Viromes vs. mixed community metagenomes: choice of method dictates interpretation of viral community ecology

AUTHORS:
James C. Kosmopoulos[1,2], Katherine M. Klier[1,3], Marguerite V. Langwig[1,3], Patricia Q. Tran[1], and Karthik Anantharaman[1,4*]

[1]Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin, USA
[2]Microbiology Doctoral Training Program, University of Wisconsin-Madison, Madison, Wisconsin, USA
[3]Freshwater and Marine Sciences Program, University of Wisconsin-Madison, Madison, Wisconsin, USA
[4]Department of Integrative Biology, University of Wisconsin-Madison, Madison, Wisconsin, USA
[*]Correspondence: karthik@bact.wisc.edu

ABSTRACT

Viruses, the majority of which are uncultivated, are among the most abundant biological entities in microbiomes and ecosystems on Earth. From selfishly altering community-wide microbial physiology to driving microbiome dynamics, viruses (particularly bacteriophage) are fundamental members of microbiomes. While the number of studies leveraging viral metagenomics (viromics) for studying uncultivated viruses is growing, standards for viromics research are lacking. Viromics can leverage computational discovery of viruses from total metagenomes of all community members (hereafter metagenomes) or use physical separation of virus-specific fractions (hereafter viromes). However, differences in the recovery and interpretation of viruses from metagenomes and viromes obtained from the same samples remains understudied. Here, we compare viral communities from paired viromes and metagenomes obtained from 51 diverse samples across human gut, soil, freshwater, and marine ecosystems. Overall, viral communities obtained from viromes were more abundant and species rich than those obtained from metagenomes, although there were some exceptions. Despite this, metagenomes still contained many viral genomes not detected in viromes. We also found notable differences in the predicted lytic state of viruses detected in viromes vs. metagenomes at the time of sequencing. Other forms of variation observed include genome presence/absence, genome quality, and encoded protein content between viromes and metagenomes, but the magnitude of these differences varied by environment. Overall, our results here show that one's choice of method can lead to differing interpretations of viral community ecology. We suggest that the choice of whether to target a metagenome or virome should be dependent on the environmental context and ecological questions being asked. However, our overall recommendation to researchers investigating viral ecology and evolution is to pair both approaches to maximize their respective benefits.

KEYWORDS:

Virus, bacteriophage, virome, viromics, metagenome

INTRODUCTION

Viruses exist in all known ecosystems and infect cells from all domains of life. Being the most abundant biological entity on Earth [1,2], viruses significantly impact the ecology and evolution of their hosts [3,4], play pivotal roles in microbial community succession [5], contribute to community-wide metabolic processes [6–8], and are a source of novel therapies being used to combat a worldwide antimicrobial resistance crisis [9,10]. Advances in these areas have been enabled by large-scale investigations into entire communities of viruses which have revealed tremendous amounts of previously unknown virus diversity in human [11–13] and environmental [14–18] systems. Since their hosts largely have not been isolated, these investigations have utilized viral metagenomics (viromics) to examine thousands of viral genomes from DNA/RNA sequence data extracted directly from host-associated and environmental samples. While the number of studies using viromics has been growing in the past decade [18–20], the sampling and analytical methods used can vary greatly [20,21]. Although there have recently been efforts to establish standards of analyzing viruses from sequence data [19–21], standards in sampling methodologies are still largely lacking.

Viral sequences can be identified among metagenomes sourced from a mixed microbial community (hereafter metagenomes), or by separation of virus-like particles (VLPs) from a sample to enrich for viral community DNA prior to sequencing (hereafter viromes). The latter method typically results in higher recovery of viruses and viral genome coverage and contiguity [22,23]. However, many studies typically use one method over the other with different justifications. Since viromes do not offer the host context that metagenomes can [24,25], investigations into virus-host relationships can benefit from metagenomes. On the other hand, predicting virus-host relationships from metagenomes alone remains difficult and can often only be achieved for a fraction of viral genomes [24,25]. Also, although rare, low-abundance viruses are diverse and have significant impacts on their communities [26–28], they are often not detected in metagenomes, but are detectable in viromes [29]. It has also been argued that that active viruses exist mostly in an intra-cellular state and therefore metagenomes are more likely to be appropriate to study viral communities [30,31], especially in low biomass or slow-growing environments. However, globally high rates of viral lysis and virion production [32] might suggest that sequences captured in viromes could potentially better reflect the active viral community, especially since most sampling efforts only capture specific temporal and environmental contexts which may not reflect the true physiology and spectrum of viral infections. Overall, most studies of viral ecology typically use either method depending on their scope and environmental context.

Although most viral ecology studies have typically utilized either viromes or metagenomes, few have leveraged both methods. Except for a small number of studies in specific environments such as soil [5,6,29], comparative analyses of viral communities across paired viromes and metagenomes has been lacking. In an agricultural soil ecosystem, the cumulative richness of viruses in viromes was orders of magnitude greater than that of metagenomes [29]. On the other hand, in a seasonally anoxic freshwater lake, viromes were richer in viruses than metagenomes [6] but the magnitude

of this difference was much smaller than that of the soil study. Viruses from the freshwater lake also primarily grouped by sample type (viromes or metagenomes) [6], while viruses assembled from viromes and metagenomes of human gut communities largely grouped by the disease status of the human host, rather than sample method [33]. These studies individually offer novel insights into the viral and prokaryotic community composition of their respective ecosystems, but they remain to be synthesized together into a broader context of method application.

The few existing studies leveraging paired viromes and metagenomes have largely paid attention to community-level differences in viruses assembled from each approach, but it remains unknown if or how this influences interpretation of ecology and evolution variation, and abundance of viruses at the genome or gene level. Considering many viral genes can be identified from viromes and metagenomes in fragmented or even unassembled viral sequences [34], strictly community-level comparisons of viromes and metagenomes may be unable to consider viral genes that are absent from contiguous viral genomes. Similarly, while differences in contiguity and genome assembly quality between the two methods have been discussed [22,23], focused comparisons of viral genomes assembled from viromes versus metagenomes are lacking. While a handful of studies have been able to compare paired viromes and metagenomes, their isolated contexts in addition to their unique motivations and analytical methods make it challenging to draw any broad conclusions of how choice of viromes versus metagenomes affect downstream results in a variety of environments and contexts. Overall, it is still not well understood how choice of these two viromics approaches influences one's interpretation of viral community ecology.

Here, we set out to directly compare paired viromes and metagenomes from multiple samples obtained from four vastly different environments: a freshwater lake, the deep sea, the human gut microbiome, and soil. After using the same, standardized analytical workflow for every sample and each environment, we first compared virome and metagenome samples with respect to their overall sequencing yield and assembly quality. We also compared the ability of each method to detect viral genomes from virome/metagenome assemblies, and we determine the number of viral species uniquely found in assemblies of one of the methods as well as the number of viral species detected in both methods. The lytic states and taxonomy of the recovered viral genomes were also predicted and were used to determine if viromes or metagenomes were biased towards lytic phage, lysogenic phage, integrated prophage, or any taxonomic group of viruses. We also compare genome assemblies for the same viral species sourced from a virome and a metagenome. Additionally, we conducted a differential analysis of viral proteins between viromes and metagenomes and discuss how protein annotations in each method could influence one's interpretation of viral community function. Lastly, we discuss the unique insights offered by each approach and suggest when to apply viromes, metagenomes, or both methods when studying viral communities in different environmental contexts.

METHODS

139 ***Data acquisition***
140
141 In effort to compare paired viromes and mixed community metagenomes from a variety
142 of environments, sequence reads from publicly available studies. We searched for short-
143 read collections that met the following criteria: (1) both viromes and metagenomes must
144 have been generated in the same sequencing project, (2) virome samples must not have
145 underwent whole-genome or multiple-displacement amplification, (3) metadata was
146 available that allowed virome and metagenome pairs originating from the same biological
147 sample to be inferred, or read filenames made it otherwise clear, and (4) no measures
148 must have been taken to result in a biased or unbalanced sequencing depth towards
149 either viromes, metagenomes, or any specific organisms.
150 Among the datasets that met the criteria, we chose collections of paired viromes
151 and metagenomes to represent four vastly different environments: a freshwater lake, a
152 marine water column, the human gut microbiome, and soil. Raw reads from virome and
153 metagenome libraries sequenced from water column samples of Lake Mendota,
154 Wisconsin, USA [6] were chosen to represent a freshwater environment, reads from soil
155 samples of an agricultural field in Davis, California, USA [29] were chosen to represent a
156 soil environment, reads from fecal samples of a cohort in Cork, Ireland [11] were chosen
157 to represent human gut samples here, and reads from deep sea water column samples
158 taken from hydrothermal plumes of East Pacific Rise and Guaymas Basin were obtained
159 to represent marine samples, together. Soil and human gut reads were obtained from
160 NCBI GenBank [35] using SRAtoolkit (hpc.nih.gov/apps/sratoolkit.html) from BioProjects
161 PRJNA545408 and PRJNA646773, respectively. All read libraries obtained are
162 composed of paired-end Illumina reads. A detailed description of data sources for this
163 study and relevant information can be found in Supplementary Table 1.
164
165 ***Sequence read quality control and assembly***
166
167 Freshwater and marine samples were previously sequenced by the Department of Energy
168 Joint Genome Institute (DOE JGI) and thus sequence reads underwent quality control
169 (QC) and were assembled into contigs within the DOE JGI metagenome workflow [36].
170 To reduce biases that could have been introduced by different QC and assembly methods,
171 read QC and metagenome assembly was performed following the same assembly
172 workflow with the same sequence of software (and versions), commands, and
173 parameters. Briefly, raw reads from soil and human gut samples underwent quality
174 filtering and trimming with BBDuk and BBMap using rqcfilter.sh which were then error-
175 corrected with bbcms. Filtered, error-corrected reads were split into separate mates and
176 singletons using reformat.sh and read-pairs were imported to metaSPAdes v3.13.0 [37]
177 for contig assembly. Read lengths and counts at each step of QC were obtained with
178 readlen.sh from the BBTools suite (sourceforge.net/projects/bbmap/) and assembly
179 statistics were obtained for samples from all environments using metaQUAST v5.2.0 [38]
180 which were parsed in R [39] and plotted using ggplot2 [40] to generate Figure 2.
181
182 ***Virus identification, mapping, binning, quality-assessment, and taxonomic***
183 ***assignment with ViWrap***
184

For every sample, ViWrap v1.2.1 [41] was run with its assembled contigs and filtered reads using with the parameter "--identify_method vb" to only use VIBRANT v1.2.1 [42] to identify viral contigs as well as the options "--input_length_limit 10000" and "--reads_mapping_identity_cutoff 0.90" to adhere to established recommended minimum requirements for virus detection [20]. After using VIBRANT to identify viral contigs, ViWrap mapped reads to the input assembly using Bowtie2 v2.4.5 [43] and used the resulting coverage files to bin viral contigs into vMAGs with vRhyme v1.1.0 [44]. In this study, both binned viral contigs and un-binned singletons are referred to as vMAGs. The quality, completeness, and redundancy of resulting vMAGs were assessed with CheckV v1.0.1 [45]. Although CheckV was designed to assess the quality of viral genomes as individual contigs rather than vMAGs, we found it to be the best method available to estimate the completeness and redundancy of vMAGs [41]. Moreover, vMAGs within samples were grouped into genus-level clusters with vConTACT2 v 0.11.0 [46] and then into species-level clusters with dRep v3.4.0 [47]. ViWrap assigned taxonomy to vMAGs by aligning proteins to NCBI RefSeq viral proteins [48], the VOG HMM database v97 [49], and IMG/VR v4.1 high-quality vOTU representative proteins with DIAMOND v2.0.15 [50]. Summary statistics on the number of viral contigs, vMAGs, taxonomy, and genome quality gathered by ViWrap for each sample were parsed in R and plotted using ggplot2 to generate Figure 2, Figure S2, and Figure S3.

Predicting the lytic state of vMAGs

ViWrap provides a prediction on the lytic state for all vMAGs it identifies [41], i.e., whether a vMAG is likely to represent a lytic virus, a lysogenic virus, an integrated prophage flanked with cellular DNA, or not determined. ViWrap makes these determinations based on a combination of annotation results from VIBRANT and binning results from vRhyme. Possible predictions by ViWrap include "lytic scaffold", "lytic virus", "lysogenic scaffold", "lysogenic virus", and "integrated prophage". It is worth noting that ViWrap handles instances when vRhyme bins multiple integrated prophage sequences or lytic and integrated prophage sequences together by splitting the vMAG back into individual scaffolds to avoid retaining potentially contaminated bins (see github.com/AnantharamanLab/ViWrap). Furthermore, the distinction made by ViWrap between "scaffold" and "virus" depends on the genomic context of the contigs in a vMAG [44] and the estimated completion of a vMAG [45]. Here, we simplified these predictions and did not distinguish between predictions on the "virus" or "scaffold" level and used the results predicted by ViWrap to label vMAGs as "lytic", "lysogenic", or "integrated prophage".

vMAG presence/absence analysis

Although ViWrap employed dRep to dereplicate vMAGs into species-level clusters at 95% ANI within samples, species representative vMAGs were still redundant between samples after running ViWrap on each sample. To dereplicate vMAGs across all samples, an additional ANI-based approach was taken. Redundant vMAGs from each sample were gathered and dereplicated using dRep v3.4.3 [47] with a minimum genome length of 10,000 bp in addition to the options "-pa 0.8 -sa 0.95 -nc 0.85" to set the ANI thresholds

for primary and secondary clusters to 80% and 95%, respectively, and to require a minimum covered fraction of 85%, as recommended by established benchmarks for viral community analyses [20]. The parameters "-comW 0 -conW 0 -strW 0 -N50W 0 -sizeW 1 -centW 0" were also used when running dRep so the resulting species representative vMAGs chosen were the largest vMAGs in each cluster.

Bowtie2 mapping indices were created from fasta files containing all representative vMAGs from each environment, separately, to be used in competitive alignments. For each environment, filtered reads from every sample were separately mapped to the environment's mapping index using Bowtie2 v2.5.1 with default parameters to perform an end-to-end alignment and report single best matches at a minimum of 90% identity. The resulting alignment files were sorted and indexed using SAMtools v1.17 [51]. Sorted and indexed map files were used with coverM v0.6.1 (github.com/wwood/CoverM) to obtain covered fraction (genome breadth) statistics at the vMAG level for reads mapping with at least 90% identity. A minimum breadth threshold of 75% was used to establish the detection of a vMAG in each read sample in accordance with previously established recommendations [20]. Lists of unique representative vMAG IDs determined to be present in samples this way were used to generate Figure 3 and Figure S4 with the R package eulerr (CRAN.R-project.org/package=eulerr) [52,53].

Species accumulation curves

Using multi-sample map files generated in the presence/absence analysis, coverM was used to obtain trimmed-mean read coverage statistics (removing 5% of bases with highest coverage and 5% of bases with lowest coverage) for all species-representative vMAGs in each environment. These coverage tables were used with the R package vegan [54] with the function specaccum to generate to generate 1,000 random permutations of sample order using all tested sample pairs for each environment, separated by sampling method. Resulting cumulative species richness across all permutations, and their means, were plotted using gplot2 to generate Figure 4.

Virus genome assembly comparison

To address the preexisting notion that metagenomes typically result in truncated or less-complete viral genome assemblies than viromes [21,29,55], we set out find an example of such a viral genome among our vMAGs. Using our previously generated dRep results, we identified pairs of vMAGs that met the following criteria: (1) one vMAG was assembled from a virome and the other a metagenome, (2) each vMAG in the pair was placed in the same species-level cluster, (3) both vMAGs were assembled from the same sample source, (4) the virome-assembled vMAG was a single-contig and predicted by CheckV to be complete, and (5) the metagenome-assembled vMAG was predicted by CheckV to be incomplete.

A single pair was chosen among the resulting candidates based on their respective lengths. Each genome was then subject to non-competitive mapping of filtered reads from the virome and metagenome of the same sample source. Thus, the genome assemblies of each extraction method were mapped by reads from the same as well as the opposite extraction method. The read depths at each genome position were obtained using

samtools v1.17 with the option "depth", and then normalized such that $\text{normalized depth} = \log_{10} \frac{\text{depth}}{\text{total reads in sample} \div 10^8}$. The two genomes were aligned using Mauve [56] and BLASTn v2.5.0 from the BLAST+ suite [57] to identify regions in the virome-assembled genome that were missing from the metagenome-assembled genome, as well as gaps and alternate sequences in either. This revealed the metagenome-assembled genome in the pair to be on the opposite strand as the virome-assembled genome, so downstream analyses of this genome were performed on its reverse-complement. Lastly, each genome in the chosen pair was re-annotated for gene predictions and function using Pharokka v1.4.1 with default settings. The resulting read depths by genome position and unassembled regions were plotted using ggplot2 and arrows representing gene prediction coordinates were added with gggenes v0.5.1 (wilkox.org/gggenes) to generate Figure 5.

Differential abundance of viral proteins

In a similar approach to the vMAG presence/absence analysis, we sought out to identify protein-coding viral genes that were differentially abundant across virome and metagenome assemblies. For both virome and metagenome assemblies within each environment, we combined all nucleotide sequences of protein-coding genes predicted by Prodigal [58] as implemented by VIBRANT into a database of redundant gene sequences. These databases were then dereplicated, separately by environment, using MMseqs2 v14.7e284 [59]. Briefly, we used the command "mmseqs easy-search" to estimate pairwise average nucleotide identities (ANI) for all genes in each database, with parameters "--min-seq-id 0.95 -c 0.80 –cov-mode 1" to only retain alignments with minimum ANI of 0.95 and a minimum aligned fraction to the target sequence of 0.80. A clustered graph was generated from the pairwise ANI estimates using mcl with mcxload v14-137 [60] to obtain gene-clusters, and the longest gene within each cluster was chosen to be the cluster's dereplicated representative. Bowtie2 mapping indices were separately generated from the four databases of dereplicated gene representatives of each environment. For each environment, filtered reads from all samples within each were mapped to the Bowtie2 index of dereplicated genes corresponding to the same environment, using the same parameters and filtering steps as in the vMAG presence/absence analysis above.

Tables of raw mapped read counts for each dereplicated gene representative were obtained for each environment using CoverM. These tables were used to build negative bionomial generalized models of gene counts with DESeq2 [61] to infer genes that were differentially abundant across viromes and metagenomes for each environment, separately. Extraction method (virome or metagenome) and sample source were included as factors in the models for each environment, and the DESeq2 workflow employed Wald tests to compare the counts between viromes and metagenomes. For each test, the resulting $\log_2$ fold changes reported by DESeq2 were shrunken using the function "lfcShrink" with adaptive Student's t prior shrinkage estimators. We used a false-discovery rate adjusted *P*-value cutoff of 0.05 for the Wald test results as well as a minimum shrunken $\log_2$ fold change of 0.58 (corresponding to a minimum fold change of 1.5) as requirements to determine if a given gene was enriched in either virome or metagenome

samples of a given environment. Results were visualized using ggplot2 to generate Figure 6A.

PHROG [62] functional predictions for all dereplicated gene representatives were obtained by running Pharokka v1.4.1 [63] on each dereplicate gene database. The resulting PHROG annotations and functional categories were mapped back to the DESeq2 significant genes to obtain the presence of PHROG functional categories in each enrichment (virome or metagenome). The relative abundance of PHROG categories among all genes in each enrichment group was calculated and plotted with ggplot2 to generate Figure 6B. To assess the over or underrepresentation of any particular PHROG category within either enrichment group, we performed hypergeometric tests on the genes assigned to each enrichment group for every environment, separately, using the function “phyper” from the stats R package. Resulting *P*-values were false-discovery rate adjusted, and significant results were plotted using ggplot2 to generate Figure 6C.

RESULTS

Obtaining paired viromes and metagenomes

We searched for short-read collections that met the following criteria: (1) both viromes and metagenomes must have been generated in the same sequencing project, (2) virome samples must not have underwent whole-genome or multiple-displacement amplification, (3) metadata was available that allowed virome and metagenome pairs originating from the same biological sample to be inferred, or read filenames made it otherwise clear, and (4) no deliberate measures must have been taken to result in a biased or unbalanced sequencing depth towards either viromes or metagenomes. This resulted in a collection of paired viromes and metagenomes across four different types of environments with a total of 102 samples (Table 1): a seasonally anoxic freshwater lake [6] sampled at discrete depths and timepoints, marine water columns of hydrothermal vent plumes in the deep sea sampled at discrete depths, the human gut microbiome in a cohort of ten healthy adults [11], and a dataset from a tomato agricultural field with spatially-structured soil samples across plots with varying soil amendments, which was previously used to directly compare viromes and mixed metagenomes in another study [29].

Viromes from all samples used size filtration to enrich for viruses, however, the exact approaches differed slightly between environments (Table 1). Viromes of freshwater and soil samples were generated by 0.2 μm filtration (Table 1) which is a widely used method to generate viromes [64]. Human gut samples underwent 0.45 μm filtration in addition to additional centrifugation and polyethylene glycol (PEG) precipitation (Table 1) which has been established as optimal for extracting viromes from human fecal samples [65]. Marine virome extractions used a filter size of 0.03 μm (Table 1), much smaller than the filter sizes used to generate viromes from the other environments. Although virome extraction methods were not constant across the different environments, our main comparisons here are based on viromes versus metagenomes within the same environment, and thus this was not determined to be an issue.

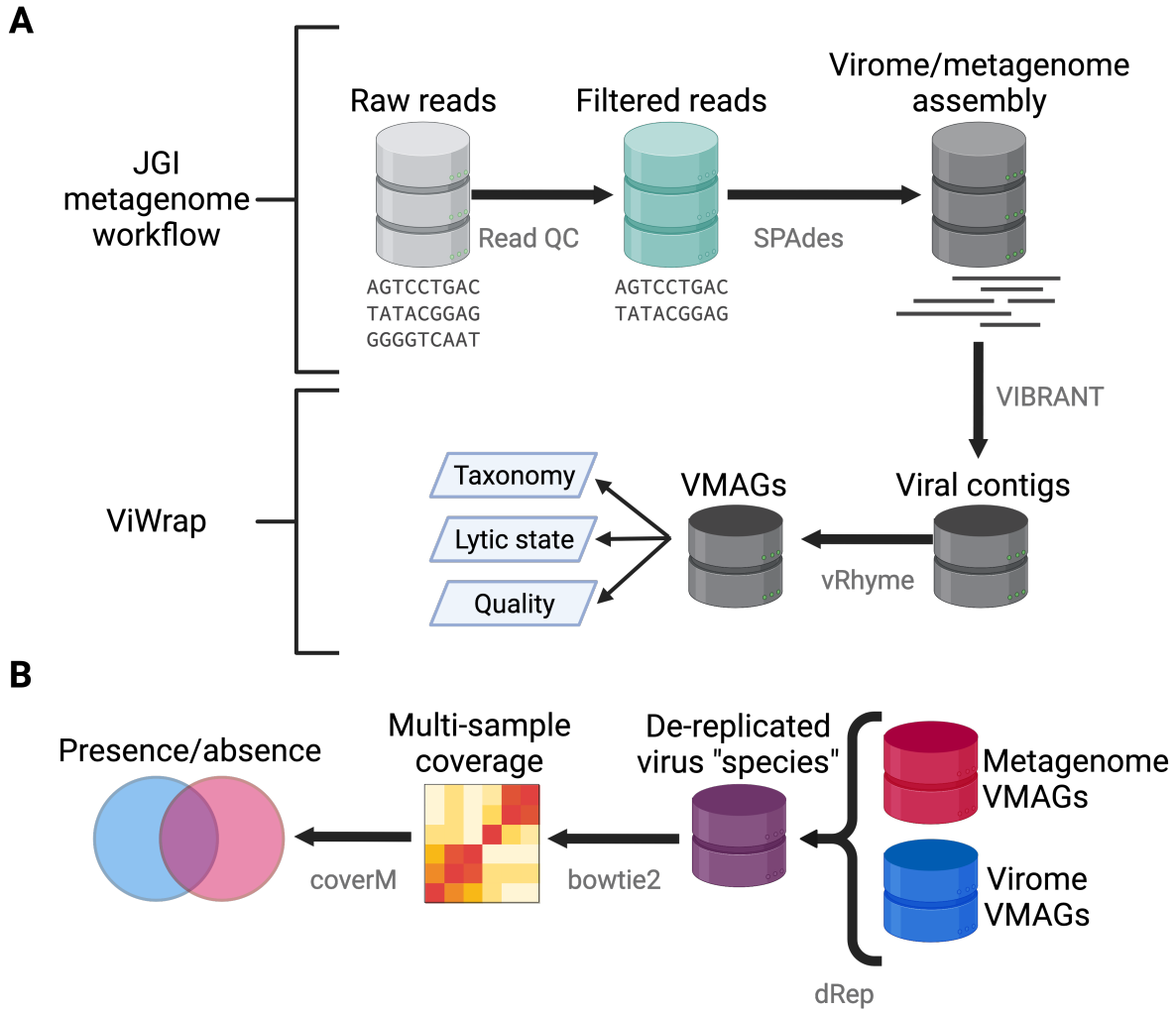


Figure 1. Summary of methods. (A) Following the Department of Energy Joint Genome Institute (JGI) metagenome assembly workflow [36], raw reads underwent quality control with the BBTools suite (sourceforge.net/projects/bbmap/) to yield filtered reads which were assembled into contigs with metaSPAdes [37]. Freshwater and marine samples underwent this workflow automatically by JGI after sequencing, while this workflow was mimicked manually using the same software, commands, and parameters for reads from human gut and soil samples obtained from NCBI. Within ViWrap [41], viral metagenome-assembled genomes (vMAGs) were generated for virome and metagenome assemblies using vRhyme [66]. (B) Resulting vMAGs were dereplicated into species-level clusters by dRep [47] and filtered reads from every sample were mapped to the resulting species representative vMAGs with Bowtie2 [43]. Mapping results were imported into CoverM (github.com/wwood/CoverM) to generate read coverage and breadth statistics that were used to determine vMAG presence/absence among samples. Figure created with BioRender.com.

Table 1: Sources of data used in this study.

Environment	Sample origin	Source	Virus enrichment approach	# of virome-metagenome	Sample design

				sample pairs used	
Human gut	Fecal samples; Cork, Ireland	Shkoporov et al., 2019 [11]	0.45 µm filtration, centrifugation, & polyethylene glycol (PEG) precipitation	10	Individuals, timepoint
Freshwater	Oxic & anoxic water columns; Lake Mendota, Madison, WI, USA	Tran et al., 2023 [6]	0.22 µm filtration & $FeCl_3$ precipitation	14	Water column depth, timepoint
Marine	Hydrothermal vent plumes; Guaymas Basin & East Pacific Rise	-	0.03 µm filtration	12 (4 Guaymas Basin + 8 East Pacific Rise)	Water column depth
Soil	Tomato field; Davis, CA, USA	Santos-Medellin et al., 2021 [29]	Amended 1% potassium citrate (AKC) resuspension, 0.22 µm filtration, & centrifugation	15	Soil amendment, plot, timepoint

Comparing viromes and metagenomes with respect to overall yield in viral sequence data

Sequencing depth within and between viromes versus metagenomes varied widely (Figure 2A). Freshwater and human gut virome samples had a significantly higher sequencing yield than their metagenome samples, while there was no difference in depth between viromes and metagenomes of soil and marine samples (Figure 2A). Although samples varied in sequencing depth, metrics hereafter are not normalized to sequencing depth since the aim of this study is to compare viruses recovered from viromes versus metagenomes. If one method led to a greater sequencing depth in an environment, normalizing by depth could inaccurately suggest that two methods do not lead to different results. The sources of the data used here were investigated to ensure that there were no deliberate actions taken during sequencing to increase the yield of one method over the other. Additionally, while multiple environments are included in this study, the goal of this study is not to directly compare viral communities across environments. Rather, we aim to compare viral communities reconstructed from viromes to viruses recovered from metagenomes from the same environment, while investigating multiple environments for a greater resolution.

Viromes from freshwater and human gut had a higher sequencing depth compared to freshwater and human gut metagenomes. Also, reads from freshwater, human gut, and soil viromes mapped back to their assembled contigs (>10 kb) at a significantly higher rate than metagenomes. However, marine viromes were not significantly different than metagenomes in their read recruitment (Figure 2B). The inconsistency in sequencing

depth and read recruitment suggest that higher or lower depth of filtered reads between viromes and metagenomes do not always lead to a higher or lower quality assembly. There was no significant difference between sequencing depth of soil viromes and metagenomes (Figure 2A), but there was a difference in read recruitment between soil viromes and metagenomes (Figure 2B). Strikingly, soil viromes recruited upward of 25% of filtered reads while all soil metagenomes recruited less than <1% of filtered reads. Further inspection of soil metagenome assembly statistics revealed a median N50 <3000, even when only calculating statistics for contigs >2000 bp (Figure S1). The poor read recruitment of the soil metagenome assemblies is likely a result of the poor contiguity of the assembly.

Although the differences between viromes and metagenomes with respect to sequencing depth and read recruitment to all assembled contigs >10 kb varied by environment, viromes from all tested environments here had reads mapping to viral contigs at a greater rate than metagenomes (Figure 2C). All assemblies (metagenomes and viromes) except for the human gut had a greater proportion of viral to non-viral contigs overall (Figure 2D). This suggests that viromes and metagenomes from all environments were successfully enriched with viral over non-viral sequences, although downstream this may not always result in a greater proportion of viral contigs >10 kb in assemblies from environments such as the human gut (Figure 2D). Accordingly, on average, viromes yielded a greater number of viral contigs (Figure S2A) in all environments except for marine assemblies. While it is a current standard in viromics to use dereplicated viral contigs to generate viral operational taxonomic units (vOTUs) when estimating viral species richness [67], viewing individual contigs as individual viral species can overestimate viral diversity and richness and is not suitable for fragmented, multi-contig virus assemblies [21,66,68]. Indeed, the N50 for all assemblies here varied greatly, with medians sometimes lower than 5000 bp (Figure S1) suggesting many assemblies here are highly fragmented. It is for this reason that viral contigs in this study were binned into viral metagenome-assembled genomes (vMAGs) to reduce overestimations of diversity due to genome fragmentation. Hereafter, we use vMAGs to refer to individual viruses or viral genomes even in cases where vMAGs only have a single contig member.

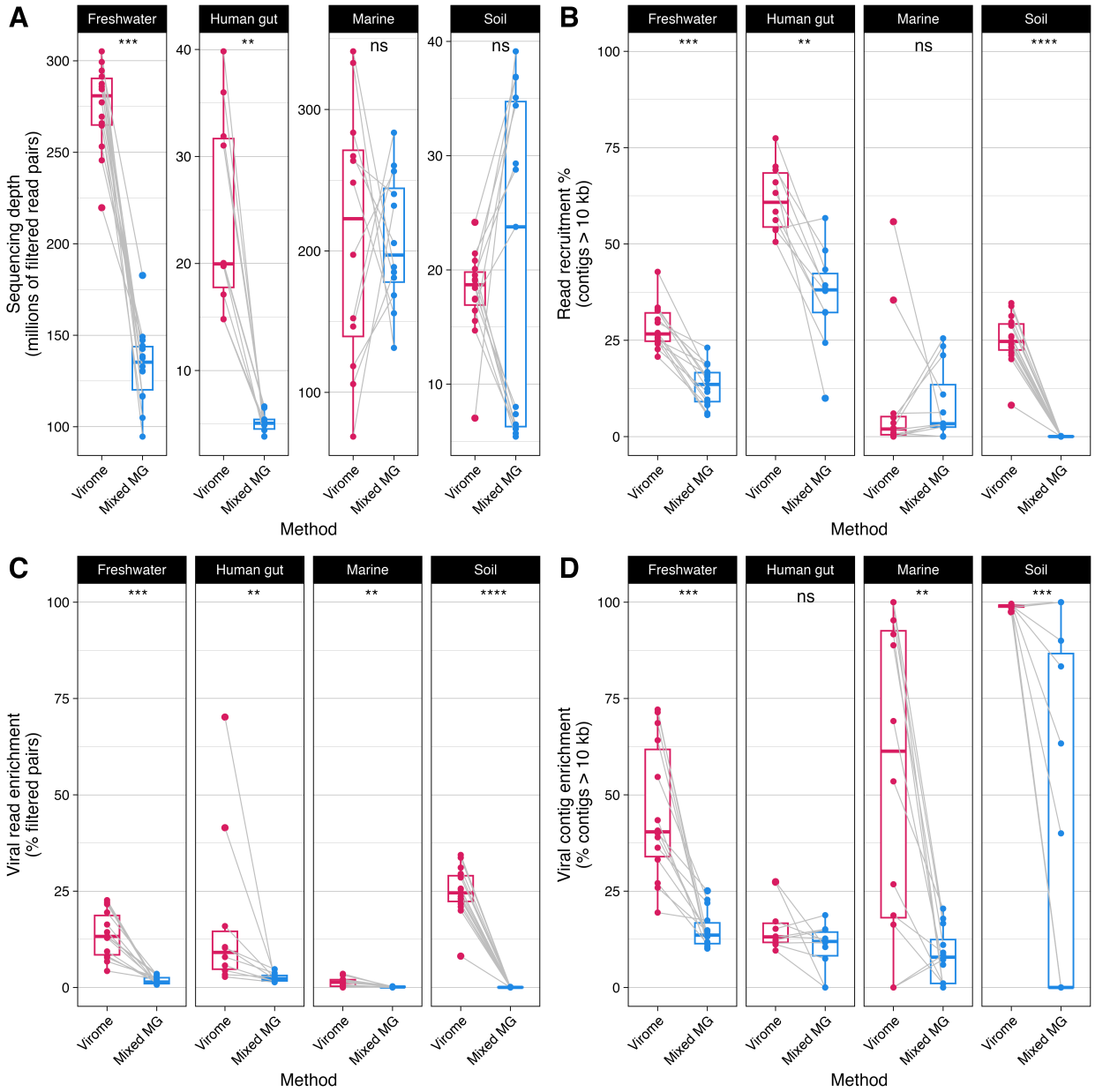


Figure 2. Differences in sequencing yield, read recruitment, viral read enrichment, and viral contig enrichment. Points indicate an individual metagenome/virome assembly, and lines connecting them signify pairs whereby the sequenced DNA was extracted from the same biological sample. Significance was inferred by Wilcoxon rank sum test: ns $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; **** $p \leq 0.0001$. (A) While virome samples yielded significantly more read pairs after quality filtering in freshwater and human gut samples, there was no difference in marine or soil samples. (B) With a minimum alignment identity cutoff of 90%, filtered read pairs from freshwater, human gut, and soil virome samples mapped back to assembled contigs >10 kb at a significantly higher rate than metagenomes, but marine samples had no difference in read recruitment. (C) In all tested environments, virome assemblies contained more read pairs mapping to viral contigs as a proportion of all quality-filtered read pairs (mapped or unmapped) than

metagenome assemblies. (D) All tested environments except human gut samples contained a greater proportion of viral contigs to all assembled contigs >10 kb.

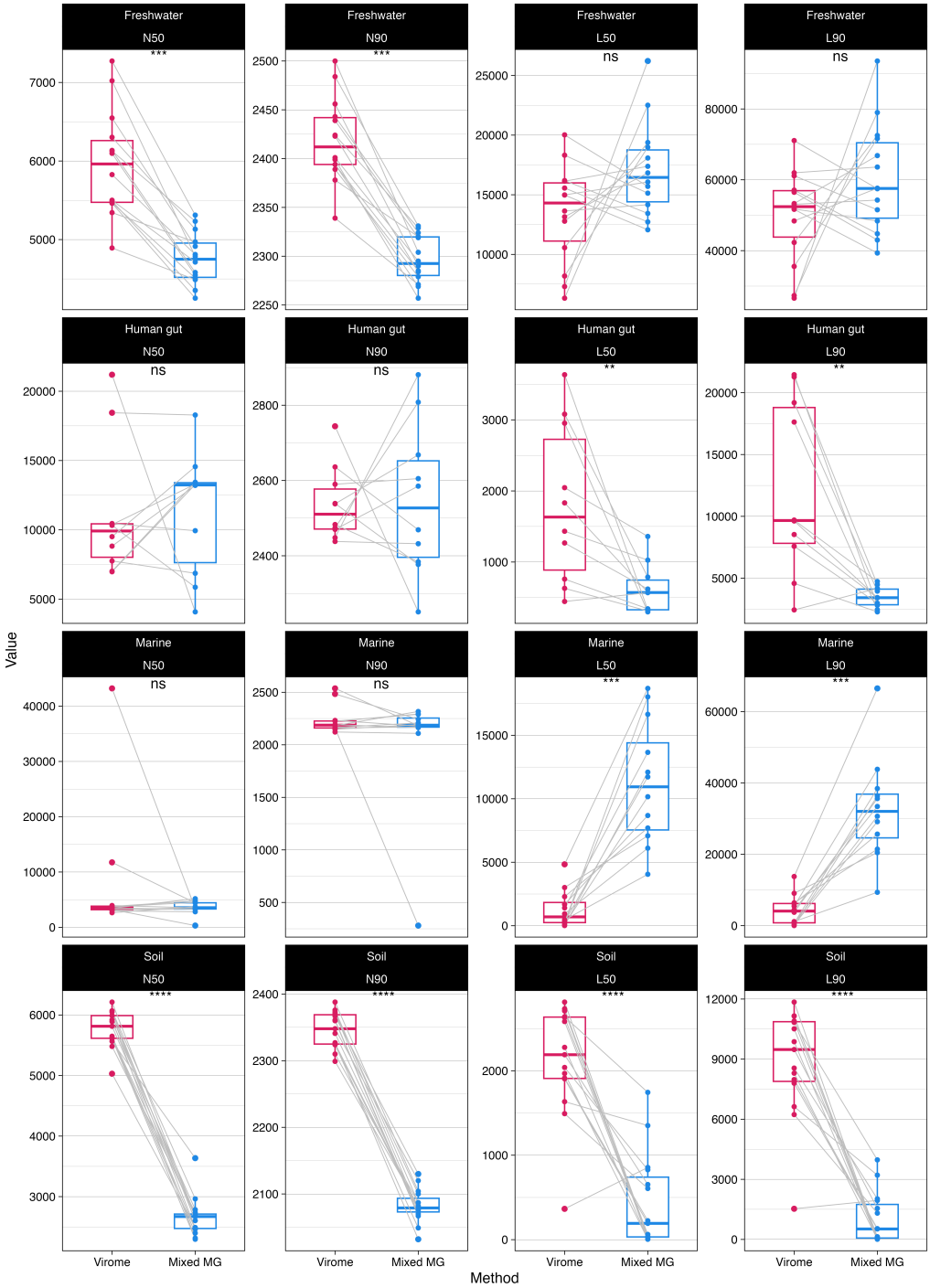


Figure S1. **Assembly statistics.** Points indicate an individual metagenome/virome assembly, and lines connecting them signify pairs whereby the sequenced DNA was extracted from the same biological sample. All statistics were calculated only using contigs >2000 bp. Note that y-axes are not the same scale across panels. Significance

was inferred by Wilcoxon rank sum test: ns $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; **** $p \leq 0.0001$.

Viromes from all environments yielded more viral contigs and vMAGs than metagenomes, with the exception of marine samples (Figure S2B). We recognize that more vMAGs gained from one method over the other is not necessarily better, since it is often that many viral genomes recovered from bulk and fragmented sequence data are incomplete or otherwise “low-quality” [45]. When considering only “high-quality” vMAGs that are estimated to represent complete or near-complete viral genomes [45], viromes, with the exception of marine assemblies, yielded significantly more high-quality vMAGs than metagenomes, although there was considerable variation within methods (Figure S2C). Similarly, after dereplicating vMAGs to species-level clusters within samples, viromes again had a higher viral species richness than metagenomes (Figure S2D), except for marine samples. The differing results of marine viromes and metagenomes with respect to the total number of viral contigs identified, vMAGs binned, high-quality genomes recovered, and viral species richness, suggests that the differences in data resulting from viromes compared to metagenomes is in some cases environment specific.

While marine metagenome samples from East Pacific Rise (EPR) used here were filtered by 0.65 µm particle size, marine metagenome samples from Guaymas basin were filtered by 0.10 µm particle size, both of which is different from metagenomes generated in freshwater, human gut, and soil samples which did not size-filter their metagenomes. It is possible that the lack of differences observed in marine samples was due to a drastic reduction in cellular community members of Guaymas metagenomes. However, the relatively unique environments from which the marine samples came have shown to be enriched in ultra-small microbes and to have communities quite distinct from other marine environments [69,70], so the reasoning behind the reduced filter sizes is justified. Although there were no differences in the total number of viral contigs, vMAGs, and species richness between marine viromes and metagenomes here, we reiterate that there was a significant difference in viral read and contig enrichment in the marine viromes and metagenomes (Figure 2C-D). Altogether, this highlights the importance of utilizing enrichment techniques that are tailored to the environment of interest and the research questions being asked.

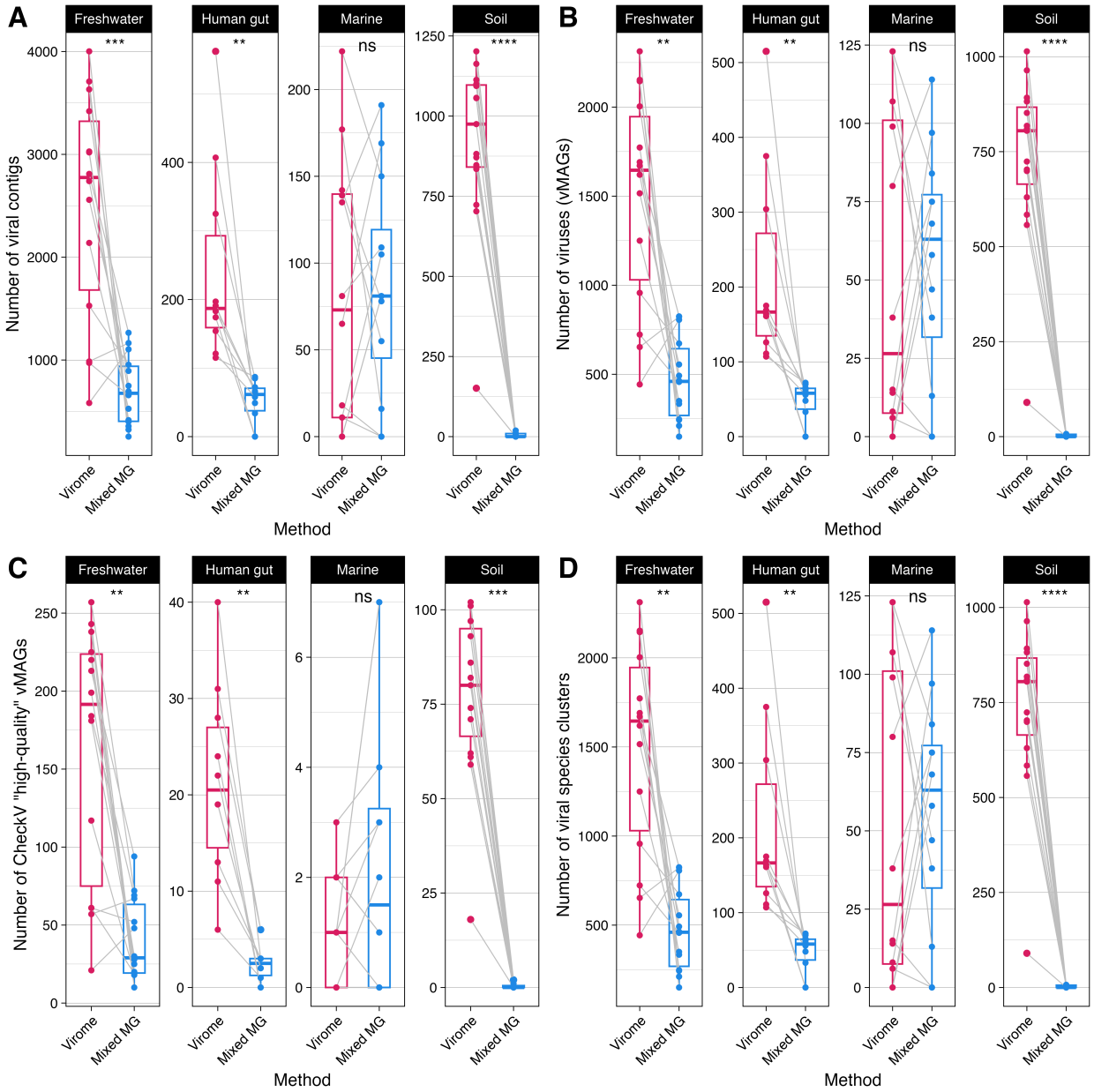
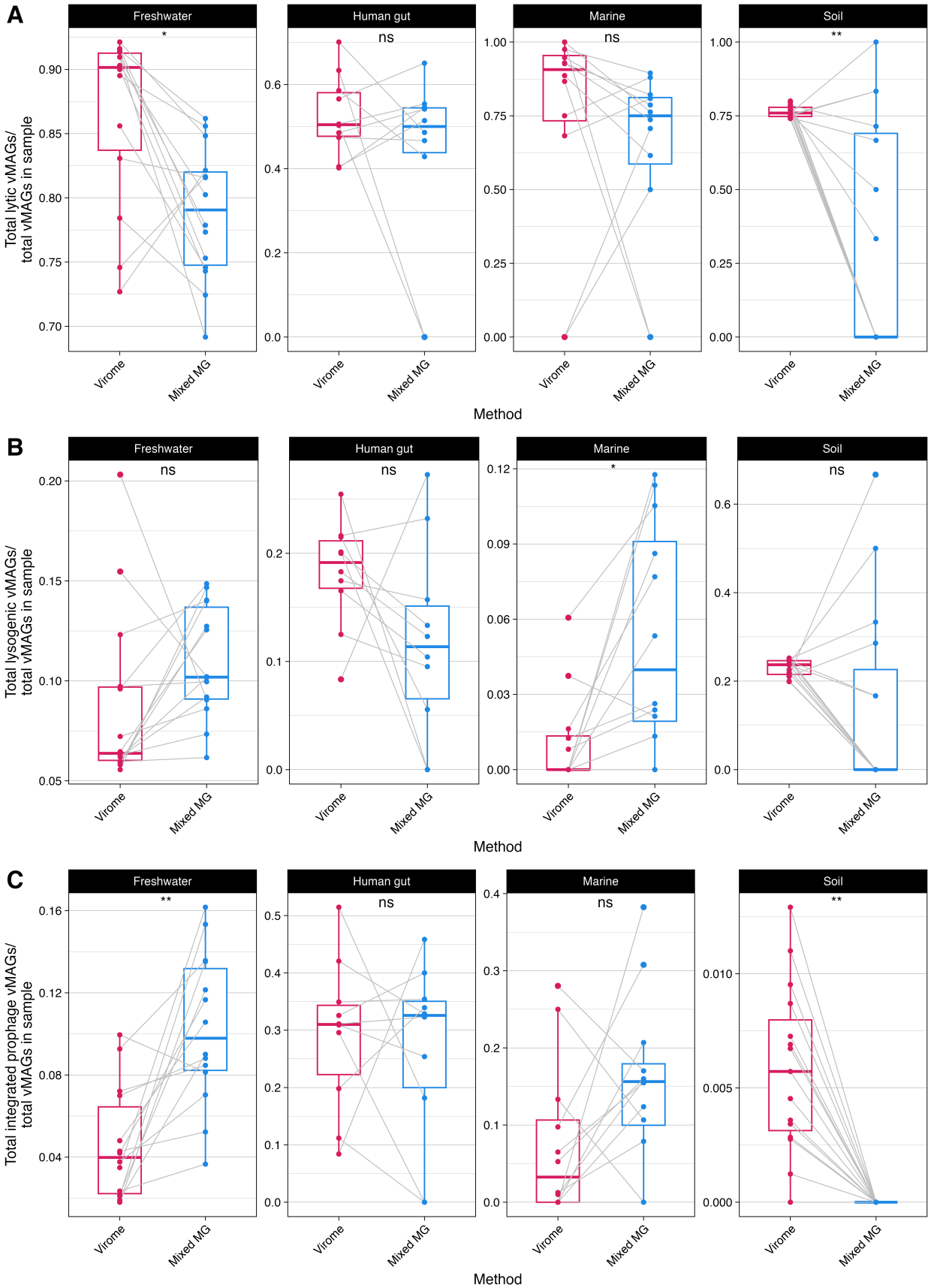


Figure S2. Viromes generally outperformed metagenomes in all tested environments except for marine samples. (A) Viromes resulted in significantly more viral contigs than metagenomes in all environments tested except for marine samples. (B) Similarly, in freshwater, human gut, and soil samples, viromes resulted in significantly more viral metagenome-assembled genomes (vMAGs) than metagenomes, whereas the number of vMAGs from marine samples was not different between viromes and metagenomes. (C) This pattern was also true for just high-quality vMAGs as determined by CheckV [45], and (D) for the number of viral species clusters at 95% average nucleotide identity. (A-D) For all panels, points indicate an individual metagenome/virome assembly, and lines connecting them signify pairs whereby the sequenced DNA was extracted from the same biological sample. Note that y-axes are not the same scale

510 across panels. Significance was inferred by Wilcoxon rank sum test: ns $p > 0.05$; * $p \leq 0.05$;
511 ** $p \leq 0.01$; *** $p \leq 0.001$; **** $p \leq 0.0001$.
512
513 ***Comparing viromes and metagenomes with respect to the lytic state of viral***
514 ***genomes***
515
516 We next asked whether there were differences between viromes and metagenomes with
517 respect to the lytic state of the viruses recovered, i.e., whether vMAGs were predicted to
518 represent lytic, lysogenic, or integrated viruses. Using these predictions generated by
519 ViWrap (see methods), we compared viromes and metagenomes in their total yield of
520 vMAGs assigned to each of these categories. Considering assemblies varied greatly in
521 the number of vMAGs recovered, we normalized the number of vMAGs belonging to each
522 lytic state in a given assembly by the total number of vMAGs in the same assembly. The
523 results varied by environment (Figure S3). Among human gut and marine assemblies,
524 there was no significant difference between the normalized number of lytic vMAGs from
525 viromes compared to metagenomes, while freshwater and soil assemblies had a higher
526 normalized number of lytic vMAGs in viromes compared to metagenomes (Figure S3A).
527 In contrast, there was no difference in the number of lysogenic vMAGs between viromes
528 and metagenomes of freshwater, human gut, and soil assemblies, while marine
529 metagenomes contained significantly more lysogenic vMAGs than viromes (Fig S3B).
530
531 Freshwater metagenomes contained significantly more vMAGs predicted to represent
532 integrated prophage, that is, an identified viral genome flanked by host DNA on each end
533 of the same contig (see methods), than viromes (Figure S3C). Interestingly, integrated
534 prophage vMAGs were found in viromes across all four environments (Figure S3C). This
535 may be expected to some degree since the virus enrichment protocols used to generate
536 viromes here are not considered 100% effective at removing cellular DNA. Indeed, soil
537 viromes contained significantly more integrated prophage vMAGs than soil
538 metagenomes. Closer inspection revealed that soil metagenomes did not contain any
539 vMAGs predicted to represent integrated prophages at all. Given that the total number of
540 vMAGs generated from soil metagenomes were so low (Figure S2B) and that soil
541 metagenomes recruited less than <1% of filtered reads overall (Figure 2B), this striking
542 difference is explained by the low virus richness in soil metagenomes overall. Lastly, while
543 there was an observable slight increase in the median normalized amount of integrated
544 prophage vMAGs from metagenomes compared to viromes in human gut and marine
545 assemblies, these differences were not significant (Figure S3C).
546
547 Considering that there were inconsistent patterns in the different lytic states of viruses
548 found between viromes and metagenomes across all environments, whether viromes or
549 metagenomes enrich for viruses of a certain lytic state likely varies. There were no
550 differences in the normalized amount of vMAGs belonging to any lytic state between
551 viromes and metagenomes of human gut samples (Figure S3A-C). However, freshwater
552 and soil metagenomes contained more lytic vMAGs than viromes (Figure S3A), with no
553 differences in the normalized amount of lysogenic vMAGs (Figure S3B), and the two
554 environments showed opposite patterns with respect to the proportion of integrated
555 prophage (Figure S3C). On the other hand, marine metagenomes contained significantly

556 more lysogenic viruses (Figure S3B) than marine viromes, but there was no difference in
557 lytic vMAGs (Figure S3A) and had no difference in integrated prophages (Figure S3C).
558 Considering that there were inconsistent patterns in the different lytic states of viruses
559 found between viromes and metagenomes across all environments, whether viromes or
560 metagenomes enrich for viruses of a certain lytic state likely varies by the environment in
561 question.
562


A
Freshwater
Human gut
Marine
Soil
Total lytic vMAGs/
total vMAGs in sample
*
ns
ns
**
Virome
Mixed MG
Method
B
Total lysogenic vMAGs/
total vMAGs in sample
ns
ns
*
ns
C
Total integrated prophage vMAGs/
total vMAGs in sample
**
ns
ns
**

563

Figure S3. The recovery of predicted lytic, lysogenic, and integrated viruses in metagenomes vs. viromes. Points indicate an individual metagenome/virome assembly, and lines connecting them signify pairs whereby the sequenced DNA was extracted from the same biological sample. Values represent the number of vMAGs in an assembly belonging to the facet category (i.e., lytic state ~ environment) normalized by the total number of vMAGs in the same assembly. Note that y-axes are not the same scale across panels in (C). Significance was inferred by Wilcoxon rank sum test: ns $p > 0.05$; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$.

Viruses shared between and unique to viromes and metagenomes

While viromes generally yielded more viral contigs, vMAGs, and viral species, it was not known how many vMAGs were unique to either virome or metagenome assemblies or how many were shared. To address this, we took a genome-similarity and read mapping approach (Figure 1B). After dereplicating all vMAGs across all environments and methods to species-level representatives at 95% average nucleotide identity (ANI), and then mapping reads from each sample on to an index of all unique species representatives, we used a minimum read breadth cutoff of 75% to determine whether a vMAG was present in a sample as recommended by community standards [20]. Therefore, if at least 75% of the bases on a vMAG were covered by reads from a given sample, we decided the vMAG was present in the sample, otherwise we decided the vMAG was absent.

Our dereplication and read-mapping approach yielded 24,761 unique species-representative vMAGs found in freshwater assemblies, 9,039 found in soil assemblies, 2,271 found in human gut assemblies, and 1,174 found in marine assemblies, with a total of 37,245 unique vMAGs identified across all environments (Figure 3A). Out of the total 37,245 unique vMAGs, 1,455 were found only in metagenome assemblies, 20,722 were found only in virome assemblies, and 15,068 were found in both (Figure 3B). Overall, freshwater, soil, human gut, and marine virome assemblies contained more unique vMAGs than metagenome assemblies (Figure 3C). Soil virome assemblies contained nearly all vMAGs detected in soil metagenomes, except only one vMAG was found unique to soil metagenomes (Figure 3C). Although more vMAGs were unique to viromes than metagenomes in each environment, even more vMAGs were detected in both viromes and metagenomes of freshwater, human gut, and marine samples (Figure 3C).

We previously observed inconsistent patterns in the different lytic states of viruses found between viromes and metagenomes across all environments (Figure S3A-C). We next asked whether there were similar inconsistent patterns in the number of unique and shared vMAGs between viromes and metagenomes. The detection of specifically lytic vMAGs followed the same trends as all vMAGs, with more lytic vMAGs being detected only in viromes than lytic vMAGs being detected only in metagenomes (Figure 3D), except more lytic vMAGs from human gut assemblies were detected only in viromes than were detected in only metagenomes, or both methods combined (Figure 3D). Lysogenic vMAGs also followed a similar pattern of detection (Figure 3E), however, more lysogenic vMAGs were unique to marine metagenomes than were detected in only viromes or both methods combined (Figure 3E). The single soil vMAG detected only in metagenomes was

predicted to be lysogenic and therefore was the only lysogenic soil vMAG detected only in soil metagenomes (Figure 3E). Lastly, we report many freshwater and human gut vMAGs classified as integrated prophage detected in viromes, too (Figure 3F). However, this may be due to virome reads originating from excised lysogenic/temperate virus genomes mapping to vMAGs assembled from metagenomes that were originally flanked by host DNA.

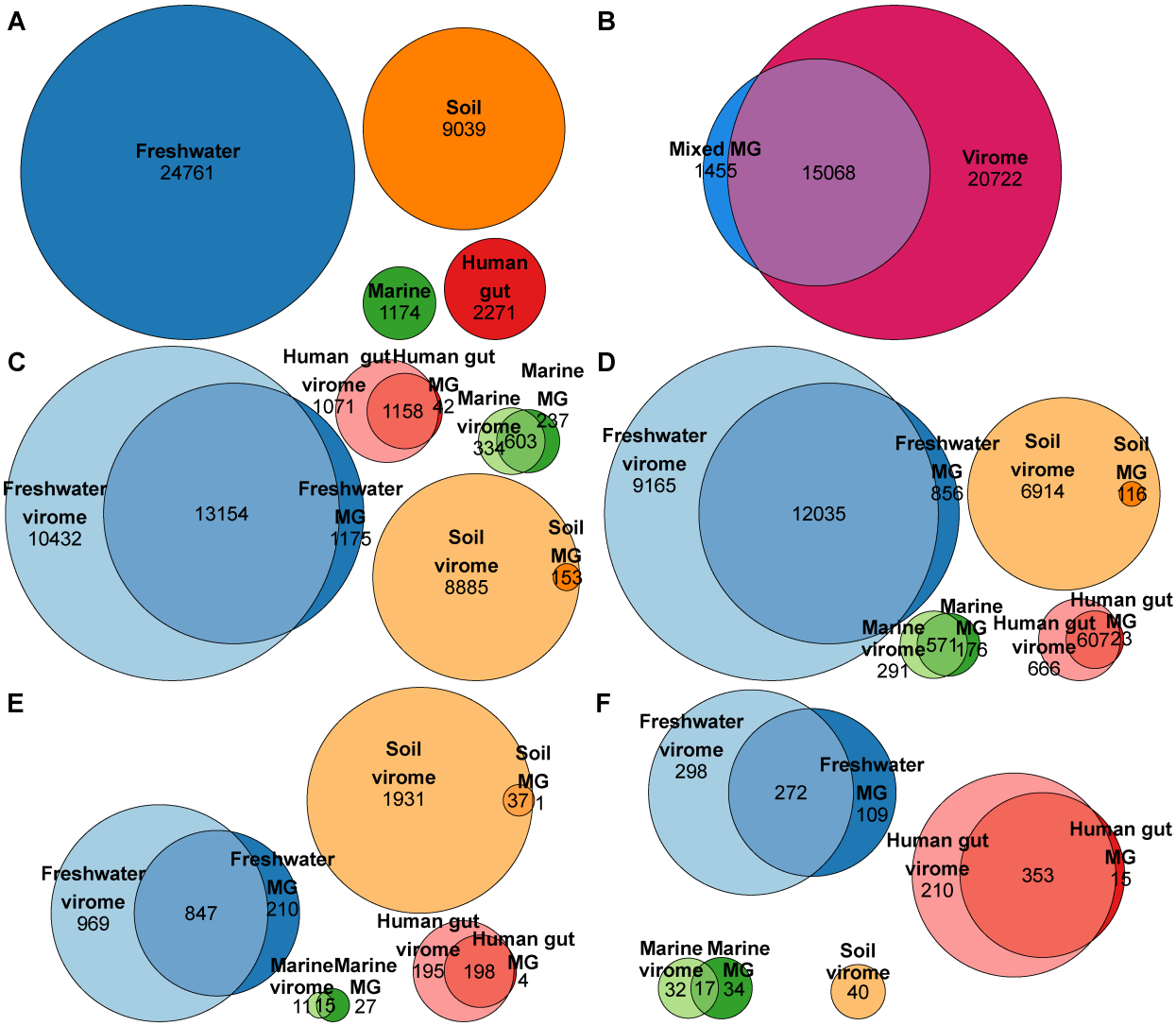


Figure 3. vMAGs assembled from viromes were not detected in most metagenome samples. Euler diagrams generated using eulerr (CRAN.R-project.org/package=eulerr) [52,53] with IDs of unique species-level vMAGs detected in the labeled category; quantities within areas are given beneath labels. An individual vMAG was marked as detected in a virome/metagenome if reads from the virome/metagenome mapped to the contigs in the vMAG with a minimum breadth of 75% across the entire vMAG. (A) Total number of vMAGs in each environment, regardless of method. (B) All vMAGs and environments, separated by method. (C) All vMAGs, separated by environment and method. (D) Predicted lytic vMAGs, separated by environment and method. (R) Predicted

lysogenic vMAGs, separated by environment and method. (F) Predicted integrated prophage vMAGs, separated by environment and method.

Viral families shared and unique to viromes and metagenomes

Having examined the degree of overlap in species-level vMAG clusters between viromes and metagenomes, we were interested in whether higher-level, defined viral taxa were uniquely found in viromes or metagenomes. Although only 1,012 out of all 37,245 unique representative vMAGs here could successfully be assigned to a described viral family (Figure S4), we used the previously generated presence/absence results in conjunction with taxonomy results generated by ViWrap (Figure 1A) to assess which known viral families were found in both viromes and metagenomes or unique to either environment.

Using the names of each viral family assigned to vMAGs from viromes and metagenomes, we found that both methods contained assemblies with many of the same viral families (Figure S4). Although 10,432 species-level vMAGs were unique to freshwater viromes compared to just 1,175 being shared with freshwater metagenomes (Figure 3C), 17 viral families were assigned to vMAGs from both freshwater viromes and metagenomes out of a total of 21 unique families assigned to vMAGs across both methods (Figure S4A). Steigviridae and Suoliviridae were unique to freshwater viromes, while Iridoviridae and Poxviridae were unique to freshwater metagenomes (Figure S5). All other environments had a complete overlap in the number of unique viral families assigned to vMAGs present in viromes and metagenomes, with 7 families assigned to marine vMAGs (Figure S4B), 11 families assigned to human gut vMAGs (Figure S4C), and 8 families assigned to soil vMAGs (Figure S4D).

Overall, of the vMAGs that could be assigned to a known viral family, identified viral families were generally present in both viromes and metagenomes. While there were two viral families unique to freshwater viromes and another two unique to freshwater metagenomes, the vMAGs assigned to these families were very rare relative to the other vMAGs in the environment (Figure S5). Since only a small fraction of vMAGs across all examined environments could be successfully assigned to a described virus family, there are likely many vMAGs within the same hypothetical “family” that were left out of the analysis shown in Figure S4. However, the near-complete overlap in vMAGs assigned to viral families known at the time of this analysis suggests that viromes and metagenomes may both contain a large proportion of vMAGs within the same virus family once virus taxonomy is expanded in the future.

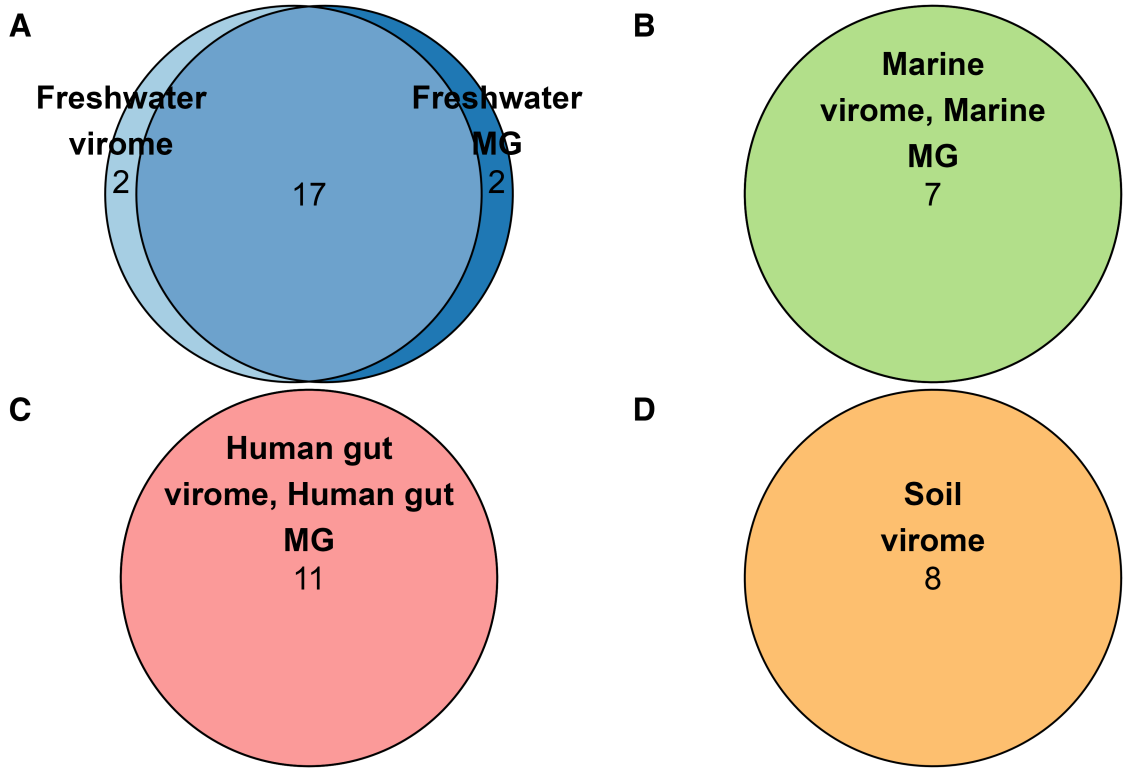


Figure S4. Families of vMAGs found in viromes and metagenomes were largely found in both methods. Euler diagrams generated using eulerr (CRAN.R-project.org/package=eulerr) [52,53] with names of unique, known viral families assigned to vMAGs detected in the labeled category; quantities within areas are given beneath labels. These Euler diagrams disregard overlap in families found in multiple environments and instead only considers the overlap in viral families found in viromes and metagenomes of (A) freshwater samples, (B) marine samples, (C) human gut samples, and (D) soil samples. Except for freshwater samples, there was a complete overlap in the unique viral families identified in viromes and metagenomes.

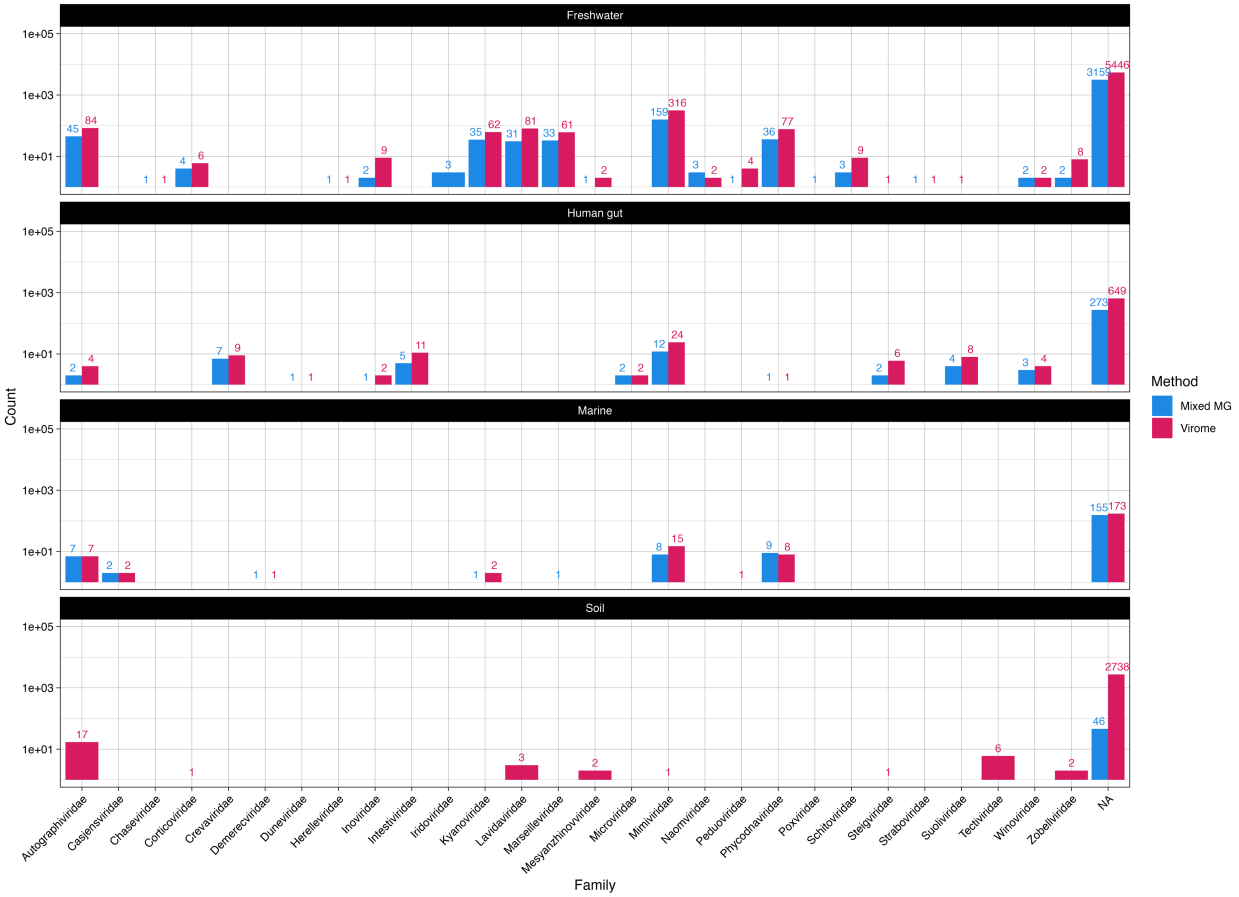


Figure S5. Viral family assignments to species representative vMAGs. Taxonomic assignments obtained from ViWrap pipeline [41]. Presence/absence of a vMAG in virome or metagenome assemblies determined by analysis summarized in Figure 1B.

Sampling effort and viral species accumulation

Using 1,000 random permutations of sampling order, species accumulation curves show that how sampling effort influenced observed viral species richness between viromes and metagenomes depended on the environment in question. Freshwater assemblies resulted in a higher cumulative richness in metagenomes than viromes, which remained consistent for all 14 sample pairs (Figure 4). This contrasts the overall greater absolute viral species richness observed in freshwater viromes compared to metagenomes (Figure S2B). On the other hand, human gut viromes consistently resulted in a higher cumulative richness than metagenomes for all 10 sample pairs, while marine assemblies had little difference in richness between sampling methods in the first three samples before viromes surpassed metagenomes in the remaining nine samples (Figure 4). The only difference was among soil assemblies (Figure 4), which, consistent with previous findings [29], an increasing sampling effort had little influence on the observed cumulative richness in metagenomes relative to viromes.

Overall, in the first random sample, random permutations of vMAG coverages resulted in a considerable difference in the observed richness between viromes and metagenomes

in freshwater, human gut, and especially soil assemblies, but not marine assemblies (Figure 4). Sampling after the first quickly began to shrink the differences in richness between sampling methods in freshwater and human gut assemblies, while increasing sampling effort enhanced this difference in marine assemblies (Figure 4). Additionally, increasing sampling effort up to 15 samples did little to bring the observed richness between soil viromes and metagenomes closer together, consistent with previous results of the same data [29]. Collectively, these species accumulation curves suggest that increasing the number of sample replicates has a large effect on observed viral species richness. The degree that sampling effort impacts the differences in richness between viromes and metagenomes varies by environment.

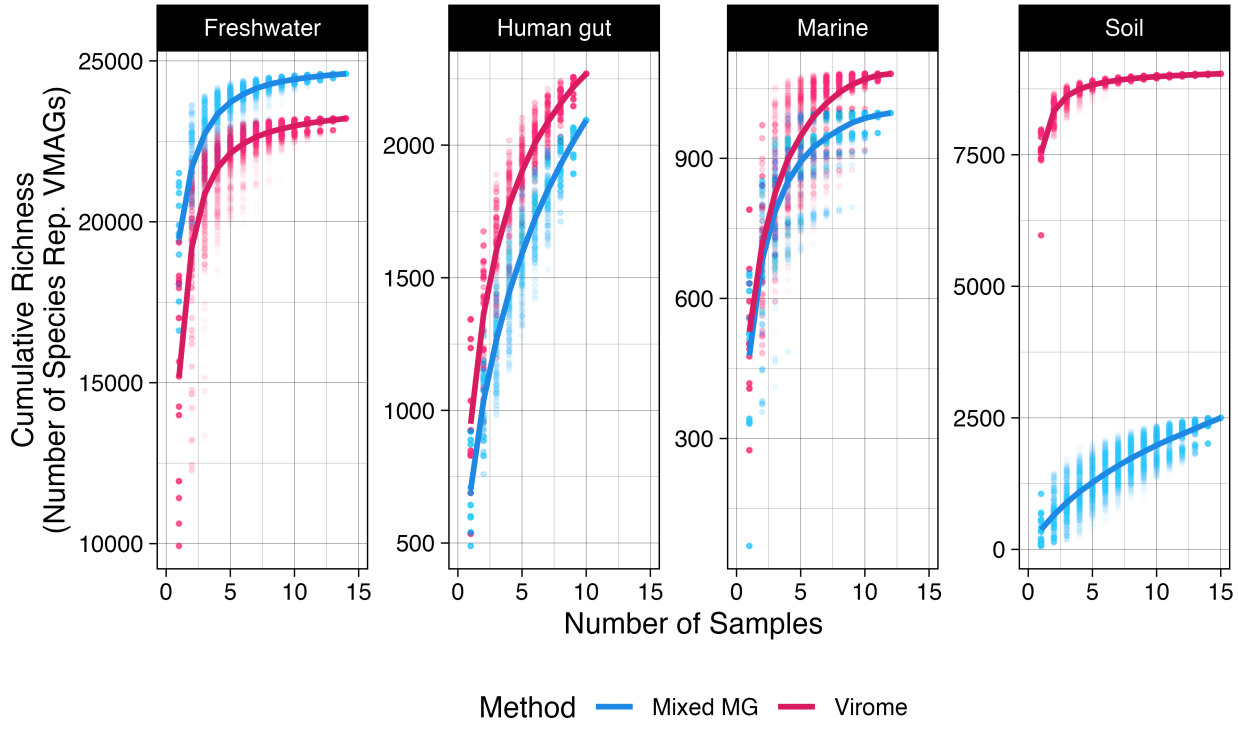


Figure 4. The effects of sampling effort on cumulative viral species richness. Points indicate the cumulative viral species richness across 1,000 random permutations of sampling order at each total number of samples along the x-axis. The average cumulative richness for each environment across all permutations for each sampling effort are represented by curves over points. The maximum sampling effort for each environment reflects the actual number for samples included in the analysis, with n = 14 for freshwater virome-metagenome pairs, n = 10 for human gut pairs, n = 12 for marine pairs, and n = 15 for soil pairs.

Comparison of virome- and metagenome-assembled viral genomes

Past arguments in favor of utilizing virome extractions to study viral communities have cited a tendency to assemble more complete viral genomes with greater depth than those assembled from metagenomes [21,29,55]. We set out to identify a pair of viral genome

assemblies, one from a virome and one from a metagenome, that highlights this case. We searched our previously generated viral species clusters for viruses assembled in both viromes and metagenomes of the same sample source. From this, we found two genomes representing a single viral species from a freshwater sample with notable differences across extraction methods. The virome-assembled viral genome was nearly 38 kb in length with 70 gene predictions (Figure 5, Table S4), and was predicted to be complete by CheckV due to the presence of direct terminal repeats. The metagenome-assembled viral genome, however, was predicted by CheckV to be incomplete and was nearly 5 kb shorter than the virome assembly and contained only 57 gene predictions (Figure 5, Table S4).

The missing regions in the metagenome-assembled viral genome spanned both ends of the contig (Figure 5). These regions covered eleven genes with unknown functions that were present in the virome but not the metagenome assembly, as well as the first 527 bases of a phage portal protein (Figure 5, Figure S4). Additionally, the virome-assembled viral genome contained a 130 bp region spanning two genes predicted to encode a hypothetical protein and a tail protein (Figure 5, Table S4). This 130 bp region was absent from the metagenome assembly, resulting in a single, fused gene prediction for a phage tail protein (Figure 5, Table S4). The only region we identified in the metagenome-assembled viral genome that was absent from the virome assembly was a single 3 bp sequence over the aforementioned portal protein (Table S4). Finally, although this genome was incompletely assembled from the metagenome, metagenome reads mapped over the entire length of virome-assembled genome (Figure 5, Table S5). Virome reads also mapped to both assemblies of the same genome with a depth up to two orders of magnitude greater than metagenome reads (Figure 5, Table S5). Overall, these results show that virome extractions can confer much greater sequencing depth of viral genomes than metagenomes and can lead to more complete or otherwise different genome assemblies.

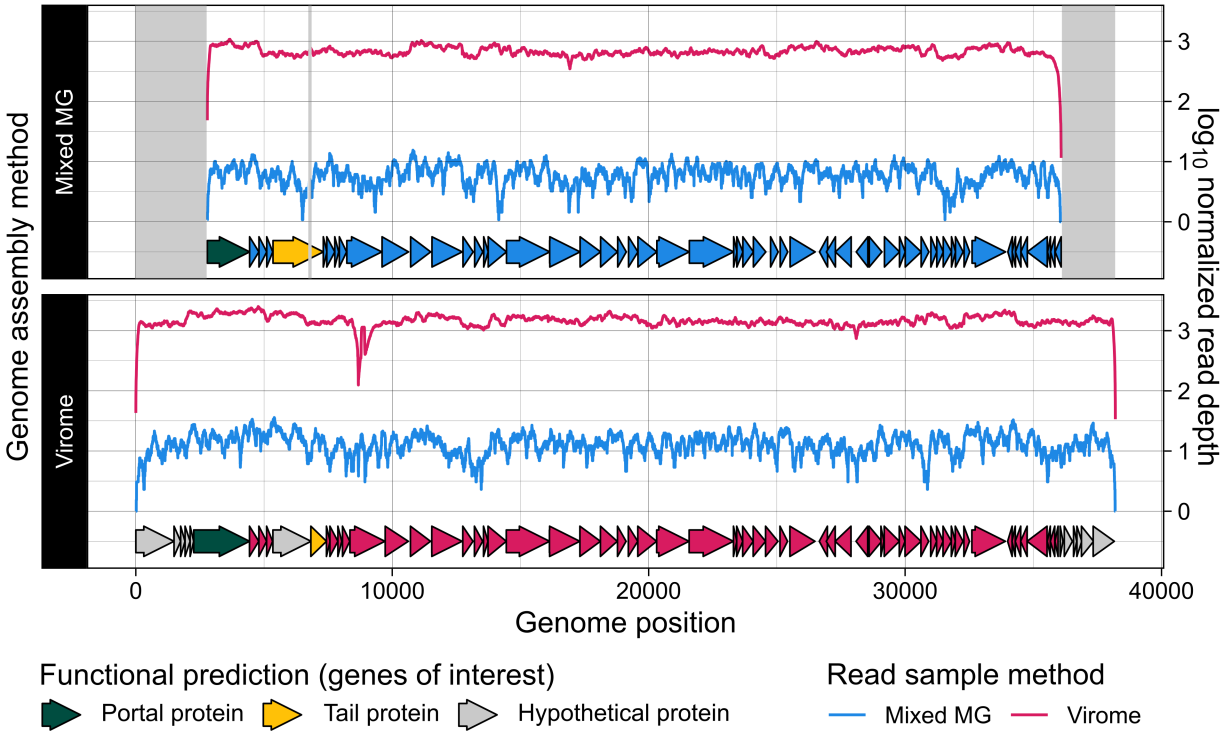


Figure 5. An incomplete metagenome-assembled viral genome was complete in its corresponding virome. A single-contig, complete viral genome identified from a virome assembly was detected but was incompletely assembled in the sample's corresponding metagenome. Areas highlighted in grey represent regions in the virome-assembled genome that were absent from the metagenome-assembled genome. Reads yielded from the virome and metagenome of the same sample source were each mapped to both versions of the genome assembly. Arrows along the x-axis represent predicted genes which are colored by the extraction method of their genome's origin, except for a selection of genes of interest that are colored by their functional predictions.

Differential abundance of protein-coding viral genes

After annotating, dereplicating, and mapping reads to protein-coding viral genes from all assembled viral contigs >10 kb, we used DESeq2 [61] to test if any genes were differentially abundant between virome and metagenome samples. A total of 303,343 genes resulted from dereplication across all environments and extraction methods, 13,099 of which came from human gut assemblies, 206,127 from freshwater assemblies, 5,463 from marine assemblies, and 78,654 were from soil assemblies (Table 2, Table S2). Out of all dereplicated genes, a total of 67,603 unique genes were differentially abundant across extraction methods (Wald test P <0.05, FDR adjusted) (Table 2, Table S2), only 55 of these genes were from the human gut, while 64,999 genes were from freshwater samples, 1,242 from marine samples, and 1,306 from soil samples (Table 2, Table S2).

Of these differentially abundant genes, using a minimum fold change cutoff of 1.5 in either direction, 64,879 in total were determined to be enriched in either virome or metagenome samples (Table 2, Table S2, Figure 6A). The remaining 2,724 were differentially abundant

782 but did not meet the minimum fold change of 1.5 (Table 2, Table S2, Figure 6A). We did
783 not identify any genes that were enriched in either virome or metagenome samples from
784 the human gut (Table 2, Figure 6A). However, 37,683 and 25,328 genes were enriched
785 in viromes and metagenomes from freshwater samples, respectively (Table 2, Table S2,
786 Figure 6A). Among marine samples, only 96 genes were enriched in viromes whereas
787 749 were enriched in metagenome samples (Table 2, Table S2, Figure 6A). Lastly, 432
788 genes were enriched in soil viromes and 591 were enriched in soil metagenomes (Table
789 2, Table S2, Figure 6A).
790
791 **Table 2. Number of genes throughout the differential abundance (DA) workflow.**

Environment	Number of genes before dereplication	Number of genes after dereplication (% of before)	Differentially abundant genes (% of dereplicated)	Virome-enriched genes (% of DA)	Metagenome-enriched genes (% of DA)
Human gut	8.39×10^4	1.31×10^4 (16%)	55 (0.004%)	0	0
Freshwater	1.02×10^6	2.06×10^5 (20%)	6.50×10^4 (32%)	3.77×10^4 (58%)	2.53×10^4 (39%)
Marine	3.61×10^4	5.46×10^3 (15%)	1.24×10^3 (23%)	96 (7.7%)	749 (60%)
Soil	4.42×10^5	7.87×10^4 (18%)	1.31×10^3 (1.7%)	432 (33%)	591 (45%)
Total	**1.58×10^6**	**3.03×10^5 (19%)**	**6.76×10^4 (22%)**	**3.82×10^4 (57%)**	**2.67×10^4 (39%)**

792
793 We sought out to predict potential functions for the differentially abundant genes that were
794 enriched in either viromes or metagenomes. Using PHROG [62] functional categories
795 predicted by Pharokka [63], out of the 64,879 unique genes enriched in viromes or
796 metagenomes across all environments, Pharokka assigned PHROG functional
797 categories to a total of 10,236 genes (16%), 6,234 in viromes and 4,002 in metagenomes
798 (Table S3). This small fraction of genes with predicted functions is consistent with the
799 current challenge of obtaining functional predictions to viral genes from complex,
800 environmental samples [71–73].
801
802 Since predicted PHROG functional categories were largely present in both virome- and
803 metagenome- enriched genes across the three environments with them (Figure 6B), we
804 performed hypergeometric tests on enriched genes from each environment to determine
805 whether any functional categories were over or underrepresented in viromes or
806 metagenomes. We found seven PHROG categories that were significantly over- or under-
807 represented between viromes and metagenomes across freshwater, marine, and soil
808 samples (hypergeometric test P <0.05, FDR adjusted) (Figure 6C). Generally, genes
809 encoding viral structural proteins such as head-tail connectors, packaging proteins, and
810 tail proteins were under-represented in metagenomes and overrepresented in viromes
811 across the three environments, except tail proteins in marine samples which displayed
812 the opposite pattern (Figure 6C). Integration and excision coding genes were
813 overrepresented in freshwater and marine metagenomes but where underrepresented in
814 marine viromes (Figure 6C). Conversely, in freshwater assemblies, lysis genes were
815 underrepresented in metagenomes and overrepresented in viromes. These results show

that the differential abundance of genes between viromes and metagenomes may be the result of structural variation or differences in replication strategies of viruses detected in either method. Overall, the enrichment of genes in viromes or metagenomes leads to varying enrichments of inferred gene functions as well.

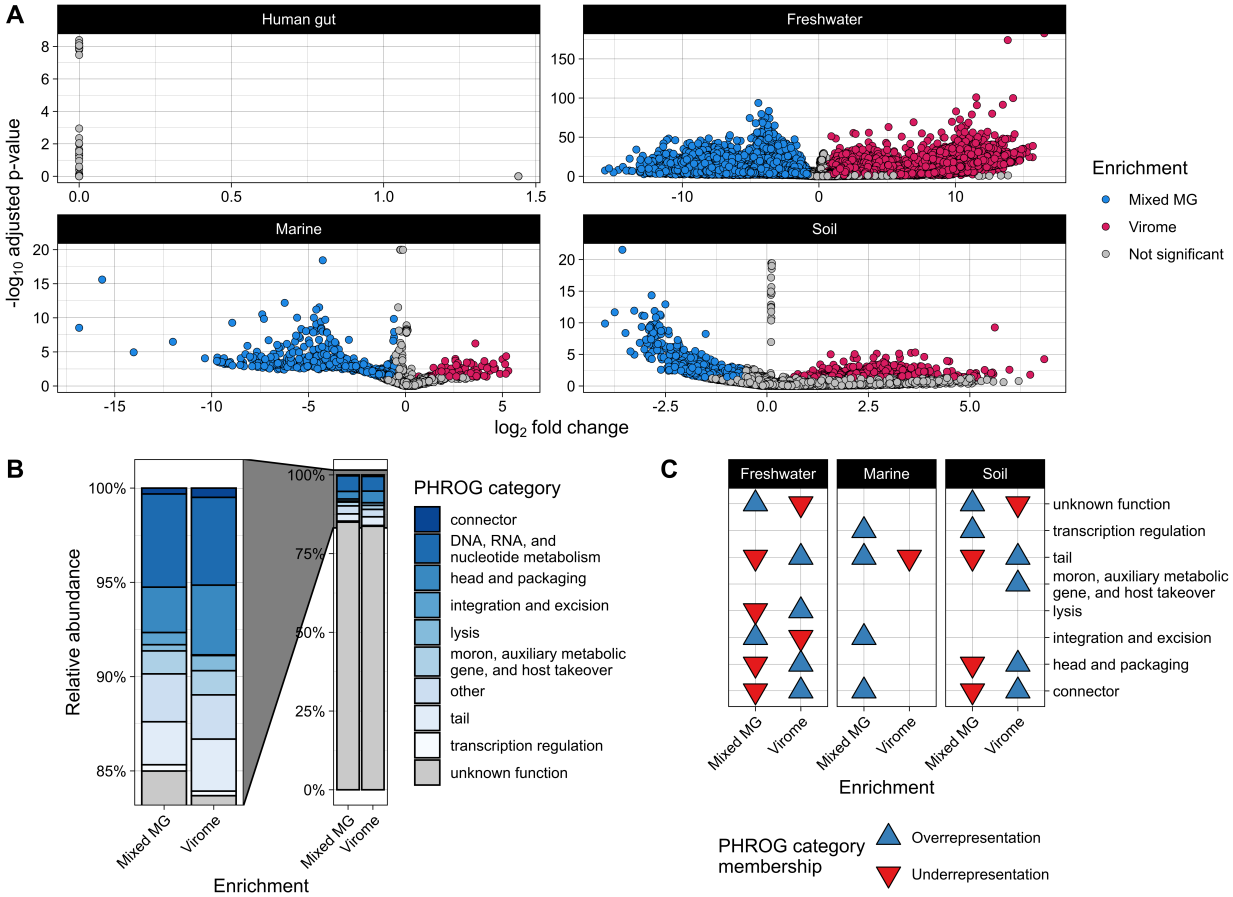


Figure 6. Protein-coding viral genes are differentially abundant across viromes and metagenomes and have predictable functions. (A) Differential abundance of protein-coding viral genes as inferred by a DESeq2 [61] workflow. Points indicate unique, dereplicated protein-coding viral genes that were annotated from viral contigs assembled from the environment indicated by the panel labels. Enrichment of a given gene in virome or metagenome samples was determined if the resulting fold change was at least 1.5. (Wald test P <0.05, FDR adjusted). No protein-coding viral genes were determined to be significantly enriched in neither virome nor metagenome human gut assemblies. (B) Relative abundance and (C) over/under-representation of PHROG [62] functional categories assigned to differentially abundant genes displayed in (A) (hypergeometric test P <0.05, FDR adjusted). Categories without an arrow in a given environment or method were not significantly over or underrepresented in that environment/method.

DISCUSSION

838 The expansion of viral genome databases and viral community ecology studies in the
839 recent years has resulted in an explosion of known viral diversity that has been enabled
840 by sequencing of whole virus communities [12,13,16,17,74,75]. Assembly of virus
841 communities can be achieved either by sequencing extracted DNA from the total, mixed
842 community of prokaryotes, eukaryotes, and viruses within a sample to generate
843 metagenomes. Viral communities can also be assembled if extra measures are taken to
844 enrich for virus-like particle DNA during extraction to generate viromes. Although viromes
845 can generally offer a more focused view of viruses in a sample compared to
846 metagenomes [23], the consequences of choosing one sampling method over the other
847 have relatively unexplored and limited to individual study ecosystems [5,6,29]. Here, we
848 applied the same analytical methods to collections of paired virome and metagenome
849 sequence reads to directly infer the unique and shared results gained from each sample
850 method. We assembled, annotated, and analyzed 51 pairs of viromes and metagenomes
851 across four different environments and found that the uniqueness and sharedness of
852 results from each method varied across environmental origin.
853
854 ***Community-level impacts***
855
856 Viromes, by design, typically allow more viral species and genome coverage to be
857 observed than what might be found in metagenomes [23]. Indeed, compared to
858 metagenomes, virome assemblies here generally contained more viral contigs overall,
859 more binned viral metagenome-assembled genomes (vMAGs), higher species richness,
860 and greater read recruitment to vMAGs. Interestingly, marine metagenomes and viromes
861 were an exception. We observed no difference in the number of viral contigs, vMAGs, or
862 viral species richness. While this exception may be influenced by the smaller filter size
863 used to sample marine viromes compared to those used to generate the other datasets
864 used here, the environments from which marine samples were collected are perhaps the
865 most distinguished out of the others owing to their origin from hydrothermal vent plumes
866 in the deep sea. These environments have been shown to be composed of ultra-small
867 microbes that can have cell sizes <0.22 μm [69,70,76–78], requiring modified virome
868 extraction methods, just as the other environmental sources here require in other respects
869 [64,65].
870
871 We suggest that the lack of significant differences between the absolute viral genome
872 counts in the marine viromes and metagenomes here are likely due to the slow- growing,
873 low-biomass nature of deep-sea water column [79]. It is presumed that lysogeny or
874 chronic infection is a more common viral replication strategy in the deep-sea rather than
875 lysis [80–82], and a recent systemic analysis of marine viruses indeed found that more
876 than half of prokaryotic genomes in the deep sea represented lysogens [83]. While we
877 observed no difference the proportion of lytic and integrated vMAGs, there was a
878 significant increase in the proportion of lysogenic vMAGs in marine metagenomes
879 compared to viromes. Overall, while viromes generally yield more viral genomes in an
880 assembly and offer greater coverage, this trend may not be the case in all environments,
881 especially low-biomass and slow-growing environments that are currently understudied
882 such as the deep sea.
883

Our observation that more viral genomes were detected in viromes than metagenomes is not surprising given previous reports of viromes yielding more viral genomes than metagenomes [23]. Here, we show that viromes detect a considerable number of viral genomes not detected in metagenomes across all four environments examined. We also found a smaller but still noteworthy amount of vMAGs detected in metagenomes but not viromes. Overall, freshwater and marine assemblies each had a greater proportion of vMAGs unique to either viromes or metagenomes than human gut and soil assemblies. On the other hand, only a small fraction of human gut vMAGs were found only in metagenomes, and nearly all vMAGs present in soil metagenomes were also present in viromes. The degree of separation between viromes and metagenomes in terms of vMAG presence/absence therefore seems to vary by environment.

Interestingly, we also observed that an even larger number of viral genomes can be detected in both viromes and metagenomes than can be detected only in either method on their own. Soil assemblies were an exception with only one vMAG being detected in metagenomes and not viromes, yet so few viral genomes were found at all in soil metagenomes. Our analyses only included vMAGs with a genome size of at least 10 kb, viral sequences <10 kb that were not binned into a vMAG were not considered in the presence/absence analysis. It is possible that the degree of overlap and separation between vMAGs detected in viromes or metagenomes may be different when considering smaller and much more fragmented viral sequences. However, including viral sequences at such a low quality is not practical for functional, population, or community-level analyses given the high probability of false-positives [20,21,45]. Overall, while we find a significant amount of vMAGs detected only in viromes, and to a lesser extent metagenomes too, most vMAGs were still detected in both viromes and metagenomes.

The freshwater and soil datasets analyzed here were originally generated by studies that in addition to raising ecological questions about their respective ecosystems, directly compared the results they gained from the two sampling methods [6,29]. The freshwater study concluded that viromes contained a greater proportion of viral contigs than metagenomes and a greater absolute count of viral genomes (vMAGs) in viromes [6]. Similarly, the soil study found that viromes contained much more viral contigs and viral species (vOTUs) than metagenomes [29]. While the exact counts and means of the freshwater and soil datasets here differed from those in the original studies, likely due to differences in analytical methods and/or thresholds for retaining viral genomes prior to analysis, we observed the same enrichment in viral contigs, vMAGs, and viral species in virome assemblies compared to metagenomes.

Genome-level impacts

While there have been a handful of studies in the past that have examined viral community data resulting from viromes in comparison to metagenomes [6,11,29,84,85], even fewer have taken a closer look at specific genome-level differences that result across the two methods. Here, we examined the effects that choice of extraction method has on resulting viral genome assemblies. While we only focused on one viral species in this context, we found that a virome assembly resulted in a more-complete viral genome

than the genome assembled from a metagenome of the same sample. Both virome-assembled and metagenome-assembled viral genomes had orders of magnitude higher depth by virome reads that metagenome reads. Furthermore, the metagenome sample contained reads that mapped over the entire length of the complete version of the genome. Overall, although some viral genomes may be incompletely assembled in metagenomes, their full sequences can still be recovered if the metagenome reads were mapped to a higher quality virome assembly or reference genome.

Gene-level impacts

Viral genomes are very heterogeneous and mosaic, and viruses overall exhibit significant intra-population variation on the nucleotide level as well as on a gene-by-gene basis [4,86–91]. Consequently, if the overall viral community composition inferred from viromes is different than that of metagenomes, we reasoned that we would also see a difference on the protein-coding gene-level. Our results show that one's choice of extraction method does indeed influence gene annotations, but the significance and magnitude of differences varies between environments.

Out of all dereplicated protein-coding viral genes, 22% were differentially abundant (DA) between viromes and metagenomes. However, the vast majority of these DA genes were from freshwater assemblies alone, which is consistent with the larger size of the freshwater database compared to the other environments. Overall, freshwater and marine assemblies contained the greatest number of genes enriched in one method or the other (97% of freshwater DA genes, 68% of marine DA genes). Among the small fraction of soil genes that were DA, 78% were enriched in either method, while no human gut genes at all were enriched in viromes or metagenomes. We find this discrepancy between freshwater/marine assemblies and gut/soil assemblies to be consistent with the presence/absence of vMAGs in these environments. We first saw that freshwater and marine assemblies contained more vMAGs unique to viromes or metagenomes than human gut and soil assemblies. Likewise, we have found freshwater and marine assemblies to contain the most differentially abundant genes. In summary, observed community-level differences in vMAG abundance between viromes and metagenomes lead to differences in gene abundance as well.

Whether the purpose is for assigning taxonomy [92], revealing mechanisms to avoid host defenses [93], identifying auxiliary metabolic genes [94], or investigating mobile reservoirs for antimicrobial resistance genes [95,96], obtaining functional gene predictions is a critical step in analyses of viral communities. However, it can be quite challenging to assign functional predictions to viral genes annotated from metagenomic environmental data due to incredible diversity and under-characterization of viruses. Thus, annotating genes in complex viral communities often reveals a substantial amount of viral "dark matter" represented as genes with no known function that encode "hypothetical" proteins [25,97,98]. This challenge was indeed present here, as we could obtain functional predictions for only 16% of genes enriched in viromes or metagenomes. Nonetheless, we identified several functional categories across the three environments with significant genes. With this, our next aim was to assess whether enrichment of genes in either

viromes or metagenomes led to enrichments in any particular function across sample methods.

Hypergeometric testing is an effective way to infer functional enrichment from differentially abundant genes across groups [99–101], and it has been applied to the study of viral communities as well [84,102]. Our hypergeometric test results revealed several gene functional categories that were over and underrepresented in viromes and metagenomes. We found an overrepresentation of integration and excision genes in freshwater and marine metagenomes with an underrepresentation in freshwater viromes. However, lysis genes were underrepresented in freshwater metagenomes and overrepresented in freshwater viromes. On the one hand, this is consistent with our observations that freshwater metagenomes contained a greater proportion of integrated prophage vMAGs compared to viromes, and that marine metagenomes contained more lysogenic vMAGs compared to viromes. On the other hand, this contrasts our observation that there was no difference in the proportion of lysogenic vMAGs between freshwater viromes and metagenomes. Likewise, there was no significant difference between the proportion of integrated prophage vMAGs between marine viromes and metagenomes.

Together, our hypergeometric test results suggests that differences in replicative strategy of viruses may be behind the differential abundance of viral genes between viromes and metagenomes. We also found that genes encoding structural proteins were generally underrepresented in metagenomes and overrepresented in viromes. This may be the result of prophage domestication; this phenomenon leads to an unbalanced deletion of lysis and structural genes in viruses [103]. However, temperate viruses still require structural proteins to assemble upon lysis [104,105], and the enrichment of structural proteins was still different in soil assemblies even though we strangely found more prophage in soil viromes than metagenomes. We suggest that replication may not be the only reason behind the differential abundance of viral genes between viromes and metagenomes. There may exist additional processes that drive gene and genome-level differences that remain unexplored. Consequently, the choice between viromes and metagenomes can significantly influence one's interpretation of viral communities based on gene annotations.

CONCLUSIONS

In many contexts, viromes revealed more viral diversity than metagenomes at a species and genome-level scale. Because of this, extracting viromes may be more advantageous than metagenomes when studying viral communities in several contexts (Table 3). However, this may not be the case in all contexts (Table 3). Marine viromes did not prove superior to metagenomes as we observed no difference in the number of viral genomes or viral species between to the two sampling methods. Enriching for virus-like particles with virome extractions may not be able to provide a broader view of viral communities than mixed-community metagenomics in low-biomass, slow growing environments such as the understudied deep sea. If one has reason to believe that a viral community may largely be composted of virocells or viruses undergoing chronic infection, perhaps the benefits of mixed-community metagenomics such as retaining host context are greater

than the potential benefits of extracting viromes, which may not be gained. Overall, our results here show that one's choice of method can lead to differing interpretations of viral community ecology.

Nonetheless, it is often optimal to apply both extractions methods (Table 3). While a substantial number of viruses were detected only in viromes, there was a noticeable number of viruses still only detected in metagenomes in all four environments tested here. That these viral genomes were only detected in the mixed community and not in a virus-like particle sized fraction could very well be ecologically relevant. These viruses may offer insights into viral diversity that can go unseen in viromes alone. Furthermore, since the natural history, ecology, and evolution of viruses are intimately tied to that of their hosts, obtaining host context with metagenomes can maximize the utility of the viral genomes assembled from viromes (Table 3). In conclusion, we recommend that researchers investigating viral communities should extract both viromes and mixed-community metagenomes in pairs from the same biological samples. However, if one is restricted to using just one method, viromes will likely be the better option for virus-focused studies of most environments.

Table 3. Recommendations for choosing extraction methods depending on research context.

Context	Recommended method(s)	Rationale
Viral community dynamics, overall virus diversity, assembly of uncultivatable virus genomes	Virome	Viromes generally contained more viral species and greater viral sequence enrichment than metagenomes.
Bacterial/archaeal communities, no interest in viruses	Metagenomes	Viromes are unnecessary to the study of just the cellular members of communities.
Fast-growing, highly dynamic communities, and/or lytic viruses	Virome	Assuming viral lysis is prevalent due to the present biotic or abiotic conditions, viromes will enrich for lytic viruses.
Slow-growing, low-biomass communities, and/or integrated viruses	Metagenomes	Assuming lysogeny is prevalent due to the present biotic or abiotic conditions, detecting viruses integrated in the host genome require metagenomics.
Host-virus interactions	Paired viromes & metagenomes	Metagenomes are necessary to provide any host context. While metagenomes alone can yield some viral genomes, viromes are also recommended to maximize viral genome assembly.
Maximization of total virus diversity	Paired viromes & metagenomes	Both viromes and metagenomes resulted in the assembly of viral genomes not detected in the other method. Utilizing both methods can maximize the detection and assembly of as many viral genomes as possible.

ABBREVIATIONS

VLP: Virus-like particle

ANI: Average nucleotide identity
PEG: Polyethylene glycols
vMAG: Viral metagenome-assembled genome
AKC: Amended 1% potassium citrate
vOTU: Viral operational taxonomic unit
EPR: East Pacific Rise
DA: Differentially abundant

DECLARATIONS

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the following repositories: Freshwater, originally presented by Tran et al. [6] and deposited to the JGI Genome Portal under Proposal ID 506328; Human gut, originally presented by Shkoporov et al. [11] and deposited to the NCBI Sequence Read Archive under BioProject accession PRJNA545408; Soil, originally presented by Santos-Medellin et al. [29] and deposited to the NCBI Sequence Read Archive under BioProject accession PRJNA646773.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM143024, and by the National Science Foundation under grant numbers DBI2047598 and OCE2049478.

Authors' contributions

Conceptualization, J.C.K. and K.A.; Methodology, J.C.K. and K.A.; Software, J.C.K.; Validation, J.C.K.; Formal Analysis, J.C.K.; Investigation, J.C.K., K.M.K., M.V.L., P.Q.T., and K.A.; Resources, K.A.; Data curation, J.C.K., K.M.K., M.V.L., P.Q.T., and K.A.; Writing — Original Draft, J.C.K.; Writing — Review & Editing, J.C.K., K.M.K., M.V.L., P.Q.T., and K.A.; Visualization, J.C.K.; Supervision, K.A.; Project Administration, K.A.; Funding Acquisition, K.A.

Acknowledgments

We are thankful to all authors of the studies that originally generated and distributed the data analyzed here. We also gratefully acknowledge the insights provided by Cody Martin during this study.

REFERENCES

1. Wommack KE, Colwell RR. Virioplankton: Viruses in Aquatic Ecosystems † [Internet]. MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS. 2000. Available from: https://journals.asm.org/journal/mmbr
2. Cesar Ignacio-Espinoza J, Solonenko SA, Sullivan MB. The global virome: Not as big as we thought? Curr Opin Virol. Elsevier B.V.; 2013. p. 566–71.
3. Stern A, Sorek R. The phage-host arms race: Shaping the evolution of microbes. BioEssays. 2011. p. 43–51.
4. Kosmopoulos JC, Campbell DE, Whitaker RJ, Wilbanks EG. Horizontal Gene Transfer and CRISPR Targeting Drive Phage-Bacterial Host Interactions and Coevolution in “Pink Berry” Marine Microbial Aggregates. Vives M, editor. Appl Environ Microbiol [Internet]. 2023;89. Available from: https://journals.asm.org/doi/10.1128/aem.00177-23
5. Santos-Medellín C, Blazewicz SJ, Pett-Ridge J, Emerson JB. Viral but not bacterial community succession is characterized by extreme turnover shortly after rewetting dry soils. bioRxiv. 2023;
6. Tran PQ, Bachand SC, Peterson B, He S, Anantharaman K. Viral impacts on microbial activity and biogeochemical cycling in a seasonally anoxic freshwater lake. bioRxiv [Internet]. 2023; Available from: https://doi.org/10.1101/2023.04.19.537559
7. Hurwitz BL, U’Ren JM. Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol. Elsevier Ltd; 2016. p. 161–8.
8. Kieft K, Zhou Z, Anderson RE, Buchan A, Campbell BJ, Hallam SJ, et al. Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat Commun. 2021;12.
9. Fujimoto K, Kimura Y, Shimohigoshi M, Satoh T, Sato S, Tremmel G, et al. Metagenome Data on Intestinal Phage-Bacteria Associations Aids the Development of Phage Therapy against Pathobionts. Cell Host Microbe. 2020;28:380-389.e9.
10. Gordillo Altamirano FL, Barr JJ. Phage Therapy in the Postantibiotic Era. Clin Microbiol Rev [Internet]. 2019;32. Available from: http://cmr.asm.org/
11. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, et al. The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host Microbe. 2019;26:527-541.e5.
12. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184:1098-1109.e9.
13. Shah SA, Deng L, Thorsen J, Pedersen AG, Dion MB, Castro-Mejía JL, et al. Expanding known viral diversity in the healthy infant gut. Nat Microbiol. 2023;8:986–98.
14. Paez-Espino D, Zhou J, Roux S, Nayfach S, Pavlopoulos GA, Schulz F, et al. Diversity, evolution, and classification of virophages uncovered through global metagenomics. Microbiome. 2019;7.
15. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell. 2019;177:1109-1123.e14.
16. Gaïa M, Meng L, Pelletier E, Forterre P, Vanni C, Fernandez-Guerra A, et al. Mirusviruses link herpesviruses to giant viruses. Nature [Internet]. 2023; Available from: https://www.nature.com/articles/s41586-023-05962-4

17. Hillary LS, Adriaenssens EM, Jones DL, McDonald JE. RNA-viromics reveals diverse communities of soil RNA viruses with the potential to affect grassland ecosystems across multiple trophic levels. ISME Communications. 2022;2.
18. Roux S, Emerson JB. Diversity in the soil virosphere: to infinity and beyond? Trends Microbiol. 2022;30:1025–35.
19. Roux S, Adriaenssens EM, Dutilh BE, Koonin E V., Kropinski AM, Krupovic M, et al. Minimum information about an uncultivated virus genome (MIUVIG). Nat Biotechnol. 2019;37:29–37.
20. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an *in silico* evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017;5:e3817.
21. Kieft K, Anantharaman K. Virus genomics: what is being overlooked? Curr Opin Virol. Elsevier B.V.; 2022.
22. Trubl G, Hyman P, Roux S, Abedon ST. Coming-of-Age Characterization of Soil Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics. Soil Syst. 2020;4:23.
23. Roux S, Matthijnssens J, Dutilh BE. Metagenomics in Virology. Encyclopedia of Virology. Elsevier; 2021. p. 133–40.
24. Roux S, Camargo AP, Coutinho FH, Dabdoub SM, Dutilh BE, Nayfach S, et al. iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. PLoS Biol. 2023;21.
25. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. Elife. 2015;4.
26. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, VerBerkmoes NC, et al. Twelve previously unknown phage genera are ubiquitous in global oceans. Proc Natl Acad Sci U S A. 2013;110:12798–803.
27. Pascoal F, Costa R, Magalhães C. The microbial rare biosphere: Current concepts, methods and ecological principles. FEMS Microbiol Ecol. Oxford University Press; 2021.
28. Garin-Fernandez A, Pereira-Flores E, Glöckner FO, Wichels A. The North Sea goes viral: Occurrence and distribution of North Sea bacteriophages. Mar Genomics. 2018;41:31–41.
29. Santos-Medellin C, Zinke LA, ter Horst AM, Gelardi DL, Parikh SJ, Emerson JB. Viromes outperform total metagenomes in revealing the spatiotemporal patterns of agricultural soil viral communities. ISME Journal. 2021;15:1956–70.
30. Forterre P. Manipulation of cellular syntheses and the nature of viruses: The virocell concept. Comptes Rendus Chimie. Elsevier Masson SAS; 2011. p. 392–9.
31. López-Pérez M, Haro-Moreno JM, Gonzalez-Serrano R, Parras-Moltó M, Rodriguez-Valera F. Genome diversity of marine phages recovered from Mediterranean metagenomes: Size matters. PLoS Genet. 2017;13.
32. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018;3:754–66.
33. Chen C, Yan Q, Yao X, Li S, Lv Q, Wang G, et al. Alterations of the gut virome in patients with systemic lupus erythematosus. Front Immunol. 2023;13.
34. Zaragoza-Solas A, Haro-Moreno JM, Rodriguez-Valera F, López-Pérez M. Long-Read Metagenomics Improves the Recovery of Viral Diversity from Complex Natural Marine Samples. mSystems. 2022;7.

35. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022;50:D20–6.
36. Clum A, Huntemann M, Bushnell B, Foster B, Foster B, Roux S, et al. DOE JGI Metagenome Workflow. Segata N, editor. mSystems [Internet]. 2021;6:D723–33. Available from: https://journals.asm.org/doi/10.1128/mSystems.00804-20
37. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 2017;27:824–34.
38. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2016;32:1088–90.
39. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2020. Available from: https://www.R-project.org/
40. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.
41. Zhou Z, Martin C, Kosmopoulos JC, Anantharaman K. ViWrap: A modular pipeline to identify, bin, classify, and predict viral–host relationships for viruses from metagenomes. iMeta [Internet]. 2023; Available from: https://onlinelibrary.wiley.com/doi/10.1002/imt2.118
42. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8:90.
43. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
44. Kieft K, Adams A, Salamzade R, Kalan L, Anantharaman K. vRhyme enables binning of viral genomes from metagenomes. Nucleic Acids Res. 2022;50:e83–e83.
45. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39:578–85.
46. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019;37:632–9.
47. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–8.
48. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44:D733–45.
49. Grazziotin AL, Koonin E V., Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res. 2017;45:D491–8.
50. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
51. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10.
52. Wilkinson L. Exact and Approximate Area-Proportional Circular Venn and Euler Diagrams. IEEE Trans Vis Comput Graph. 2012;18:321–31.

53. Micallef L, Rodgers P. eulerAPE: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses. PLoS One. 2014;9:e101717.
54. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D. vegan: Community Ecology Package. 2018;
55. Roux S, Matthijnssens J, Dutilh BE. Metagenomics in Virology. Encyclopedia of Virology. Elsevier; 2021. p. 133–40.
56. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 2004;14:1394–403.
57. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
58. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
59. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9:2542.
60. Van Dongen S. Graph Clustering Via a Discrete Uncoupling Process. SIAM Journal on Matrix Analysis and Applications. 2008;30:121–41.
61. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
62. Terzian P, Olo Ndela E, Galiez C, Lossouarn J, Pérez Bucio RE, Mom R, et al. PHROG: families of prokaryotic virus proteins clustered using remote homology. NAR Genom Bioinform. 2021;3.
63. Bouras G, Nepal R, Houtak G, Psaltis AJ, Wormald P-J, Vreugde S. Pharokka: a fast scalable bacteriophage annotation tool. Bioinformatics. 2023;39.
64. Wommack KE, Williamson KE, Helton RR, Bench SR, Winget DM. Methods for the Isolation of Viruses from Environmental Samples. 2009. p. 3–14.
65. Shkoporov AN, Ryan FJ, Draper LA, Forde A, Stockdale SR, Daly KM, et al. Reproducible protocols for metagenomic analysis of human faecal phageomes. Microbiome. 2018;6:68.
66. Kieft K, Adams A, Salamzade R, Kalan L, Anantharaman K. VRhyme enables binning of viral genomes from metagenomes. Nucleic Acids Res. 2022;50.
67. Roux S, Adriaenssens EM, Dutilh BE, Koonin E V., Kropinski AM, Krupovic M, et al. Minimum information about an uncultivated virus genome (MIUVIG). Nat Biotechnol. 2019;37:29–37.
68. Johansen J, Plichta DR, Nissen JN, Jespersen ML, Shah SA, Deng L, et al. Genome binning of viral entities from bulk metagenomics data. Nat Commun. 2022;13.
69. Kuppa Baskaran DK, Umale S, Zhou Z, Raman K, Anantharaman K. Metagenome-based metabolic modelling predicts unique microbial interactions in deep-sea hydrothermal plume microbiomes. ISME Communications [Internet]. 2023;3:42. Available from: https://www.nature.com/articles/s43705-023-00242-8
70. Zhou Z, St. John E, Anantharaman K, Reysenbach AL. Global patterns of diversity and metabolism of microbial communities in deep-sea hydrothermal vent deposits. Microbiome. 2022;10.
71. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Patterns and ecological drivers of ocean viral communities. Science (1979). 2015;348.

72. Hurwitz BL, Sullivan MB. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. PLoS One. 2013;8:e57355.
73. Brum JR, Ignacio-Espinoza JC, Kim E-H, Trubl G, Jones RM, Roux S, et al. Illuminating structural proteins in viral “dark matter” with metaproteomics. Proceedings of the National Academy of Sciences. 2016;113:2436–41.
74. Camargo AP, Nayfach S, Chen IMA, Palaniappan K, Ratner A, Chu K, et al. IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res. 2023;51:D733–43.
75. Sunagawa S, Acinas SG, Bork P, Bowler C, Babin M, Boss E, et al. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol. Nature Research; 2020. p. 428–45.
76. Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Sci Rep. 2013;3:2471.
77. He C, Keren R, Whittaker ML, Farag IF, Doudna JA, Cate JHD, et al. Genome-resolved metagenomics reveals site-specific diversity of episymbiotic CPR bacteria and DPANN archaea in groundwater ecosystems. Nat Microbiol. 2021;6:354–65.
78. Chaudhari NM, Overholt WA, Figueroa-Gonzalez PA, Taubert M, Bornemann TL V., Probst AJ, et al. The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers. Environ Microbiome. 2021;16:24.
79. Lutz RA, Kennish MJ. Ecology of deep-sea hydrothermal vent communities: A review. Reviews of Geophysics. 1993;31:211.
80. Williamson SJ, Cary SC, Williamson KE, Helton RR, Bench SR, Winget D, et al. Lysogenic virus–host interactions predominate at deep-sea diffuse-flow hydrothermal vents. ISME J. 2008;2:1112–21.
81. Cai L, Weinbauer MG, Xie L, Zhang R. The smallest in the deepest: the enigmatic role of viruses in the deep biosphere. Natl Sci Rev. 2023;10.
82. Correa AMS, Howard-Varona C, Coy SR, Buchan A, Sullivan MB, Weitz JS. Revisiting the rules of life for viruses of microorganisms. Nat Rev Microbiol. 2021;19:501–13.
83. Yi Y, Liu S, Hao Y, Sun Q, Lei X, Wang Y, et al. A systematic analysis of marine lysogens and proviruses. Nat Commun [Internet]. 2023;14:6013. Available from: https://www.nature.com/articles/s41467-023-41699-4
84. Santos-Medellín C, Blazewicz SJ, Pett-Ridge J, Firestone MK, Emerson JB. Viral but not bacterial community successional patterns reflect extreme turnover shortly after rewetting dry soils. Nat Ecol Evol. 2023;
85. ter Horst AM, Santos-Medellín C, Sorensen JW, Zinke LA, Wilson RM, Johnston ER, et al. Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. Microbiome. 2021;9.
86. Ramos-Barbero MD, Gómez-Gómez C, Sala-Comorera L, Rodríguez-Rubio L, Morales-Cortes S, Mendoza-Barberá E, et al. Characterization of crAss-like phage isolates highlights Crassvirales genetic heterogeneity and worldwide distribution. Nat Commun. 2023;14:4295.

87. Gao R, Naushad S, Moineau S, Levesque R, Goodridge L, Ogunremi D. Comparative genomic analysis of 142 bacteriophages infecting Salmonella enterica subsp. enterica. BMC Genomics. 2020;21:374.
88. Pope WH, Bowman CA, Russell DA, Jacobs-Sera D, Asai DJ, Cresawn SG, et al. Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. Elife. 2015;4.
89. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of Highly Mosaic Mycobacteriophage Genomes. Cell. 2003;113:171–82.
90. Grose JH, Casjens SR. Understanding the enormous diversity of bacteriophages: The tailed phages that infect the bacterial family Enterobacteriaceae. Virology. 2014;468–470:421–43.
91. Mavrich TN, Hatfull GF. Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol. 2017;2:17112.
92. Moreno-Gallego JL, Reyes A. Informative Regions In Viral Genomes. Viruses. 2021;13:1164.
93. Gao Z, Feng Y. Bacteriophage strategies for overcoming host antiviral immunity. Front Microbiol. 2023;14.
94. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. 2020;48:8883–900.
95. Moon K, Jeon JH, Kang I, Park KS, Lee K, Cha C-J, et al. Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. Microbiome. 2020;8:75.
96. Strange JES, Leekitcharoenphon P, Møller FD, Aarestrup FM. Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage. Sci Rep. 2021;11:1600.
97. Hurwitz BL, Sullivan MB. The Pacific Ocean Virome (POV): A Marine Viral Metagenomic Dataset and Associated Protein Clusters for Quantitative Viral Ecology. PLoS One. 2013;8:e57355.
98. Brum JR, Ignacio-Espinoza JC, Kim EH, Trubl G, Jones RM, Roux S, et al. Illuminating structural proteins in viral “dark matter” with metaproteomics. Proc Natl Acad Sci U S A. 2016;113:2436–41.
99. Park J, Taslim C, Lin S. BOG: R-package for Bacterium and virus analysis of Orthologous Groups. Comput Struct Biotechnol J. 2015;13:366–9.
100. Cao J, Zhang S. A Bayesian extension of the hypergeometric test for functional enrichment analysis. Biometrics. 2014;70:84–94.
101. Zhou X, Su Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. BMC Genomics. 2007;8:246.
102. Santos-Medell In C, Estera-Molina K, Yuan M, Pett-Ridge J, Firestone MK, Emerson JB. Spatial turnover of soil viral populations and genotypes overlain by cohesive responses to moisture in grasslands. 2022; Available from: https://doi.org/10.1073/pnas.2209132119
103. Bobay LM, Touchon M, Rocha EPC. Pervasive domestication of defective prophages by bacteria. Proc Natl Acad Sci U S A. 2014;111:12127–32.
104. Birge EA. Genetics of temperate bacteriophages. Bacterial and bacteriophage genetics. 2006. p. 261–303.

1363 105. Owen S V., Canals R, Wenner N, Hammarlöf DL, Kröger C, Hinton JCD. A window
1364 into lysogeny: revealing temperate phage biology with transcriptomics. Microb Genom.
1365 2020;6.
1366
1367