10.1101/710228.

Breitbart M, Bonnain C, Malki K, Sawaya NA. 2018. Phage puppet masters of the marine microbial realm. *Nature Microbiology* 3:754–766. DOI: 10.1038/s41564-018-0166-y.

Breitbart M, Thompson L, Suttle C, Sullivan M. 2007. Exploring the Vast Diversity of Marine Viruses. *Oceanography* 20:135–139. DOI: 10.5670/oceanog.2007.58.

Brum JR, Ignacio-espinoza JC, Roux S, Doucier G, Acinas SG, Alberti A, Chaffron S. 2015. Patterns and ecological drivers of ocean viral communities. *Science* 348:1261498-1–11. DOI: 10.1126/science.1261498.

Callanan J, Stockdale SR, Shkoporov A, Draper LA, Ross RP, Hill C. 2020. Expansion of known ssRNA phage genomes: From tens to over a thousand. *Science Advances* 6. DOI: 10.1126/sciadv.aay5981.

Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:1–13. DOI:
Clokie MRJ, Shan J, Bailey S, Jia Y, Krisch HM, West S, Mann NH. 2006. Transcription of a “photosynthetic” T4-type phage during infection of a marine cyanobacterium. *Environmental Microbiology* 8:827–835. DOI: 10.1111/j.1462-2920.2005.00969.x.

Clooney AG, Sutton TDS, Shkoporov AN, Plevy SE, Ross RP, Hill C, Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, Regan OO. 2019. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. *Cell Host and Microbe* 15. DOI: 10.1016/j.chom.2019.10.009.

Conceição-Neto N, Zeller M, Lefrère H, De Bruyn P, Beller L, Deboutte W, Yinda CK, Lavigne R, Maes P, Ranst M Van, Heylen E, Matthijssens J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: A reproducible protocol for virome analysis. *Scientific Reports* 5:1–14. DOI: 10.1038/srep16532.

Coutinho FH, Silveira CB, Gregoracci GB, Thompson CC, Edwards RA, Brussaard CPD, Dutilh BE, Thompson FL. 2017. Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nature Communications* 8:15955. DOI: 10.1038/ncomms15955.

Dammeyer T, Bagby SC, Sullivan MB, Chisholm SW, Frankenlag-Dinkel N. 2008. Efficient Phage-Mediated Pigment Biosynthesis in Oceanic Cyanobacteria. *Current Biology* 18:442–448. DOI: 10.1016/j.cub.2008.02.067.

Deng L, Ignacio-Espinoza JC, Gregory AC, Poulos BT, Weitz JS, Hugenholtz P, Sullivan MB. 2014. Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* 513:242–245. DOI: 10.1038/nature13459.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Furlan M, Desnues C, Haynes M, Li L, McDaniell L, Moran MA, Nelson KE, Nilsson C, Olson R, Paul J, Brito BR, Ruan Y, Swan BK, Stevens R, Valentine DL, Thurber RV, Wegley L, White BA, Rohwer F. 2008. Functional metagenomic profiling of nine biomes. *Nature* 452:629–632. DOI: 10.1038/nature06810.

Duhaime MB, Deng L, Poulos BT, Sullivan MB. 2012. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: A rigorous assessment and optimization of the linker amplification method. *Environmental Microbiology* 14:2526–2537. DOI: 10.1111/j.1462-2920.2012.02791.x.

Dutilh BE, Reyes A, Hall RJ, Whiteson KL. 2017. *Virus Discovery by Metagenomics: The (Im)possibilities*. DOI: 10.3389/978-2-88945-308-5.

Edwards RA, McNair K, Faust K, Raes J, Dutilh BE. 2016. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiology Reviews* 40:258–272. DOI: 10.1093/femsre/fuv048.

Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft B., Jang HB, Singleton CM, Salden LM, Naas AE, Boyd JA, Hodgkins SB, Wilson RM, Trubl G, Li C, Froliking S, Pope PB, Wrighton KC, Crill PM, Chant J, Sullivan MB. 2018. Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* 3:870–880. DOI: 10.1038/s41564-018-0190-y.

Enault F, Brier A, Bouteille L, Roux S, Sullivan MB, Petit MA. 2017. Phages rarely encode antibiotic resistance genes: A cautionary tale for virome analyses. *ISME Journal* 11:237–247. DOI: 10.1038/ismej.2016.90.

Enav H, Kirzner S, Lindell D, Mandel-Gutfreund Y, Beja O. 2018. Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. *bioRxiv*:261479. DOI: 10.1101/261479.
Fernandes MA, Verstraete SG, Phan T, Deng X, Stekol E, Lamere B, Lynch S V., Heyman MB, Delwart E. 2019. Enteric Virome and Bacterial Microbiota in Children with Ulcerative Colitis and Crohn Disease. *Journal of Pediatric Gastroenterology and Nutrition* 68:30–36. DOI: 10.1097/MPG.0000000000002140.

Fuhrman JA. 1999. Marine viruses and their biogeochemical and ecological effects. *Nature* 399:541–548. DOI: 10.1038/21119.

Galiez C, Siebert M, Enault F, Vincent J, Söding J. 2017. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* 33:3113–3114. DOI: 10.1093/bioinformatics/btx383.

Gazitúa MC, Vik DR, Roux S, Gregory AC, Bolduc B, Widner B, Mulholland MR, Hallam SJ, Ulloa O, Sullivan MB, Sullivan MB. 2020. Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. *The ISME Journal*. DOI: 10.1038/s41396-020-00825-6.

Gregory AC, Gerhardt K, Zhong Z-P, Bolduc B, Temperton B, Konstantinidis KT, Sullivan MB. 2020a. MetaPop: A pipeline for macro- and micro-diversity analyses and visualization of microbial and viral metagenome-derived populations. *bioRxiv*. DOI: 10.1101/2020.11.01.363960.

Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, Sudek S, Maitland A, Chittick L, Santos F dos, Weitz JS, Worden AZ, Woyke T, Sullivan MB. 2016. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* 17:930. DOI: 10.1186/s12864-016-3286-x.

Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB, Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B. 2020b. The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut Resource The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host and Microbe*:1–17. DOI: 10.1016/j.chom.2020.08.003.

Gregory AC, Zayed AA, Sunagawa S, Wincker P, Sullivan MB, Ferland J, Kandels L, Liu Y, Marec C, Vik D. 2019. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *cell*:1–15. DOI: 10.1016/j.cell.2019.03.040.

Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, Darzi Y, Audic S, Berline L, Brum JR, Coelho LP, Espinoza JC1, 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 SG, Bork P, De Vargas C, Iudicone D, Sullivan MB, Raes J, Karsenti E, Bowler C, Gorsky G. 2016. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 532:465–470. DOI: 10.1038/nature16942.

Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. 2019. Prokaryotic Population Dynamics and Viral Predation in a Marine Succession Experiment Using Metagenomics. *Frontiers in Microbiology* 10:1–14. DOI: 10.3389/fmicb.2019.02926.

Howard-Varona C, Lindback MM, Bastien GE, Solonenko N, Zayed AA, Jang H, Andreopoulos B, Brewer HM, Rio TG del, Adkins JN, Paul S, Sullivan MB, Duhaime MB, Received: 2020. Phage-specific metabolic reprogramming of virocells. *Isme J*. DOI: 10.1038/s41396-019-0580-z.

Hurwitz BL, Brum JR, Sullivan MB. 2015. Depth-stratified functional and taxonomic niche specialization in the ‘core’ and ‘flexible’ Pacific Ocean Virome. *The ISME Journal* 9:472–484. DOI: 10.1038/ismej.2014.143.

Hurwitz BL, Deng L, Poulos BT, Sullivan MB. 2013. Evaluation of methods to concentrate and
puriﬁy ocean virus communities through comparative, replicated metagenomics.

*Environmental Microbiology* 15:1428–1440. DOI: 10.1111/j.1462-2920.2012.02836.x.

Hurwitz BL, Hallam SJ, Sullivan MB. 2013. Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biology* 14. DOI: 10.1186/gb-2013-14-11-r123.

Ignacio-espinoza JC, Ahlgren NA, Fuhrman JA. 2019. Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nature Microbiology*. DOI: 10.1038/s41564-019-0628-x.

Jang H Bin, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* 37:632. DOI: 10.1038/s41587-019-0100-8.

Jin M, Guo X, Zhang R, Qu W, Gao B, Zeng R. 2019. Diversities and potential biogeochemical impacts of mangrove soil viruses. *Microbiome* :1–15.

Jurtz VI, Villarroel J, Lund O, Voldby Larsen M, Nielsen M. 2016. MetaPhinder - Identifying bacteriophage sequences in metagenomic data sets. *PLoS ONE* 11:1–14. DOI: 10.1371/journal.pone.0163111.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* 30:3059–3066. DOI: 10.1093/nar/gkf436.

Kelly LA, Mezulis S, Yates C, Wass M, Sternberg M. 2015. The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols* 10:845–858. DOI: 10.1038/nprot.2015-053.

Kieft K, Zhou Z, Anantharaman K. 2020. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8:90. DOI: 10.1186/s40168-020-00867-0.

Kieft K, Zhou Z, Anderson RE, Buchan A, Campbell BJ, Hallam SJ, Hess M, Sullivan MB, Walsh DA, Roux S, Anantharaman K. 2020. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *bioRxiv*. DOI: 10.1017/CBO9781107415324.004.

Laber CP, Hunter JE, Carvalho F, Collins JR, Hunter EJ, Schieler BM, Boss E, More K, Frada M, Thamtrakoln K, Brown CM, Haramaty L, Ossolinski J, Fredricks H, Nissimov JI, Vandzura R, Sheyn U, Lehahn Y, Chant RJ, Martins AM, Coolen MJL, Vardi A, Ditullio GR, Van Mooy BAS, Bidle KD. 2018. Coccolithovirus facilitation of carbon export in the North Atlantic. *Nature Microbiology* 3:537–547. DOI: 10.1038/s41564-018-0128-4.

Lara E, Vaqué D, Sà EL, Boras JA, Gomes A, Borrull E, Diez-Vives C, Teira E, Pernice MC, Garcia FC, Forn I, Castillo YM, Peiró A, Salazar G, Morán XAG, Massana R, Catalá TS, Luna GM, Agustí S, Estrada M, Gasol JM, Duarte CM. 2017. Unveiling the role and life strategies of viruses from the surface to the dark ocean. *Science Advances* 3. DOI: 10.1126/sciadv.1602565.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research*:1–4. DOI: 10.1093/nar/gkz239.

Li Y, Liu H, Pan H, Zhu X, Liu C, Zhang Q, Luo Y, Di H, Xu J. 2019. T4-type viruses: Important impacts on shaping bacterial community along a chronosequence of 2000-year old paddy soils. *Soil Biology and Biochemistry* 128:89–99. DOI: 10.1016/j.soilbio.2018.10.007.

Lim ES, Zhou Y, Zhao G, Bauer IK, Droit L, Ndao IM, Warner BB, Tarr PI, Wang D, Holtz LR. 2015. Early life dynamics of the human gut virome and bacterial microbiome in infants.
Lindell D, Jaffe JD, Coleman ML, Futschik ME, Axmann IM, Rector T, Kettler G, Sullivan MB, Steen R, Hess WR, Church GM, Chisholm SW. 2007. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature* 449:83–86. DOI: 10.1038/nature06130.

Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438:86–89. DOI: 10.1038/nature04111.

Luo E, Eppley JM, Romano AE, Mende DR, DeLong EF. 2020. Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. *The ISME Journal*. DOI: 10.1038/s41396-020-0604-8.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research* 29:4724–4735. DOI: 10.1093/nar/29.22.4724.

Mann NH, Cook A, Millard A, Bailey S, Clokie M. 2003. Bacterial photosynthesis genes in a virus. *Nature* 424:741. DOI: 10.1038/424741a.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution* 1:1–5. DOI: 10.1093/ve/vev003.

Millard AD, Zwiglmaier K, Downey MJ, Mann NH, Scanlan DJ. 2009. Comparative genomics of marine cyanomyoviruses reveals the widespread occurrence of Synechococcus host genes localized to a hyperplastic region: Implications for mechanisms of cyanophage evolution. *Environmental Microbiology* 11:2370–2387. DOI: 10.1111/j.1462-2920.2009.01966.x.

Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. 2013. Expanding the Marine Virosphere Using Metagenomics. *PLoS Genetics* 9. DOI: 10.1371/journal.pgen.1003987.

Moniruzzaman M, Wurch LL, Alexander H, Dyhrman ST, Sullivan MB, Edgcomb VP. 2020. Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *The ISME Journal*. DOI: 10.1038/s41396-020-00739-3.

Martín DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution* 1:1–5. DOI: 10.1093/ve/vev003.

Moraru C, Varsani A, Kropinski AM. 2020. VIRIDIC — A Novel Tool to Calculate the Intergenomic Similarities of. *Viruses* 12:1268. DOI: 10.3390/v12111268.

Nayfach S, Camargo AP, Eloevadrosh E, Roux S, Kyrpides N, Berkeley L. 2020. CheckV: assessing the quality of metagenome-assembled viral genomes. *bioRxiv*.

Nghien LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32:268–274. DOI: 10.1093/molbev/msu300.

Nishimura Y, Watai H, Honda T, Mihara T, Omae K, Roux S, Blanc-Mathieu R, Yamamoto K, Hingamp P, Sako Y, Sullivan MB, Goto S, Ogata H, Yoshida T. 2017a. Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *American society for microbiology* 2.

Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. 2017b. ViPTree: The viral proteomic tree server. *Bioinformatics* 33:2379–2380. DOI: 10.1093/bioinformatics/btx157.

Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleschner P, Stappenbeck TS, McGovern DPB, Keshavarzian A, Mutlu EA, Sauk J,
Cite references here.
Sakowski EG, Arora-williams K, Tian F, Zayed AA, Zablocki O, Sullivan MB, Preheim SP. 2021. Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR. *Nature Microbiology*. DOI: 10.1038/s41564-021-00873-4.

Santos-Medellin C, Zinke LA, ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. 2020. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. DOI: 10.1017/CBO9781107415324.004.

Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P. 2013. Genomic variation landscape of the human gut microbiome. *Nature* 493:45–50. DOI:

Scola V, Ramond J-B, Frossard A, Zablocki O, Adriaenssens EM, Johnson RM, Seely M, Cowan DA. 2017. Namib desert soil microbial community diversity, assembly, and function along a natural xeric gradient. *Microbial Ecology* 1:193–203. DOI: 10.1007/s00248-017-1009-8.

Shaffer M, Borton MA, Mcgivern BB, Zayed AA, Leanti S, Rosa L, Solden LM, Liu P, Narrowe AB, Daly RA, Bolduc B, Gazit MC, Smith GJ, Vik DR, Pope PB, Sullivan MB, Roux S, Wrighton KC. 2020. DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research*:1–18. DOI: 10.1093/nar/gkaa621.

Sharon I, Tzahor S, Williamson S, Shmoish M, Man-Aharonovich D, Rusch DB, Yooseph S, Zeidner G, Golden SS, MacKey SR, Adir N, Weingart U, Horn D, Venter JC, Mandel-Guttfreund Y, Béja O. 2007. Viral photosynthetic reaction center genes and transcripts in the marine environment. *ISME Journal* 1:492–501. DOI: 10.1038/ismej.2007.67.

Sigrist CJA, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2013. New and continuing developments at PROSITE. *Nucleic Acids Research* 41:344–347. DOI: 10.1093/nar/gks1067.

Solonenko SA, Sullivan MB. 2013. *Preparation of metagenomic libraries from naturally occurring marine viruses*. Elsevier Inc. DOI: 10.1016/B978-0-12-407863-5.00008-3.

Sommers P, Fontenele RS, Kringen T, Kraberger S, Porazinska DL, Darcy JL, Schmidt SK, Varsani A. 2019. Single-Stranded DNA Viruses in Antarctic Cryoconite Holes. *Viruses*.

Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK. 2019. Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proceedings of the National Academy of Sciences*. DOI: 10.1101/597468.

Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, DeFrancesco AS, Kern SE, Thompson LR, Young S, Yandava C, Fu R, Krastins B, Chase M, Sarracino D, Osburne MS, Henn MR, Chisholm SW. 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environmental Microbiology* 12:3035–3056. DOI: 10.1111/j.1462-2920.2010.02280.x.

Sullivan MB, Lindell D, Lee JA, Thompson LR, Bielawski JP, Chisholm SW. 2006. Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biology* 4:1344–1357. DOI: 10.1371/journal.pbio.0040234.

Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: A genome comparison visualizer. *Bioinformatics* 27:1009–1010. DOI: 10.1093/bioinformatics/btr039.

Trubl G, Jang H Bin, Roux S, Emerson JB, Solonenko N, Vik DR, Solden L, Ellenbogen J, Runyon AT, Bolduc B, Woodcroft BJ, Saleska SR, Tyson GW, Wrighton KC, Sullivan MB, Rich VI. 2018. Soil viruses are underexplored players in ecosystem carbon processing.
American society for microbiology 3:e00076-18. DOI: 10.1101/338103.

Trubl G, Roux S, Solonenko N, Li Y-F, Bolduc B, Eloe-Fadrosh E, Rich V, Sullivan M. 2019. Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. PeerJ Preprints 7:e7265. DOI: 10.7287/peerj.preprints.27640.

Tzahor S, Man-Aharonovich D, Kirkup BC, Yogev T, Berman-Frank I, Polz MF, Béjà O, Mandel-Gutfreund Y. 2009. A supervised learning approach for taxonomic classification of core-photosystem-II genes and transcripts in the marine environment. BMC Genomics 10. DOI: 10.1186/1471-2164-10-229.

Villarroel J, Kleinheinz KA, Jurtz VI, Zschach H, Lund O, Nielsen M, Larsen MV. 2016. HostPhinder: A Phage Host Prediction Tool. Viruses:1–22. DOI: 10.3390/v8050116.

Wang W, Ren J, Tang K, Dart E, Ignacio-Espinoza JC, Fuhrman JA, Braun J, Sun F, Ahlgren NA. 2020. A network-based integrated framework for predicting virus–prokaryote interactions. NAR Genomics and Bioinformatics 2:1–19. DOI: 10.1093/nargab/lqaa044.

Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB, Temperton B. 2018. Long-read viral metagenomics enables capture of abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ 345041. DOI: 10.1101/345041.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, De Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T. 2018. SWISS-MODEL: Homology modelling of protein structures and complexes. Nucleic Acids Research 46:W296–W303. DOI: 10.1093/nar/gky427.

Wilhelm SW, Suttle CA. 1999. Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs. BioScience 49:781–788. DOI: 10.1161/CIRCULATIONAHA.111.030536.

Wommack KE, Nasko DJ, Chopyk J, Sakowski EG. 2015. Counts and sequences, observations that continue to change our understanding of viruses in nature. Journal of Microbiology 53:181–192. DOI: 10.1007/s12275-015-5068-6.

Yaffe E, Relman DA. 2020. Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation. Nature Microbiology 5:343–353. DOI: 10.1038/s41564-019-0625-0.

Yang J, Zhang Y. 2015. I-TASSER server: New development for protein structure and function predictions. Nucleic Acids Research 43:W174–W181. DOI: 10.1093/nar/gkv342.

Yilmaz S, Allgaier M, Hugenholtz P. 2010. Multiple displacement amplification compromises quantitative analysis of metagenomes. Nature Methods 7:943–944. DOI: 10.1038/nmeth1210-943.

Zablocki O, Adriaenssens EM, Cowan D. 2015. Diversity and ecology of viruses in hyperarid desert soils. Applied and Environmental Microbiology 82:770–777. DOI: 10.1128/AEM.02651-15.

Zablocki O, Michelsen M, Burris M, Solonenko N, Warwick-Dugdale J, Ghosh R, Pett-Ridge J, Sullivan MB, Temperton B. 2021. VirIION2: a short- And long-read sequencing and informatics workflow to study the genomic diversity of viruses in nature. PeerJ. DOI: 10.1101/2020.10.28.359364.

Zheng T, Li J, Ni Y, Kang K, Misiakou M-A, Imamovic L, Chow BKC, Rode AA, Bytzer P, Sommer M, Panagiotou G. 2019. Mining, analyzing, and integrating viral signals from metagenomic data. Microbiome 7:42. DOI: 10.1186/s40168-019-0657-y.
Zhong Z, Rapp JZ, Wainaina JM, Solonenko NE, Maughan H, Carpenter SD, Cooper ZS, Jang B, Bolduc B, Deming JW, Sullivan B. 2020. Viral Ecogenomics of Arctic Cryopeg Brine and Sea Ice. *mBio* 5:1–17.

Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology* 430:2237–2243. DOI: 10.1016/j.jmb.2017.12.007.

Zolfo M, Pinto F, Asnicar F, Manghi P, Tett A, Bushman FD, Segata N. 2019. Detecting contamination in viromes using ViromeQC. *Nature Biotechnology*. DOI: 10.1038/s41587-019-0334-5.
Figure 1. The framework of dataset simulation strategies. First, the viral RefSeq, prokaryote, eukaryote, and plasmid genome sequences were fragmented, from 5' to 3' end direction, into non-overlapping fragments of different lengths, i.e. $L = 500$ bp, 1 kbp, 3 kbp, 5 kbp, 10 kbp, and 20 kbp fragments. Then, these non-overlapping fragments were randomly sub-sampled to obtain simulated input datasets. For virus identification analysis, these simulated datasets were designed to resemble mock communities with different ratios of viral, prokaryote, eukaryote and plasmid sequences, i.e. virome_1 (10:1:0.1:0.01), virome_2 (10:1:0.01:1), bulk_1 (1:10:0.1:10), and bulk_2 (1:10:1:1). For viral classification analysis, simulated inputs were exclusively composed of fragmented viral genomes.

Figure 2. The viral identification analysis across datasets. (A-D) Pie-charts of the composition of the datasets depicted the different fragment sizes of the (i) virome_1, (ii) virome_2, (iii) bulk_1, and (iv) bulk_2. (E-H) The viral identification efficiency was calculated as Matthew’s correlation coefficient (MCC), where 1 represents perfect efficiency, (I-L) percent of recall (%), and (M-P) percent of false-discovery (%) of DeepVirFinder, MetaPhinder, MARVEL, VIBRANT, and VirSorter. For DeepVirFinder, three cutoffs were evaluated, i.e. score $\geq 0.7$, $\geq 0.9$, $\geq 0.95$, and p-value $\leq 0.05$. For MARVEL, two cutoffs were used, i.e. scores of $\geq 70\%$ and $\geq 90\%$. Next, we use two different versions of VirSorter, i.e. v1.05 and v1.10, and two cutoffs, i.e. category 1,2,3,4,5,6 and category 1,2,4,5. The upper error bars represent the mean of the replicates.

Figure 3. Viral classification analysis. (A) Percentage of the input sequences in a vConTACT v2 cluster and (B) Percentage of sequences assigned by VFP-class to a genus. The performance of VPF-class was calculated using confidence score (CS) and membership ratio (MR) thresholds of $\geq 0.2$ (Pons et al., 2021).

Figure 4. Proposed workflow and curation step for AMG identification and validation. The recommended steps of a candidate AMGs include, (i) viral contig identification and quality assessment, (ii) AMG identification, genomic context assessment, and functional annotation, and (iii) further investigation of putative AMGs function.
Supplementary figure 1. Additional metrics for viral identification analysis across datasets. (A) The viral identification accuracy, (B) F1, (C) PVV, and (D) specificity of DeepVirFinder, MetaPhinder, MARVEL, VIBRANT, and VirSorter on the different fragment sizes of the (i) virome_1, (ii) virome_2, (iii) bulk_1, and (iv) bulk_2 (the composition of the datasets depicted as pie-charts). For DeepVirFinder, three cutoffs were evaluated, i.e. score $\geq 0.7$, $\geq 0.9$, $\geq 0.95$, and p-value $\leq 0.05$. For MARVEL, two cutoffs were used, i.e. scores of $\geq 70\%$ and $\geq 90\%$. Next, we use two different versions of VirSorter, i.e. v1.05 and v1.10, and two cutoffs, i.e. category 1,2,3,4,5,6 and category 1,2,4,5. The upper error bars represent the mean of the replicates (see more in Materials and Methods).

Supplementary figure 2. Results from viral identification analysis for VIBRANT with no virome flag, (A) MCC, (B) recall, (C) false discovery rate, (D) accuracy, (E) F1, (F) PVV, and (G) specificity of (i) virome_1, (ii) virome_2, (iii) bulk_1, and (iv) bulk_2 datasets. The upper error bars represent the mean of the replicates.

Supplementary figure 3. False-discovery rate of viral sequence identification tools when challenged with Archaea, Eukaryote and Plasmid datasets. In addition to the parameters used in (Figure 2 and Figure S1), two additional settings of VIBRANT, i.e. without virome flag, were added, and two VirSorter, i.e. without virome flag (standard setting) were also applied. A color gradient represents the viral identification tool. The upper error bars represent the mean of the replicates (see more in Materials and Methods).

Supplementary figure 4. Pairwise analyses of the combined viral identification tools efficiency parameters for varying genome sizes. Metrics used included (A) MCC, (B) recall, (C) false discovery rate, (D) accuracy, (E) F1, (F) PVV, and (G) specificity of virome_1, virome_2, bulk_1, and bulk_2 datasets. The Wilcoxon test was used to compare fragment sizes (non-parametric) of the overall performance of viral identification. The 20k dataset was used as reference data (see more in Materials and Methods). Only significant comparison is shown as follows, *: $p$-value $\leq 0.05$; **: $p$-value $\leq 0.01$; ***: $p$-value $\leq 0.001$; ****: $p$-value $\leq 0.0001$. 

PeerJ reviewing PDF | (2020:11:55638:2:0:NEW 20 Apr 2021)
Supplementary figure 5. Gene-sharing network topology and performances. (A) the gene-sharing network for genome 3 kbp size, and (B) complete-genome. (C) the count of new viral clusters and clusters with reference genomes, and (D) the percentage of the overlapping, outliers, and singletons.

Supplementary figure 6. Percentage of ‘correct’ genus-assignments. (A) vConTACT v2 and (B) VFP-class. The analysis was performed by evaluating the correct genus assignment in comparison to the complete genomes. Only the genus-affiliated genomes were considered in the evaluation for vConTACT v2.

Supplementary figure 7. Comparison of the performances of viral gene-sharing network-based classification, vConTACT v2 across fragment size, (A) Accuracy (Acc), (B) Clustering-wise separation (Sep), (C) clustering-wise PPV (PPV), (D) clustering-wise sensitivity (Sn), (E) cluster-wise separation (Sep_{cl}), and (F) complex-wise sensitivity (Sep_{co}). The complete formula can be seen in Jang et al., (2019).

Supplementary figure 8. Performance evaluation of VipTree and VIRIDIC across fragment sizes. (A) ViPTree, and (B) VIRIDIC. Because VIRIDIC and ViPTree do not provide a taxonomic affiliation per se but a distance matrix, the differences in performance between different fragment sizes were calculated by comparing the pairwise distances obtained from the fragment sizes with the distances obtained from the complete genomes for the same pair, expressed as a percentage relative to the distance obtained from complete genomes. While these percentage reflects differences in the similarity matrix obtained with different fragment sizes for a given tool, these cannot be directly compared between tools or to the performance metrics used to assess the impact of genome fragmentation on the taxonomic classification obtained by vConTACT v2 and VPF-Class.
Table 1

The comparison of the commonly-used viral identification tools
Table 1. The comparison of the commonly-used viral identification tools

| Tool       | Approach                                                                 | Basic mode                                      | Ability to process modern-scale (viral) metagenomes | Reference                          |
|------------|---------------------------------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-----------------------------------|
| VirSorter  | Gene-content-based tool. Features include enrichment in viral-like genes, depletion in PFAM hits, enrichment in short genes, and depletion in coding strand changes | Permissive cut-off category 1-6                  | Scalability                                    | (Roux et al., 2015)               |
|            |                                                                           | Conservative category 1245                      | Setting for -virome, enable virome decontamination mode |                                   |
|            |                                                                           | Yes                                             |                                                |                                   |
| MARVEL     | Gene-content-based tool. Features include average gene length, average spacing between genes, density of genes, frequency of strand shifts between neighboring genes, ATG relative frequency, and fraction of genes with significant hits against the pVOGs database. | Permissive cut-off ≥70%                         | Yes                                             | (Amgarten et al., 2018)           |
| VIBRANT    | Gene-content-based tool. Features include ratio of KEGG hits, ratio of VOG hits, ratio of PFAM hit, as well as presence of key viral-like genes (e.g. nucleases, integrase, etc) | Default                                         | Yes                                             | (Kieft, Zhou & Anantharaman, 2020) |
| MetaPhinder| Integrated analysis of BLASTn hits to a, bacteriophage database, no gene prediction or amino acid-level comparison | Default                                         | Yes                                             | (Jurtz et al., 2016)              |
| DeepVirFinder | K-mer based similarity to viral and host databases, no gene prediction or amino acid-level comparison | Permissive cut-off score ≥0.7, Medium ≥0.90, Conservative ≥0.95, and p-value ≤0.05 | Yes                                             | (Ren et al., 2019)               |
| VirMiner   | Gene-content-based tool. Features include ratio of hits to KO, ratio of hits to POGs, ratio of hits to PFAM, and presence of hallmark genes | Default. Web server: http://147.8.185.62/VirMiner/ | No                                              | (Zheng et al., 2019)             |
| PHASTER    | Gene-content-based tool. Features include number of phage-like genes, with additional annotation of e.g. tRNA to better predict prophage boundaries. | Default. Web server: https://phaster.ca         | No                                              | (Arndt et al., 2017)             |
Table 2

Auxiliary metabolic genes (AMGs) curation guidelines
| Parameters                                           | Analysis program | Cut-off\(^a\) | Note                                                                 | Reference                                                                 |
|------------------------------------------------------|------------------|---------------|----------------------------------------------------------------------|----------------------------------------------------------------------------|
| Viral assembled contig quality assessment            | CheckV           | Complete viral contigs | -                                                                   | (Nayfach et al., 2020)                                                    |
| ViromeQC                                             | Default          | -             | -                                                                   | (Zolfo et al., 2019)                                                      |
| AMG identification                                   | VIBRANT          | Default       | -                                                                   | (Kieft, Zhou & Anantharaman, 2020)                                       |
| DRAM-v                                               | Default          | Putative AMG criteria: AMG score 1-3, and -M and -F flag. | (Shaffer et al., 2020)                                                  |
| Conserved residues and active sites                  | PROSITE          | Default       | PROSITE collection of motifs (ftp://ftp.expasy.org/databases/prosite/prosite.dat) database | (Sigrist et al., 2013)                                                   |
| HHPred                                               | Default          | database: PDB_mmCIF70_23_Jul | -                                                                   | (Zimmermann et al., 2018)                                                |
| BPROM                                                | Linear discriminant function (LDF) >2.75 | Bacteria σ-70 Promoters. In intergenic region or within 10 bp of start or stop of ORF | (Richardson & Watson, 2013)                                              |
| TransTermHP                                          | Confidence score > 90% | Terminators search | -                                                                   | (Kingsford, Ayanbule & Salzberg, 2007)                                   |
| ARNold                                               | Default          | Terminators search | -                                                                   | (Macke et al., 2001)                                                     |
| Protein structural                                   | Phyre\(^2\)      | 100% confident and ≥70% alignment coverage | Secondary and tertiary structure search | (Kelly et al., 2015)                                                    |
| SWISS-MODEL                                          | Global Model Quality Estimation (GMQE) score above 0.5 | Quaternary structure | -                                                                   | (Waterhouse et al., 2018)                                                |
| I-TASSER                                             | Default          | Protein structural | -                                                                   | (Yang & Zhang, 2015)                                                     |
| TMHMM                                                | Default          | Transmembrane domain | -                                                                   | (Krogh et al., 2001)                                                     |
| Synonymous and non-synonymous mutation               | MetaPop          | <0.3 represent strong purifying selection | Calculate the \(pN/pS\) | (Schloissnig et al., 2013; Gregory et al., 2020) |

\(^a\) The recommendation cut-offs that can be used in each step of AMGs curation
First, the viral RefSeq, prokaryote, eukaryote, and plasmid genome sequences were fragmented, from 5’ to 3’ end direction, into non-overlapping fragments of different lengths, i.e. $L = 500$ bp, $1$ kbp, $3$ kbp, $5$ kbp, $10$ kbp, and $20$ kbp fragments. Then, these non-overlapping fragments were randomly sub-sampled to obtain simulated input datasets. For virus identification analysis, these simulated datasets were designed to resemble mock communities with different ratios of viral, prokaryote, eukaryote and plasmid sequences, i.e. virome_1 ($10:1:0.1:0.01$), virome_2 ($10:1:0.01:1$), bulk_1 ($1:10:0.1:10$), and bulk_2 ($1:10:1:1$). For viral classification analysis, simulated inputs were exclusively composed of fragmented viral genomes.
Figure 2

The viral identification analysis across datasets

**Figure 2. The viral identification analysis across datasets. (A-D)** Pie-charts of the composition of the datasets depicted the different fragment sizes of the (i) virome_1, (ii) virome_2, (iii) bulk_1, and (iv) bulk_2. **(E-H)** The viral identification efficiency was calculated as Matthew’s correlation coefficient (MCC), where 1 represents perfect efficiency, **(I-L)** Percent of recall (%), and **(M-P)** Percent of false-discovery (%) of DeepVirFinder, MetaPhinder, MARVEL, VIBRANT, and VirSorter. For DeepVirFinder, three cutoffs were evaluated, i.e. score $\geq 0.7$, $\geq 0.9$, $\geq 0.95$, and p-value $\leq 0.05$. For MARVEL, two cutoffs were used, i.e. scores of $\geq 70\%$ and $\geq 90\%$. Next, we use two different versions of VirSorter, i.e. v1.05 and v1.10, and two cutoffs, i.e. category 1,2,3,4,5,6 and category 1,2,4,5. The upper error bars represent the mean of the replicates.
Figure 3

Viral classification analysis

(A) Percentage of the input sequences in a vConTACT v2 cluster and (B) Percentage of sequences assigned by VFP-class to a genus. The performance of VPF-class was calculated using confidence score (CS) and membership ratio (MR) thresholds of ≥0.2 (Pons et al., 2021).
Figure 4

Proposed workflow and curation step for AMG identification and validation

The recommended steps of a candidate AMGs include, (i) viral contig identification and quality assessment, (ii) AMG identification, genomic context assessment, and functional annotation, and (iii) further investigation of putative AMGs function.
Manuscript to be reviewed

Identification of candidate AMGs
(i) AMG score of 1-3.
(ii) AMG_flag of M and F.

VIBRANT:
* AMG's identification and metabolic pathways assignment

Genomic context assessment of candidate AMGs
(i) Location of AMG's with respect to viral hallmark and viral-like genes.
(ii) RNA, attachment sites.
(iii) Promoters o-TG and p-terminators.

Candidate AMG genomic context - uncertain or unlikely viral

Candidate AMG genomic context - likely viral

AMG functional analyses:
(i) Conserved protein region, active sites
(ii) Protein structure, and
(iii) phylog.

Likely metabolic

Unlikely metabolic

Putative AMGs

Additional (optional) analysis of putative AMGs

Ecological analysis

Phylogeny analysis

Experimental validation

Genuine AMGs

Further investigations of putative AMGs

Transcriptomic analysis

Distribution

Host linkage

Hypothesized metabolic functions