frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fmicb.2022.953500

OPEN ACCESS

EDITED BY
Karthik Anantharaman,
University of Wisconsin-Madison,
United States

REVIEWED BY
Niti B. Jadeja,
Ashoka Trust for Research in Ecology and the Environment, India
Manuel Martinez Garcia,
University of Alicante,
Spain

*CORRESPONDENCE
Ali H. A. Elbehery
ali.elbehery@fop.usc.edu.eg
Li Deng
li.deng@helmholtz-muenchen.de

SPECIALTY SECTION
This article was submitted to Phage Biology, a section of the journal Frontiers in Microbiology

RECEIVED 26 May 2022
ACCEPTED 11 August 2022
PUBLISHED 28 September 2022

CITATION
Elbehery AHA and Deng L (2022) Insights into the global freshwater virome. *Front. Microbiol.* 13:953500. doi: 10.3389/fmicb.2022.953500

Insights into the global freshwater virome

Ali H. A. Elbehery[1]* and Li Deng[2,3]*

[1]Department of Microbiology and Immunology, Faculty of Pharmacy, University of Sadat City, Sadat City, Egypt, [2]Helmholtz Centre Munich – German Research Centre for Environmental Health, Institute of Virology, Neuherberg, Germany, [3]Chair of Microbial Disease Prevention, School of Life Sciences, Technical University of Munich, Freising, Germany

Viruses are by far the most abundant life forms on this planet. Yet, the full viral diversity remains mostly unknown, especially in environments like freshwater. Therefore, we aimed to study freshwater viruses in a global context. To this end, we downloaded 380 publicly available viral metagenomes (>1TB). More than 60% of these metagenomes were discarded based on their levels of cellular contamination assessed by ribosomal DNA content. For the remaining metagenomes, assembled contigs were decontaminated using two consecutive steps, eventually yielding 273,365 viral contigs longer than 1,000 bp. Long enough contigs (≥ 10 kb) were clustered to identify novel genomes/genome fragments. We could recover 549 complete circular and high-quality draft genomes, out of which 10 were recognized as being novel. Functional annotation of these genomes showed that most of the annotated coding sequences are DNA metabolic genes or phage structural genes. On the other hand, taxonomic analysis of viral contigs showed that most of the assigned contigs belonged to the order *Caudovirales*, particularly the families of *Siphoviridae*, *Myoviridae,* and *Podoviridae*. The recovered viral contigs contained several auxiliary metabolic genes belonging to several metabolic pathways, especially carbohydrate and amino acid metabolism in addition to photosynthesis as well as hydrocarbon degradation and antibiotic resistance. Overall, we present here a set of prudently chosen viral contigs, which should not only help better understanding of freshwater viruses but also be a valuable resource for future virome studies.

KEYWORDS

freshwater, virome, metagenome, bacteriophages, auxiliary metabolic genes

Introduction

Viruses are the most numerous biological entities on Earth, with a global estimate of 4.80×10^{31} virus like particles (VLPs; Güemes et al., 2016). Freshwater is estimated to contain 1.76×10^{27} VLPs, which represents 0.0037% of VLPs on the globe. This number of VLPs in freshwater outnumbers the total number of prokaryotes in freshwater by 14 times (Güemes et al., 2016). Most of these viruses are viruses which infect prokaryotes (Edwards and Rohwer, 2005). Based on the mass of an average phage with a 50 kb genome and an icosahedral capsid, which is 0.0823 femtograms (Güemes et al., 2016), the total mass of

viruses in freshwater could be estimated to 144,848 tons, which is more than the weight of 1,800 blue whales; the weight of an average mature female blue whale is 79 tons (Krogh, 1934). In comparison, marine ecosystems are estimated to contain 1.29×10^{30} VLPs (Güemes et al., 2016), which is almost three orders of magnitude higher than VLPs in freshwater. Yet, viral communities are distinct between these two ecosystems (Logares et al., 2009; Roux et al., 2012), denoting biome-specific diversity.

Phages play important ecological roles. For example, phages, through lysis of bacteria, contribute to carbon cycling, where dissolved organic matter released from lysed bacterial cells become available to other bacteria in a process known as viral shunt (Wilhelm and Suttle, 1999). In addition, the ability of phages to kill bacteria allow them to regulate and control the size and diversity of microbial communities. One of the most popular models by which viruses are believed to control microbial populations, is "killing the winner," where viruses selectively kill abundant bacterial taxa, allowing rare taxa to grow and bringing back balance between bacteria taxa in the ecosystem (Thingstad and Lignell, 1997). Such model was previously reported in studies concerned with freshwater viruses, where bacterial diversity increased with increases in viral abundance (Auguet et al., 2009; Meunier and Jacquet, 2015). Moreover, viruses contribute to the functional diversity of microbial communities through gene transfer, which occurs through transduction, where chromosomal or plasmid DNA can be transferred among bacteria by means of viruses (Morrison et al., 1978; Ripp et al., 1994; Replicon et al., 1995; Kenzaka et al., 2010). Interestingly, many phages own host genes that are expressed during infection to overcome host metabolic bottlenecks and enhance viral production. These genes are known as auxiliary metabolic genes (AMGs; Thompson et al., 2011). AMGs span a wide range of metabolic processes, including photosynthesis (Ruiz-Perez et al., 2019), carbon metabolism (Hurwitz et al., 2013), nucleic acid synthesis (Chevallereau et al., 2016), nitrogen metabolism (Gazitúa et al., 2020), sulfur metabolism (Mara et al., 2020) and other metabolic pathways (Warwick et al., 2019). Although most virome studies have been concerned with marine environments, there is an increasing effort to study viruses in freshwater (Jacquet et al., 2010), assessing viral communities in diverse freshwater biomes. They addressed different aspects, such as the prevalence of *Podoviridae* and *Siphoviridae* in the largest lake in Ireland (Skvortsov et al., 2016), the characteristics and prevalence of particular phage, infecting *Fonsibacter*, a bacterioplankton abundant in freshwater ecosystems (Chen et al., 2019), new roles of freshwater viruses in carbon fixation and methylotrophy (Coutinho et al., 2020). Yet, freshwater viruses diversity and function are completely far from being fully known.

In this study, we aimed to collect publicly available freshwater viral metagenomes from all across the globe to get a better understanding of viruses in freshwater. We assembled sequence reads and were able to decontaminate them, relying on both sequence similarity and a machine learning-based tool, thus eventually obtaining viral contigs. We studied the abundance, novelty, function and taxonomy of these contigs in an effort to improve our knowledge of viruses in this virtually untapped environment.

Materials and methods

Downloading publicly available metagenomes

Freshwater viral metagenomes were downloaded from the National Center for Bioinformatics Information (NCBI) Sequence Read Archive (SRA)[1] using SRAdb R package (Zhu et al., 2013) on March 1, 2018, while search_terms = freshwater AND (virus OR viral OR virome OR phage OR phageome) AND (metagenome OR "metagenomic"). These search criteria resulted in 1474 records, but visual inspection of metadata for non-specific results reduced the number of viral metagenomes to 380 (Supplementary Table S1).

Quality control of sequence reads

Quality control of viral metagenomes was performed by first removing adapters using Cutadapt v.1.16 (Martin, 2011). Cutadapt was also used for N-end trimming (--trim-n) and quality trimming from both ends to a Phred score of at least 15 (--q 15,15). Moreover, all reads with more than two ambiguous nucleotides (N) and/or shorter than 50 nucleotides were removed (--max-n 2, −m 50). Prinseq v.0.20.4 (Schmieder and Edwards, 2011b) was used for the removal of low complexity sequences (--lc_threshold 50, --lc_method entropy), dereplication (--derep 12345) and filtering sequences with noniupac characters and/or with an average Phred score lower than 20 (--noniupac, --min_qual_mean 20). Metagenomes were then filtered from sequences matching PhiX 174 bacteriophage [a known control in Illumina sequencing and a potential contaminant (Mukherjee et al., 2015)], if any, using Deconseq v.0.4.3 (Schmieder and Edwards, 2011a) using default parameters.

Sequence reads assembly

Illumina reads that passed quality control were assembled, each sample separately, using megahit v1.1.1 (Liu et al., 2015) with the parameters (-t 20 -m 0.8 --preset meta-sensitive), whereas 454 reads were assembled using Newbler v.2.9 (Roche) using default parameters, since Newbler has better performance than other assemblers for Roche 454 reads (Kumar and Blaxter, 2010). Produced contigs were renamed according to the following pattern: freshwater_SRRAccession.contig000XXXXXX, where SRRAccession is the SRA Run accession number of the particular

1 https://www.ncbi.nlm.nih.gov/sra

sample used for assembly and XXXXXX is a serial number. Metagenomes that produced no contigs longer than or equal to 1,000 bp were removed from further analysis.

Assessing and removing cellular contamination

Contamination of viral metagenomes (viromes) with cellular sequences is unavoidable (Roux et al., 2013). A diagram explaining the assessment and removal of cellular contamination is shown in Supplementary Figure S1. We assessed the level of cellular contamination by detection of ribosomal DNA reads using meta_rna (Huang et al., 2009), which reports reads matching 5S, 16S and 23S ribosomal genes. We also sought to remove cellular-like sequences without false positive removal of true viral sequences. Hence, we downloaded Refseq Archaea, Bacteria, Fungi and Protozoa on April 19, 2018 (release March 12, 2018).[2] We also downloaded Refseq and non-Refseq prokaryotic viruses from NCBI as previously described (Grazziotin et al., 2017; Elbehery et al., 2018). Moreover, we predicted prophages from RefSeq bacterial and archaeal genomes (complete genomes only; 9,355 bacterial and 283 archaeal genomes) using PhiSpy v.2.3 with default parameters (Akhter et al., 2012). Predicted bacterial and archaeal prophages were shredded using the Shred tool from the BBTools suite [Joint Genome Institute (JGI)][3] with a fragment length of 80, an overlap of 40 and a minimum length of 70 nucleotides. These fragments were mapped to RefSeq bacterial sequences using BBMap (BBTools, JGI)[4] with a minimum identity of 85% and maximum indels of 2. Mapping files created from this step were used to mask viral matches in bacterial sequences. Masking of viral matches in addition to low complexity regions was carried out using BBMask (BBTools, JGI) with an entropoy threshold of 0.7 as suggested in BBMask Guide.[5] The same was done for RefSeq archaeal sequences to mask matches for archaeal and bacterial viruses and prophages. All 347 virome datasets—producing 1,000 bp or longer contigs—were mapped to the RefSeq sequences (bacterial and archaeal sequences, masked for viral matches, in addition to fungal and protozoal sequences) and the proportion of bases mapping to cellular sequences was calculated.

In their assessment of the influence of cellular contamination of viromes, Roux and colleagues (Roux et al., 2013) found that the highest ratio of rDNA in 67 published viromes was 5.3‰, which was considered non-negligible extent of contamination. Therefore, we considered any virome with a ratio of rDNA greater than 5‰ to be excessively contaminated and was removed from further consideration. We also combined rDNA ratio with another strongly correlated measure, the ratio of reads mapped to potential cellular contaminants. Hence, viromes with this ratio exceeding 2% were also excluded. The remaining viromes (143) had their contigs cleaned up from cellular sequences. To clean up contigs, clean reads of each virome were mapped to corresponding contigs and those contigs showing less than 95% coverage were discarded. Dereplication of contigs using cd-hit-est v4.7 (Li and Godzik, 2006) was done to obtain a non-redundant set of contigs (options: -c 0.95 -n 10 -d 0). Generally, cd-hit keeps the longest contig when redundancy occurs. The non-redundant contigs were processed using VIBRANT (Virus Identification By iteRative ANnoTation; Kieft et al., 2020) v1.2.0, a novel tool that identifies viral contigs by combining a supervised machine learning approach with viral scoring obtained by annotations using Pfam, Virus Orthologous Groups (VOG) and Kyoto Encyclopedia of Genes and Genomes (KEGG). VIBRANT was used with the default settings.

Viral clustering

Contigs longer than or equal to 10 kbp, identified by VIBRANT as being viral were clustered into viral operational taxonomic units (vOTUs) using cd-hit-est v4.7 (Li and Godzik, 2006) according to the guidelines suggested by Roux and colleagues (Roux et al., 2019), which suggested using 95% identity (also referred to as average nucleotide identity, ANI) and 85% coverage (also referred to as alignment fraction, AF) of the shorter sequence (options: -c 0.95 -n 10 -d 0 -G 0 – aS 0.85).

We sought to identify which of these vOTUs could belong to potentially novel viral genera/subfamilies. To this end, we downloaded Integrated Microbial Genome/Virus (IMG/VR v.2.0[6]; Paez-Espino et al., 2018) database on July 2, 2019. IMG/VR database is the largest and most comprehensive database of viral genome sequences. It was composed of 760,453 viral genomic fragments. We clustered VIBRANT vOTU representatives together with sequences from IMG/VR database into viral clusters (VCs) of genus-rank using vConTACT2 (v.0.9.19; Bin Jang et al., 2019). Since the number of sequences was huge, we had to split IMG/VR database into 39 files, adding VIBRANT vOTU representatives to each of these files and running vConTACT2, separately for each of these combinations (options: --rel-mode Diamond --db None --pcs-mode MCL --vcs-mode ClusterONE --threads 28). The 40th vConTACT2 run was similarly done, but using VIBRANT vOTU representatives alone, while vConTACT2 database was set to Refseq - release 201 (released on July 22, 2020) prokaryotic viruses (vConTACT2 option --db ProkaryoticViralRefSeq201-Merged). Generated VCs were merged by adding members of these clusters to a NetworkX (v.2.4; Hagberg et al., 2008) nondirected graph then calling upon the graph's connected components using a custom Python script. vConTACT2 (Bin Jang et al., 2019) clusters viral genomes into viral clusters (VCs) approximately equivalent to genus

2 ftp://ftp.ncbi.nlm.nih.gov/refseq/

3 https://jgi.doe.gov/data-and-tools/bbtools/

4 https://sourceforge.net/projects/bbmap/

5 https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbmask-guide/

6 https://genome.jgi.doe.gov/portal/IMG_VR/IMG_VR.home.html

level as defined by the International Committee on Taxonomy of Viruses (ICTV). vConTACT2 assigns genomes different statuses based on their clustering according to their shared gene content. Genomes can be referred to as (a) clustered, when clustering has occurred with high-confidence, (b) singleton, when they have few or no shared genes, (c) overlap when they have shared genes with genomes from more than one VC, or (d) outlier, when they have shared genes with other genomes, but not enough for high-confidence clustering into genus-level VC.

Genome quality and completeness

Genome quality and completeness were obtained using VIBRANT v1.2.0, which relies on four different parameters for genome completeness: (a) the contig/scaffold being circular, (b) VOG annotations, (c) total VOG nucleotide replication proteins, and (d) total VOG viral hallmark proteins (Kieft et al., 2020).

Host prediction for viral genomes

Host prediction was done using CrisprOpenDB tool, which relies on matches between viral genomes and an extended database of clustered regularly interspaced palindromic repeats (CRISPR) spacers, comprising more than 11 million spacers (Dion et al., 2021). The tool was run with a maximum number of mismatches of two. Taxonomic lineages of the predicted hosts were extracted from NCBI taxonomy (Schoch et al., 2020), with the help of ncbitax2lin tool v2.3.2 (Xue, 2022).

Abundance of selected genomes/genome fragments

Clean reads pertaining to each freshwater biome were mapped to the selected 19,210 genomes/genome fragments longer than 10 kbp using BBMap (BBTools, JGI)[7] with minid = 0.9, i.e., minimum identity set to 90%. Abundance was calculated based on the number of reads mapping to each genome/genome fragment and normalized to the total number of reads of each freshwater biome as well as contig length; so, relative abundance was expressed as number of reads per million reads per million bases. Intersections between biomes based on genomes/genome fragments present in each biome was calculated by combining genomes/genome fragments with nonzero abundances in each biome into a list of sets in R v.4.0.2 (R Development Core Team, 2020). Intersections between biomes based on shared genomes/genome fragments with nonzero abundances, were plotted using upset function of UpSetR package v.1.4.0 (nsets = 8, order.by = "freq," nintersects = 50, all other arguments set to default; Conway et al., 2017). Phylogenomic analysis of the contigs with nonzero abundance among all biomes was carried out by the VICTOR web service,[8] a method for the genome-based phylogeny and classification of prokaryotic viruses (Meier-Kolthoff and Göker, 2017). All pairwise comparisons of the nucleotide sequences were conducted using the Genome-BLAST Distance Phylogeny (GBDP) method (Meier-Kolthoff et al., 2013) under settings recommended for prokaryotic viruses (Meier-Kolthoff and Göker, 2017). The resulting intergenomic distances were used to infer a balanced minimum evolution tree with branch support *via* FASTME including subtree pruning and regrafting (SPR) postprocessing (Lefort et al., 2015) for the formula D4, which was selected because of its robustness for incomplete genomic sequences (Meier-Kolthoff and Göker, 2017). Branch support was inferred from 100 pseudo-bootstrap replicates each. The tree was rooted at the midpoint (Farris, 1972) and visualized with ggtree (Yu, 2020). Taxon boundaries at the species, genus and family levels were estimated with the OPTSIL program (Göker et al., 2009), the recommended clustering thresholds (Meier-Kolthoff and Göker, 2017) and an *F* value (fraction of links required for cluster fusion) of 0.5 (Meier-Kolthoff et al., 2014).

Genome annotations and detection of AMGs

Genome annotation was done using VIBRANT (Kieft et al., 2020), which integrates annotations from KEGG, Pfam, and VOG databases. Detection of AMGs was also done using VIBRANT (Kieft et al., 2020), which relies on KEGG annotations in the identification of AMGs. Contigs used for detection of AMGs were the ones initially identified by VIBRANT as viral, i.e., the 273,365 contigs. AMGs annotated as the psbA photosystem II protein D1 (PsbA) were further studied to evaluate the relatedness of these AMGs with previously reported PsbA reference sequences. To this end, we collected the AMGs identified in this study as PsbA (331 sequences) as well as viral psbA reference sequences from NCBI protein database using this search query: {(psba[Gene Name]) AND viruses[Organism] AND refseq[filter]},[9] which gave 22 sequences. Both datasets (Freshwter and RefSeq) were dereplicated using cd-hit (Li and Godzik, 2006; parameters: -c 0.95 -n 5 -d 0) to generate 161 and 12 sequences, respectively. Sequences were combined, then aligned using mafft v7.490 (Katoh and Standley, 2013; parameters: --globalpair --maxiterate 1,000). Columns with more than 50% gaps were removed using trimal v1.4 (Capella-Gutiérrez et al., 2009). The phylogenetic tree was inferred using fasttree v2.1.11 (Price et al., 2010) with default parameters. The tree was visualized using the Interactive Tree of Life (iTOL) online tool (Letunic and Bork, 2021).

7 https://sourceforge.net/projects/bbmap/

8 https://victor.dsmz.de

9 https://www.ncbi.nlm.nih.gov/protein/

TABLE 1 Freshwater biomes of the metagenomes used in this study.

Freshwater biome	Number of metagenomes	
	Before data analysis	After data analysis
Ballast water	5	–
Catchment	18	–
Estuary	16	10
Fish pond	3	3
Groundwater	8	8
Harbor water	3	–
Lake	84	37
Pure water (MilliQ)	26	–
Reclaimed water	6	3
Water reservoir	26	4
Watershed	182	75
Wastewater treatment plant	3	3
Total	380	143

Taxonomy of viral contigs

On January 7, 2021, both taxdump and prot.accession2taxid.FULL files were downloaded from NCBI Taxonomy FTP site.[10] These files were used to extract NCBI protein accessions belonging only to Viruses Superkingdom. These accessions were then used to extract viral protein sequences from the non-redundant protein (nr) database fasta file, downloaded from NCBI BLAST database FTP site[11] on January 7, 2021. Viral proteins extracted from nr were used to make a custom DIAMOND (Buchfink et al., 2015) database, here referred to as viral nr, which included 3,267,301 protein sequences. Open reading frames (ORFs) were called from contigs identified by VIBRANT as viral (273,365 contigs), using prodigal v2.6.3 (Hyatt et al., 2010) in its meta mode. ORFs called from contigs were aligned to viral nr using DIAMOND (Buchfink et al., 2015; Command: blastp; Options: --top 50 --matrix BLOSUM62 --evalue 0.001 --block-size 2 --index-chunks 4 --tmpdir.). Taxonomoy for each contig was identified using CAT v5.2.1 (von Meijenfeldt et al., 2019), which relies on a voting-based method that considers all ORFs of a particular contig and assigns taxonomy to a contig based on the lowest common ancestor, accordingly. Taxonomy assigned to contigs based on less than 50% of ORFs was disregarded. Additionally, during counting of contigs assigned to different taxonomic ranks, taxonomic assignments marked with an asterisk (a feature of CAT to mark suggestive taxonomic assignments, where conflict between classifications was not enough to allow more confident classification), were also removed.

10 ftp://ftp.ncbi.nih.gov/pub/taxonomy/

11 https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/

Results

Dataset overview

Quality control and sequence assembly

We downloaded 380 freshwater viral metagenomes from SRA samples from several countries/locations throughout the globe, encompassing four out of seven continents (Supplementary Table S2, Figure S2). These viromes were collected from diverse freshwater biomes, e.g., lakes, groundwater, wastewater, etc. (Table 1). Raw sequences of these 380 metagenomes amounted to over 614 gigabases (1.28 terabytes). Only 476 gigabases (0.96 terabytes) passed quality control (Supplementary Table S3). Out of 380 metagenomes, 347 (91.3%) generated contigs longer than 1,000 bp and those were the ones used for further study. The total number of contigs longer than 1,000 bp was 3,697,298 (Supplementary Table S4).

Assessment and removal of cellular contamination

We carried out several steps to assess and remove cellular contamination (Supplementary Figure S1). Only two metagenomes (SRR1658890 and SRR1658891) showed no detected ribosomal DNA sequences or any reads mapped to possible cellular contaminants (Supplementary Table S5). Generally, the extent of detected ribosomal DNA sequences and level of reads mapped to possible cellular contaminants, those two measures of cellular contamination, agreed remarkably with strong correlation (Supplementary Figure S3). For six metagenomes out of 347 (0.9%), percent bases mapped to cellular sequences exceeded 90%, attributed mainly to ribosomal DNA sequences (Supplementary Table S5). We decided to proceed only with metagenomes with no more than 2% bases mapped to cellular sequences and less than 5‰ predicted rDNA bases. Only 143 viral metagenomes out of 347 (41%) fulfilled these criteria. Out of these 143 metagenomes, 10 (7%) showed no predicted rDNA, 80 (56%) showed less than 0.2‰ rDNA and 53 (37%) showed more than 0.2‰ rDNA and up to less than 5‰. The total number of contigs for the selected 143 viromes was 1,778,319. After the removal of contigs showing less than 95% coverage with clean reads, the number of contigs was reduced to 1,763,349. Then, after dereplication and elimination of likely redundant contigs, the number of contigs became 1,226,122 (≥1,000 bp). VIBRANT (Kieft et al., 2020) was able to predict 273,365 contigs out of this non-redundant contig pool to belong to phages.

Viral clusters

Contigs predicted to be viral that were longer than or equal to 10,000 bp (20,107 contigs) were clustered into vOTUs. vOTUs representatives (19,210 contigs) were clustered based on their shared proteins together with JGI IMG/VR sequences as well as RefSeq prokaryotic viruses release 201. Out of 19,210 contigs 5,547 (28.9%) were clustered into clusters of two or more members (Figure 1). The total number of these clusters was 406. Out of these 406 clusters, only 31 clusters (7.6%), were solely composed of contigs from the studied freshwater metagenomes (totaling 138

contigs) i.e., these 31 clusters did not include any sequence from the JGI IMG/VR or RefSeq prokaryotic viruses. Less contigs were identified as singletons or outliers (~2 and ~3%, respectively; Figure 1). Most of the contigs, however, were denoted as overlaps (~66%; Figure 1) meaning that they were ascribed to more than one cluster, obscuring their cluster affiliation.

We considered contigs, which did not cluster with any previously known sequences, as novel genus-rank contigs. These included clustered contigs that did not include in their clusters any sequence from the JGI IMG/VR or RefSeq prokaryotic viruses, in addition to singletons and outliers, amounting to 1,206 contigs (6%; Table 2).

Recovered viral genomes

Out of the 19,210 vOTUs representatives, 549 complete and near-complete (high quality draft) genomes could be recovered. Out of these 549 genomes, novel genus-rank genomes, as defined above, were 10 (one complete and nine high quality draft genomes, Table 2). For these novel genomes, genome size ranged between 15,100 bp and 156,997 bp (Mean = 48,712.6 bp, Median = 34,324.5 bp). Coding sequences (CDS) were 561, 237 (42.2%) of which were completely unknown, i.e., had no hits in any of the KEGG, Pfam, or VOG databases. Moreover, 82 CDS (14.6%), despite having VOG hits, the proteins were of unknown function, hence designated hypothetical proteins. The remaining CDS (242 gene product, 43.1%) were largely dominated by proteins involved in DNA binding, replication, recombination, repair and metabolism in addition to phage structural proteins (Supplementary Table S6). As for the full record of viral genomes/genome fragments (273,365 contigs), out of 2,119,106 CDS, only 429,341 (20.2%) could be functionally annotated. For a complete list of annotations of all recovered genomes/genome fragments, please refer to Data Availability Section.

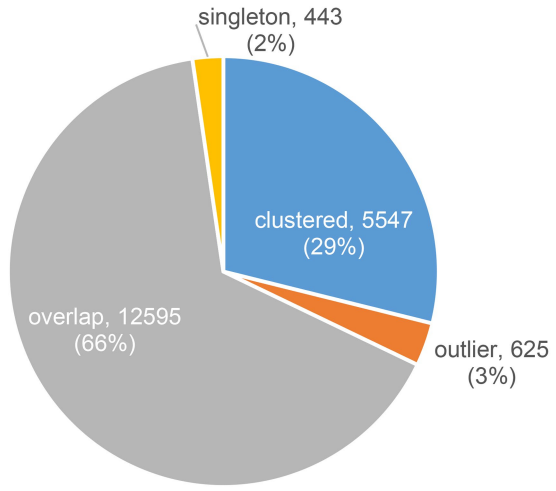

FIGURE 1
Distribution of contigs according to their clustering result based on their shared proteins using vConTACT2.

Predicted hosts of viral genomes

Out of 19,210 viral genomes, hosts could be successfully predicted for only 343 genomes (1.79%). All predicted hosts belonged to Bacteria (Supplementary Table S7). These hosts are members of nine different bacterial phyla, the most abundant of which were Proteobacteria (222 genomes), followed by Bacteroidetes (52 genomes) and Actinobacteria (34 genomes). The total number of detected genera was 141, with the most abundant genus being Flavobacterium (23 genomes), followed by Pectobacterium (19 genomes) and Salinispora (14 genomes).

Abundance of vOTUs representatives in the studied biomes

The abundance of the 19,210 vOTUs representatives was calculated for each biome and intersections between biomes based on common contigs with nonzero abundance among different biomes was determined (Figure 2). Only 29 contigs (< 0.2%) were found to have nonzero abundance in all studied biomes. A list of all common contigs among different biomes is included in Supplementary Table S8. The abundance of the 29 contigs common among all studied biomes is shown in Figure 3. It is worth noting that these 29 contigs originally belonged to only three biomes, namely estuary, groundwater and WWTP (Figure 4). Figure 4 shows the phylogenomic tree for these 29 contigs, yielding an average support of 53%. Although these contigs could be clustered into two major clades (Figure 4), no particular pattern could be observed concerning the clustering of the contigs based on the biomes they originated from. The OPTSIL clustering yielded 15 species clusters, eight genus clusters and four family level clusters.

Interestingly, one contig, namely freshwater_SRR107147.contig000000001 showed the highest abundance among all abundance values (Supplementary Table S9; Figure 3). Its abundance was highest in groundwater, comprising more than 7% of all groundwater reads or a relative abundance of 1.6×10^6 reads per million reads per million bases. It's worth noting that this contig was classified as a medium-quality draft genome by VIBRANT. The genome size was 44,722 bp. While most of its coding sequences (CDS) were identified as hypothetical proteins, only nine out of 74 CDS (12.2%) were annotated to several functions, including helicases, DNA-cytosine methyltransferase, phage terminase, as well as capsid protein and other uncharacterized phage proteins (Figure 5). Taxonomic annotation of this genome showed that it belongs to *Caudovirales*, which means that it is a double-stranded DNA (dsDNA) phage. Host prediction for this genome showed that its potential host is *Methylomonas* (Gammaproteobacterium).

TABLE 2 Genome quality versus clustering status.

		Genome quality				Total
		Complete circular	High quality draft	Medium quality draft	Low quality draft	
Clustering status	Clustered, but not with any previously known sequence	1	7	14	116	138
	Clustered with previously known sequence(s)	23	66	603	4,717	5,409
	Singleton	–	1	15	427	443
	Outlier	–	1	23	601	625
	Overlap	70	380	1798	10,347	12,595
	Total	94	455	2,453	16,208	19,210

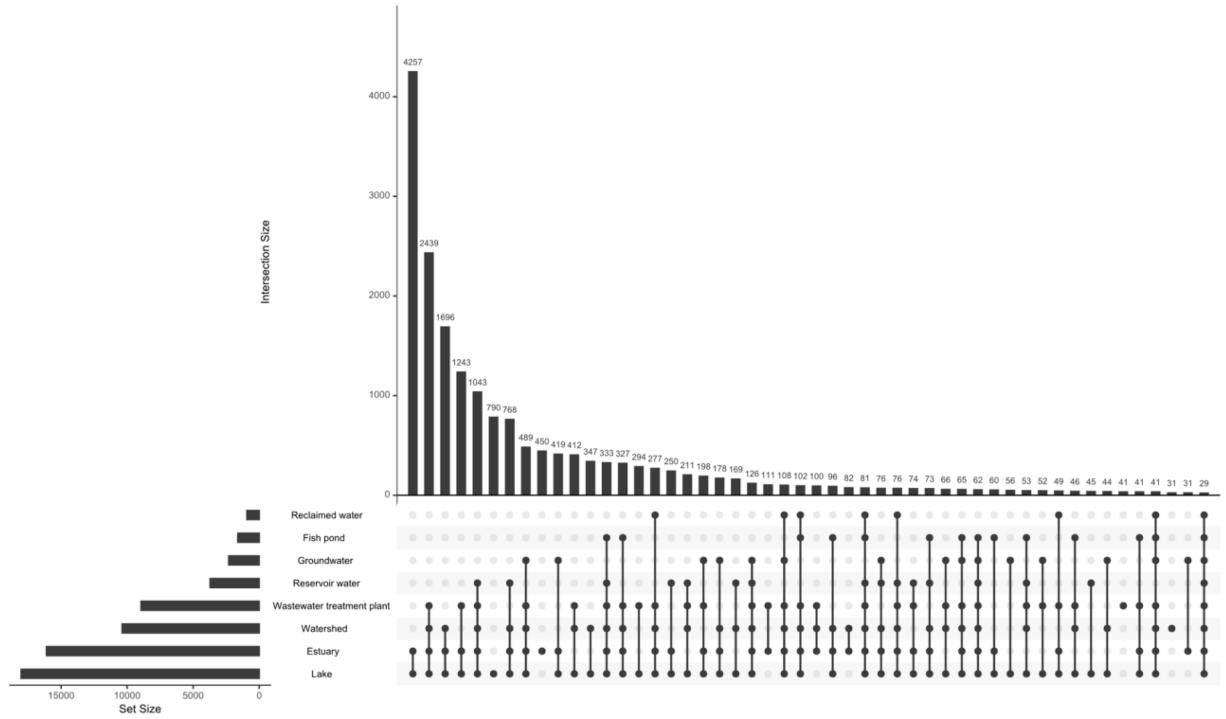

FIGURE 2
Intersection between different biomes based on common contigs with nonzero abundance. The figure was generated using upset function of UpSetR package v.1.4.0 (nsets=8, order.by="freq," nintersects=50, all other arguments set to default; Conway et al., 2017).

Auxiliary metabolic genes

A total of 17,139 AMGs (Supplementary Table S10) were detected in the contigs deemed by VIBRANT as viral (273,365 contigs). These AMGs belonged to 11 different categories of metabolism (Supplementary Figure S4, Table S11). Top three metabolic pathways with the highest numbers of AMGs pertained to carbohydrate metabolism with a total of 5,885 AMGs, followed by amino acid metabolism (4,966 AMGs), then metabolism of cofactors and vitamins (4,675 AMGs). Bottom three metabolic pathways, with the lowest numbers of AMGs, belonged to lipid metabolism, metabolism of other amino acids and xenobiotics biodegradation and metabolism with AMGs counts of 203, 167 and 71, respectively. AMGs could be detected in all studied biomes except fish ponds (Supplementary Table S10). Most of these AMGs were detected in estuary samples [15,484 out of 17,139 (90.3%)], followed by 1,282 AMGs in WWTP (7.5%). Diversity of the detected AMGs with regards to the pathways they belonged to was highest in estuary and WWTP samples (11 pathways, each), but was still relatively high in reservoir and lake samples (nine and eight pathways, respectively), despite the limited numbers of AMGs in these biomes. Numbers of detected AMGs and the pathways they belonged to were lowest in groundwater, reclaimed water and watershed (six, four and three AMGs and two, two and

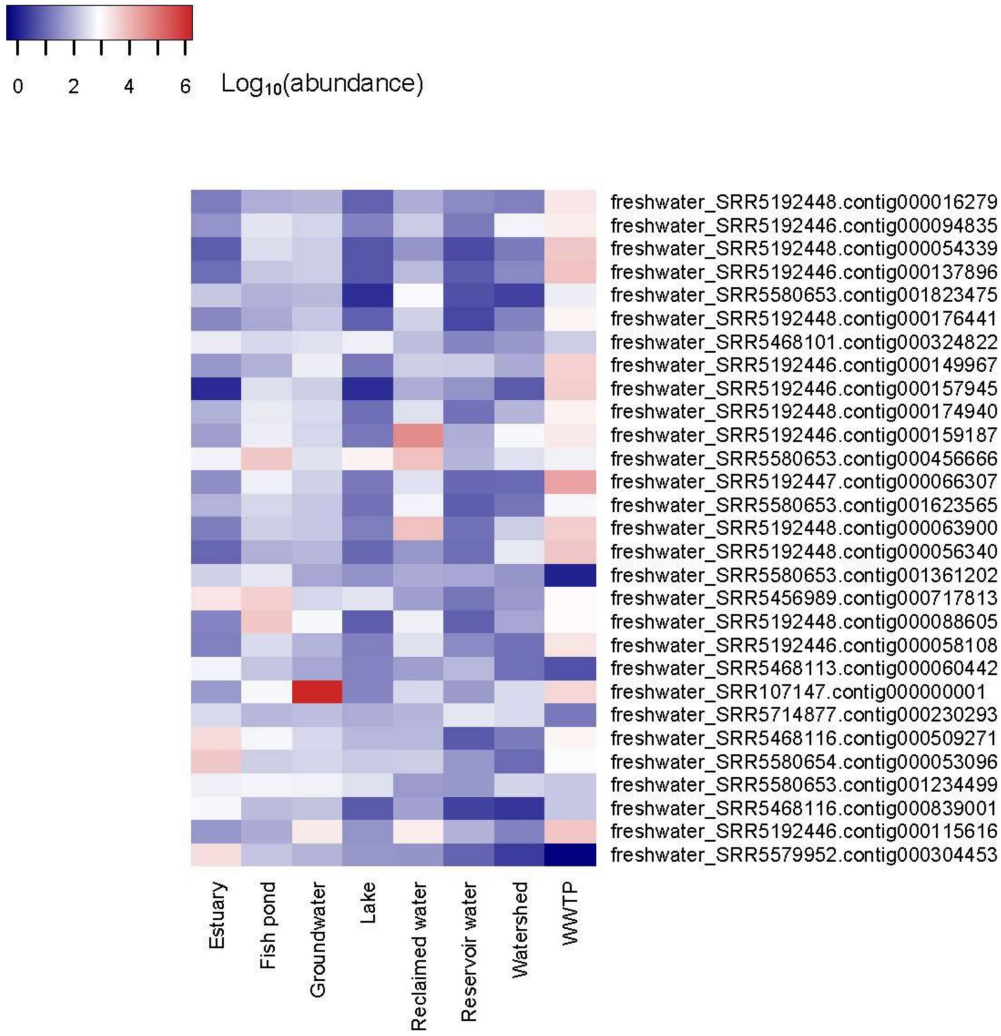

FIGURE 3
Heatmap of the abundance of 29 contigs with nonzero abundance among all studied biomes. The heatmap was generated using Heatmap3 R package v.1.1.7 (Zhao et al., 2014) in R v.4.0.2 (R Development Core Team, 2020) after $\log_{10}$ transformation of abundance values for better visualization.

three pathways, respectively). Concerning the taxonomy of these AMGs, it could be assigned at the order level for 12,240 AMGs out of 17,139 (71.4%; Supplementary Table S10). More than 12,000 of these AMGs (98.7%) belonged to the Order *Caudovirales*. On the family level, only 2,882 out of 17,139 AMGs (16.8%) could be assigned taxonomy. Out of these 2,882 AMGs, 1,488 AMGs (51.6%) belonged to *Myoviridae*.

Additionally, manual inspection of individual AMGs (Supplementary Table S10) highlighted the presence of putative antibiotic resistance proteins e.g., CDS annotated as beta-lactamase superfamily domain (five CDS), one CDS annotated as metallo-beta-lactamase superfamily and another annotated as cephalosporin hydroxylase. Similarly, 265 CDS were annotated as photosystem II protein D1 (PsbA) and 93 CDS annotated as photosystem II protein D2 (PsbD), involved in photosynthesis. Moreover, several CDS involved in xenobiotic degradation, particularly hydrocarbon degradation were detected e.g., MhpE (5 CDS) involved in the degradation of aromatic compounds, and components of the phenylacetate degradation pathway, which are also involved in the degradation of aromatic compounds e.g., PaaA (one CDS), PaaD (four CDS) and PaaK (two CDS).

The PsbA phylogenetic tree (Figure 6) reveals the clustering of sequences into two main clades (green and black branches), with 96 and 78 sequences, respectively. The green-branched clade contains all of the NCBI's RefSeq PsbA sequences (12 sequences), with particular concentration of most of these sequences along the middle subclade of the main, green-branched clade. On the other hand, the black-branched is solely made from sequences from the currently studied dataset (78 out of 162 dereplicated sequences (48.1%)).

Taxonomy of viral contigs

ORFs called from VIBRANT-identified viral contigs were aligned to viral nr and taxonomy was inferred using CAT

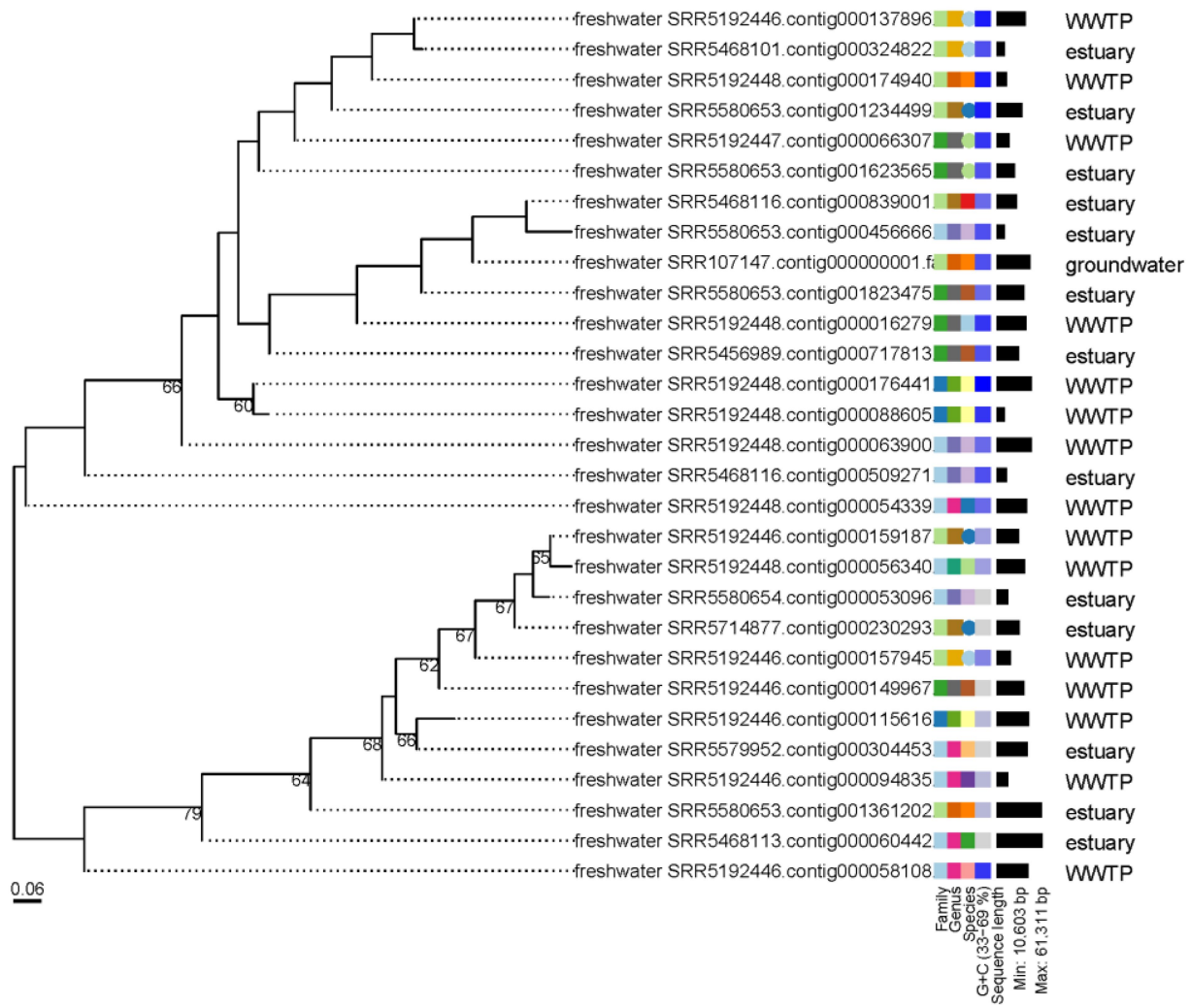

FIGURE 4
Phylogenomic tree of the 29 contigs with nonzero abundance among all studied biomes. The numbers above branches are GBDP pseudo-bootstrap support values from 100 replications. The branch lengths of the resulting VICTOR tree are scaled in terms of the D4 distance formula used.

v5.2.1 (von Meijenfeldt et al., 2019). Out of 273,365 contigs, 173,392 contigs (63%) could be classified as Viruses at the Superkingdom rank (Table 3) with at least 50% of ORFs supporting this classification. Decreasing number of contigs could be assigned to lower ranks with 43, 30, 30 and 15% of contigs assigned to known viral phyla, classes, orders and families, respectively. Uroviricota was the phylum with the highest number of assigned contigs, amounting to 116,250 (~43% of viral contigs). On the class and order levels, most assigned contigs pertained to Caudoviricetes and *Caudovirales*, respectively with exactly the same number of contigs (80,850 contigs). As for the family level, three different families, namely *Siphoviridae* (16,858 contigs), *Myoviridae* (12,840 contigs) and *Podoviridae* (7,449 contigs) recruited the highest number of assigned contigs compared to the rest of assigned families. Detailed taxonomy of all assigned contigs can be found in Supplementary Table S12.

Discussion

In this study, we sought to investigate freshwater viromes in a global context, especially given the little knowledge available about viromes in this environment. A simple PubMed search with search queries: (marine viral metagenomes) versus (freshwater viral metagenomes) shows an annual number of publications more than two times higher for marine viral metagenomes compared with freshwater viral metagenomes (Supplementary Figure S5). We downloaded 380 freshwater viral metagenomes from SRA amounting to more than one terabyte of data. We set out to leverage this big data to obtain valuable information about viruses in freshwater. In order to do that, we had to carefully consider cellular contamination in the downloaded datasets. Surprisingly, the rDNA content of many datasets approached or even exceeded 90% of the total number of reads (Supplementary Figure S2; Zolfo et al., 2019). This strikingly high rDNA content most likely means that these

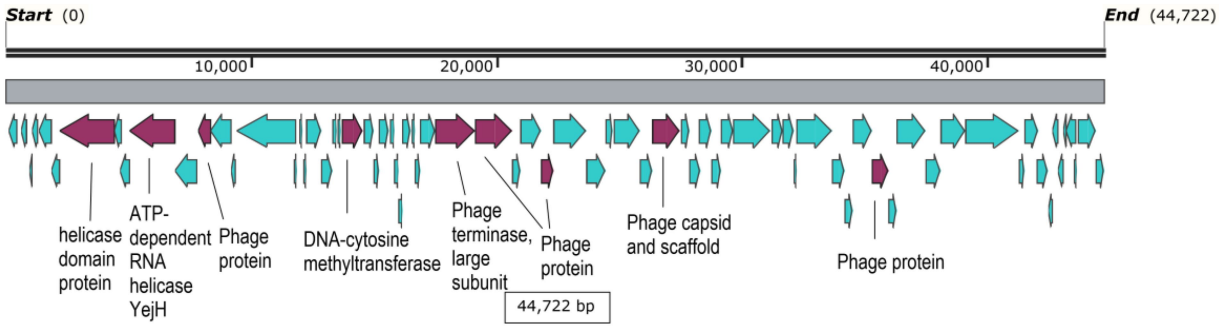

FIGURE 5
Gene map of freshwater_SRR107147.contig000000001. CDS highlighted in cyan are hypothetical proteins, while the purple ones have labeled annotations.

datasets were falsely labeled as viral metagenomes, while they are in fact rDNA amplicon sequences. It is not uncommon for viral metagenomic projects to sequence 16S rDNA amplicons to correlate viral with bacterial diversity (Parsley et al., 2010; Park et al., 2011; Skvortsov et al., 2016). Yet, the mislabeling of these amplicon sequences as viral metagenomes in metadata warrants attention to information submitted with each file in databases to avoid unnecessary processing of such files and waste of time and computational resources.

To look for novel genomes, we clustered VIBRANT viral contigs greater than 10 kb into vOTUs then using vConTACT2 (Bin Jang et al., 2019), we could cluster these contigs based on their shared proteins with JGI IMG/VR (Paez-Espino et al., 2018) and/or RefSeq prokaryotic viral genomes. Novel genomes/genome fragments, based on this approach, were 1,206, only 10 of which were complete, circular or high-quality draft genomes. The number might seem small, but it is still considerable, given the fact that clustering was done with the most comprehensive and diverse available viral resource JGI IMG/VR (Paez-Espino et al., 2018). The percentage of functionally annotated CDS of these 10 complete/high-quality draft genomes was much higher than the overall percentage of functionally annotated CDS of all viral contigs (43.1% versus 20.2%). This could possibly be explained given that VIBRANT's ability to identify genome completeness relies on the identification of viral hall mark proteins, which could also mean that complete and high-quality draft genomes would have better levels of annotation. We have previously reported similar annotation ratio (21.2%) when Pfam was used to annotate human virome protein clusters (Elbehery et al., 2018). Interestingly, Pfam is also used in VIBRANT's annotation of viral contigs, in addition to KEGG and VOG (Kieft et al., 2020). Generally, the fact that viruses are underrepresented in databases reduces the level of functional annotation of metagenomic sequences. For example, NCBI Genome resource[12] contains only 41,909 viral genomes versus 298,742 prokaryotic genomes, as of January 16, 2021.

Exploring the abundance of vOTUs showed that less than 0.2% of these contigs had nonzero abundance in all of the studied biomes. This could potentially mean that the core (shared) freshwater virome is minimal and that each freshwater environment has a distinct virome, which is site-specific. This explanation agrees with previous studies which highlighted that virome diversity differs among samples of the same environment, including those spatially close to each other. These studies suggest that viral diversity is mainly dictated by environmental variables specific to each sample site (Saxton et al., 2016; Adriaenssens et al., 2017). Looking up the taxonomy of the contigs with nonzero abundance in all studied biomes in Supplementary Table S12 shows that 41.4% of them could not be assigned taxonomy, while the rest mostly belonged to *Caudovirales* with no enough support to identify taxonomy at the family level. Phylogenomics and estimation of taxonomic boundaries of these contigs showed that they are composed of limited numbers of viral species, genera and families (Figure 4). This observation could denote a limited complexity of the freshwater core virome as previously suggested for gut and ocean viromes (Brum et al., 2015; Broecker et al., 2017).

The most abundant genome with nonzero abundance in all biomes was freshwater_SRR107147.contig000000001. The functional annotation of this genome showed several CDS with different functions, including helicases, which are important for genome replication (Frick and Lam, 2006). DNA-cytosine methyltransferase was another protein detected in this genome, whose function is to protect bacteriophages against restriction-modification in bacterial hosts (Murphy et al., 2013). Ultimately, phage terminase is another detected protein of particular importance in genome packaging at the end of the lytic cycles of infection (Shen et al., 2012). It is worth noting that *Methanococcus*, the potential host of this genome is usually isolated from freshwater habitats and that methanotrophs in general constitute a large fraction of the microbial diversity in many freshwater aquatic environments (Bowman, 2006). This probably explains why freshwater_SRR107147.contig000000001 phage genome was most abundant in a multitude of freshwater environments.

Auxiliary metabolic genes is a term that has first been suggested by Breitbart and colleagues (Breitbart et al., 2007) to describe genes

12 https://www.ncbi.nlm.nih.gov/genome/

TABLE 3 Breakdown of the different taxonomic assignments of viral contigs.

Classification	Breakdown		Total	Percentage
Superkingdom	Viruses	173,392	173,392	63%
Phylum	Cressdnaviricota	90	118,471	43%
	Hofneiviricota	1		
	Nucleocytoviricota	661		
	Peploviricota	1		
	Phixviricota	1,209		
	Preplasmiviricota	259		
	Uroviricota	116,250		
Class	Arfiviricetes	13	83,030	30%
	Caudoviricetes	80,850		
	Faserviricetes	1		
	Malgrandaviricetes	1,209		
	Maveriviricetes	258		
	Megaviricetes	660		
	Repensiviricetes	38		
	Tectiliviricetes	1		
Order	Algavirales	461	82,937	30%
	Caudovirales	80,850		
	Cirlivirales	13		
	Geplafuvirales	38		
	Imitervirales	94		
	Petitvirales	1,209		
	Pimascovirales	31		
	Priklausovirales	239		
	Tubulavirales	1		
	Vinavirales	1		
Family	Ackermannviridae	24	41,027	15%
	Adintoviridae	15		
	Autographiviridae	1754		
	Baculoviridae	1		
	Chaseviridae	8		
	Circoviridae	12		
	Corticoviridae	1		
	Cruciviridae	10		
	Demerecviridae	2		
	Drexlerviridae	4		
	Fuselloviridae	1		
	Genomoviridae	38		
	Herelleviridae	28		
	Inoviridae	1		
	Iridoviridae	27		
	Lavidaviridae	239		
	Marseilleviridae	2		
	Microviridae	1,209		
	Mimiviridae	81		
	Myoviridae	12,840		
	Phycodnaviridae	413		
	Pithoviridae	10		
	Podoviridae	7,449		
	Siphoviridae	16,858		

of bacterial host origin believed to be leftovers of previous horizontal gene transfer processes, found in bacteriophages and play important role in viral infection of bacteria. Actually, bacteriophages have been thought to utilize such genes in an integrative way to enhance infection efficiency. The variation in the number and diversity of detected AMGs among different biomes (Supplementary Table S10) could be interpreted on the merits of the differences in sequencing depth as well as the variation in environmental variables among different biomes, which warrants the presence of diverse AMGs to maximize host fitness in the face of different conditions (Warwick-Dugdale et al., 2019b). Notably, The highest numbers of AMGs were detected in the *Myoviridae* family (Supplementary Table S10), which was recently reported (Heyerhoff et al., 2022), probably indicating that phages of this particular family have special emphasis on the employment of AMGs in their life cycles potentially for an overall improved viral fitness. In this study, numerous AMGs, belonging to several metabolic pathways could be detected. For example, we detected photosynthetic core photosystem II reaction center proteins PsbA and PsbD. These proteins are widespread in cyanophages in both marine (Sullivan et al., 2006) and freshwater (Ruiz-Perez et al., 2019) aquatic systems. During their infection of cyanobacteria, cyanophages direct cyanobacteria to express these viral proteins to enhance photosynthesis and increase ATP production (Breitbart et al., 2007; Thompson et al., 2011). Similarly, we detected more AMGs belonging to carbohydrate metabolism, particularly pentose phosphate pathway, and nucleotide synthesis (Supplementary Table S11). In fact, cyanophages make use of these AMGs to eventually increase nucleotide production, which in turn increases viral replication and production (Breitbart et al., 2007; Thompson et al., 2011). Remarkably, almost half of the PsbA AMGs detected in the currently studied freshwater dataset did not cluster with any of the PsbA reference sequences included in the constructed phylogenetic tree (Figure 6), denoting their uniqueness.

In addition to AMGs, which potentially improve infection efficiency, we detected other AMGs, such as potential antibiotic resistance proteins as well as xenobiotic degradation proteins. Antibiotic resistance genes have previously been reported in several viromes, including aquaculture wastewater (Colombo et al., 2016), river (Colombo et al., 2017), and urban surface freshwater (Moon et al., 2020). Yet, there are concerns that the reported genes are overestimated; in addition, their functionality and ability to confer resistance in bacteria are questionable (Enault et al., 2017). Therefore, experimental verification of such genes is essential. On the other hand, we could detect MhpE, PaaA, PaaD and PaaK, all of which are involved in aromatic compound degradation, which could potentially be used in bioremediation. Generally, bacteriophages could have several roles in polluted environments. The presence of environmental pollutants could induce prophages in lysogenic bacteria, increasing the overall abundance of free viruses (Cochran et al., 1998). Besides, it was shown that bacteriophages control the abundances of microbial populations, following "killing the winner" pattern with a stronger influence in diesel-contaminated water systems compared to control ones (Sauret et al., 2015). Another study (Paterson et al., 2019) suggested another model in

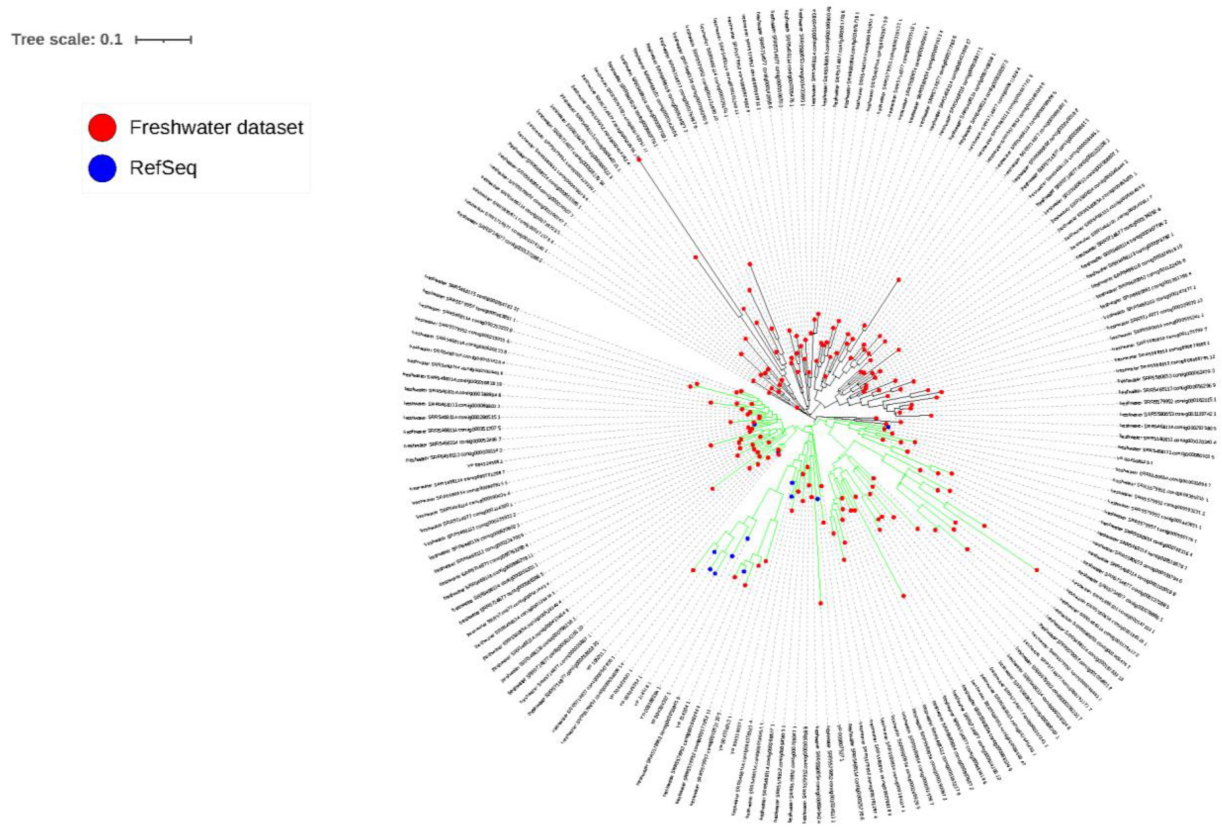

FIGURE 6
PsbA phylogenetic tree. Red circles at the tip of branches denote sequences identified in the current study, while blue circles denote RefSeq PsbA sequences collected from NCBI. Two main clades can be observed, whose branches are colored in green and black.

trichloroethene-contaminated groundwater, where viruses shift from lytic to lysogenic life cycle when bacterial hosts are persistently available in low abundances. The model was called Piggyback-the-Persistent (PtP). Furthermore, bacteriophages could harbor hydrocarbon-degrading genes (Costeira, 2019). MhpE is an aldolase, constituting a part of the 3-hydroxyphenylpropionic acid degradation pathway, where it catalyzes the cleavage of 4-hydroxy-2-ketopentanoic acid, giving pyruvate and acetaldehyde, which is then converted to acetyl coenzyme A by MhpF (Diaz et al., 2001). In contrast, PaaA, PaaD and PaaK are components of the phenylacetic acid degradation pathway, where PaaK (phenylacetate-CoA ligase) catalyzes the first step in the pathway, converting phenylacetic acid to phenylacetyl-CoA, then, PaaABCDE complex catalyzes the epoxidation of the aromatic ring (Ismail and Gescher, 2012).

Taxonomic analysis of the studied viral contigs showed that most of the assigned contigs belonged to tailed phages of the order *Caudovirales*, particularly of the families *Siphoviridae*, *Myoviridae* and *Podviridae*. This taxonomy agrees with many of the reported taxonomies of many freshwater viromes (Mohiuddin and Schellhorn, 2015; Potapov et al., 2019; Rusiñol et al., 2020; Moon et al., 2020), whose most abundant taxa also belonged to *Caudovirales*, but probably with some shuffling of the order of the most abundant family among the top three, namely *Siphoviridae*, *Myoviridae* and *Podviridae*. *Caudovirales* are tailed bacteriophages with their structures made up of icosahedral head, neck and tail (Iwasaki et al., 2018). The families of this order differ mainly in tail characteristics: *Syphoviridae* with long non-contractile tails, *Myoviridae* with long contractile tails and *Podoviridae* with short tails. *Caudovirales* generally dominate databases, which could lead to some bias in taxonomic annotation (Palermo et al., 2021). Although most metagenomic studies in aquatic environments report *Caudovirales* (tailed phages) as most abundant, a global study based on electron microscopy revealed that non-tailed viruses were most abundant in marine environments (Brum et al., 2013). A similar study for freshwater is necessary.

Taken together, we have assessed freshwater viromes from several locations all around the world, giving a global highlight of viruses in this environment. Besides, we shed light on common mislabeling and unorganized metadata in databases that could lead to profoundly misled studies, or at least waste of time, effort and computational resources. We strongly recommend the implementation of a quick pipeline through which a small subset of randomly picked sequence reads from metagenomic data should run before uploading to SRA and similar repositories. The pipeline should automatically predict whether these datasets are amplicons (e.g., 16S data), microbiomes or viromes, to verify the claims of the

uploaders. We also could recover many complete and high-quality draft genomes, although only a few of them could be considered novel. Moreover, our study emphasizes the role of viral AMGs in bacterial metabolism. Finally, our carefully decontaminated and cautiously selected set of viral contigs remains a potentially useful resource for future studies in environmental viromics.

Data availability statement

All relevant data including sequences of the viral contigs, annotations, clusters and the custom Python script are available at: https://osf.io/ucqv4/ (DOI: 10.17605/OSF.IO/UCQV4).

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This work was funded by the German Research Foundation (DFG Emmy Noether Program, project no. 273124240 and DE2360/1–1, as well as no. 391644373 and DE2360/2–1 awarded to LD).

Acknowledgments

We would like to thank the group members of LD for constructive discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2022.953500/full#supplementary-material

SUPPLEMENTARY TABLE S1
Downloaded viral metagenomes.

SUPPLEMENTARY TABLE S2
Number of metagenomes before and after decontamination.

SUPPLEMENTARY TABLE S3
Raw and quality controlled read statistics.

SUPPLEMENTARY TABLE S4
Contig statistics.

SUPPLEMENTARY TABLE S5
Percent bases mapped to cellular sequences versus rRNA reads per 1000 reads.

SUPPLEMENTARY TABLE S6
Genbank table of the 10 novel complete and high-quality draft genomes.

SUPPLEMENTARY TABLE S7
Hosts of all 19210 vOTUs.

SUPPLEMENTARY TABLE S8
List of common contigs of non-zero abundance among different biomes.

SUPPLEMENTARY TABLE S9
Abundance of all 19210 contigs (vOTUs representatives) in all studied freshwater biomes.

SUPPLEMENTARY TABLE S10
AMG individual proteins and breakdown of their counts according to biomes and taxonomy.

SUPPLEMENTARY TABLE S11
AMG pathways.

SUPPLEMENTARY TABLE S12
Taxonomy of viral contigs.

SUPPLEMENTARY FIGURE S1
Diagram explaining how cellular contamination in the metagenomic datasets was assessed and removed.

SUPPLEMENTARY FIGURE S2
World map showing the locations of the studied 143 metagenomes. The size of the circles at each location is proportional to the number of metagenomes collected from this location.

SUPPLEMENTARY FIGURE S3
Percent bases mapped to cellular sequences versus rRNA reads per 1000 reads.

SUPPLEMENTARY FIGURE S4
Numbers of detected AMGs classified according to metabolism categories.

SUPPLEMENTARY FIGURE S5
Annual number of publications for marine viral metagenomes versus freshwater viral metagenomes.

References

Adriaenssens, E. M., Kramer, R., Van Goethem, M. W., Makhalanyane, T. P., Hogg, I., and Cowan, D. A. (2017). Environmental drivers of viral community composition in Antarctic soils identified by viromics. *Microbiome* 5, 83. doi: 10.1186/s40168-017-0301-7

Akhter, S., Aziz, R. K., and Edwards, R. A. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* 40, e126. doi: 10.1093/nar/gks406

Auguet, J. C., Montanié, H., Hartmann, H. J., Lebaron, P., Casamayor, E. O., Catala, P., et al. (2009). Potential effect of freshwater virus on the structure and activity of bacterial communities in the Marennes-Oléron Bay (France). *Microb. Ecol.* 57, 295–306. doi: 10.1007/s00248-008-9428-1

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* 37, 632–639. doi: 10.1038/s41587-019-0100-8

Bowman, J. (2006). "The Methanotrophs — The families Methylococcaceae and Methylocystaceae," in *The Prokaryotes: Volume 5: Proteobacteria: Alpha and Beta Subclasses*. eds. M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer and E. Stackebrandt (New York, NY: Springer New York), 266–289.

Breitbart, M. Y. A., Thompson, L. R., Suttle, C. A., and Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. *Oceanography* 20, 135–139. doi: 10.5670/oceanog.2007.58

Broecker, F., Russo, G., Klumpp, J., and Moelling, K. (2017). Stable core virome despite variable microbiome after fecal transfer. *Gut Microbes* 8, 214–220. doi: 10.1080/19490976.2016.1265196

Brum, J. R., Ignacio-Espinoza, J. C., Roux, S., Doulcier, G., Acinas, S. G., Alberti, A., et al. (2015). Patterns and ecological drivers of ocean viral communities. *Science* 348, 1261498. doi: 10.1126/science.1261498

Brum, J. R., Schenck, R. O., and Sullivan, M. B. (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* 7, 1738–1751. doi: 10.1038/ismej.2013.67

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Chen, L. X., Zhao, Y., McMahon, K. D., Mori, J. F., Jessen, G. L., Nelson, T. C., et al. (2019). Wide distribution of phage That infect freshwater SAR11 bacteria. *mSystems* 4:19. doi: 10.1128/mSystems.00410-19

Chevallereau, A., Blasdel, B. G., De Smet, J., Monot, M., Zimmermann, M., Kogadeeva, M., et al. (2016). Next-generation "-omics" approaches reveal a massive alteration of host RNA metabolism during bacteriophage infection of Pseudomonas aeruginosa. *PLoS Genet.* 12:e1006134. doi: 10.1371/journal.pgen.1006134

Cochran, P. K., Kellogg, C. A., and Paul, J. H. J. M. E. P. S. (1998). Prophage induction of indigenous marine lysogenic bacteria by environmental pollutants. *Marine Eco. Prog. Series* 164, 125–133. doi: 10.3354/meps164125

Colombo, S., Arioli, S., Guglielmetti, S., Lunelli, F., and Mora, D. (2016). Virome-associated antibiotic-resistance genes in an experimental aquaculture facility. *FEMS Microbiol. Ecol.* 92:003. doi: 10.1093/femsec/fiw003

Colombo, S., Arioli, S., Neri, E., Della Scala, G., Gargari, G., and Mora, D. (2017). Viromes As genetic reservoir for the microbial communities in aquatic environments. *A Focus on Antimicrob.-Resist. Genes.* 8:1095. doi: 10.3389/fmicb.2017.01095

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Costeira, R. (2019). Metagenomic Characterization of Microbial Communities in Groundwater Associated with Contaminated Land. Doctor of Philosophy Doctoral Thesis, Queen's University Belfast.

Coutinho, F. H., Cabello-Yeves, P. J., Gonzalez-Serrano, R., Rosselli, R., López-Pérez, M., Zemskaya, T. I., et al. (2020). New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. *Microbiome* 8, 163. doi: 10.1186/s40168-020-00936-4

Diaz, E., Ferrandez, A., Prieto, M. A. A., and Garcia, J. L. (2001). Biodegradation of aromatic compounds by *Escherichia coli*. *J Microbiol. Molecular Biol. Rev.* 65, 523–569. doi: 10.1128/MMBR.65.4.523-569.2001

Dion, M. B., Plante, P.-L., Zufferey, E., Shah, S. A., Corbeil, J., and Moineau, S. (2021). Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res.* 49, 3127–3138. doi: 10.1093/nar/gkab133

Edwards, R. A., and Rohwer, F. (2005). Viral metagenomics. *Nat. Rev. Microbiol.* 3, 504–510. doi: 10.1038/nrmicro1163

Elbehery, A. H. A., Feichtmayer, J., Singh, D., Griebler, C., and Deng, L. (2018). The human Virome protein cluster database (HVPC): A human viral metagenomic database for diversity and function. *Ann. Dent.* 9:1110. doi: 10.3389/fmicb.2018.01110

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M. B., and Petit, M. A. (2017). Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* 11, 237–247. doi: 10.1038/ismej.2016.90

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *Am. Nat.* 106, 645–668. doi: 10.1086/282802

Frick, D. N., and Lam, A. M. I. (2006). Understanding helicases as a means of virus control. *Curr. Pharm. Des.* 12, 1315–1338. doi: 10.2174/138161206776361147

Gazitúa, M. C., Vik, D. R., Roux, S., Gregory, A. C., Bolduc, B., Widner, B., et al. (2020). Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. *ISME J.* 15, 981–998. doi: 10.1038/s41396-020-00825-6

Göker, M., García-Blázquez, G., Voglmayr, H., Tellería, M. T., and Martín, M. P. (2009). Molecular taxonomy of Phytopathogenic fungi: A case study in Peronospora. *PLoS One* 4:e6319. doi: 10.1371/journal.pone.0006319

Grazziotin, A. L., Koonin, E. V., and Kristensen, D. M. (2017). Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* 45, D491–D498. doi: 10.1093/nar/gkw975

Güemes, A. G. C., Youle, M., Cantú, V. A., Felts, B., Nulton, J., and Rohwer, F. (2016). Viruses as Winners in the Game of Life. *Annu. Rev. Virol.* 3, 197–214. doi: 10.1146/annurev-virology-100114-054952

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy 2008)*. eds. G. Varoquaux, T. Vaught and J. Millman (CA, USA: Pasadena).

Heyerhoff, B., Engelen, B., and Bunse, C. (2022). Auxiliary metabolic gene functions in pelagic and benthic viruses of the Baltic Sea. *Front. Microbiol.* 13, 13. doi: 10.3389/fmicb.2022.863620

Huang, Y., Gilna, P., and Li, W. (2009). Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics (Oxford, England)* 25, 1338–1340. doi: 10.1093/bioinformatics/btp161

Hurwitz, B. L., Hallam, S. J., and Sullivan, M. B. (2013). Metabolic reprogramming by viruses in the sunlit and dark ocean. *Genome Biol.* 14, R123. doi: 10.1186/gb-2013-14-11-r123

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinfo.* 11, 119. doi: 10.1186/1471-2105-11-119

Ismail, W., and Gescher, J. (2012). Epoxy coenzyme A Thioester pathways for degradation of aromatic compounds. *J App. Environ. Microbiol.* 78, 5043–5051. doi: 10.1128/AEM.00633-12

Iwasaki, T., Yamashita, E., Nakagawa, A., Enomoto, A., Tomihara, M., and Takeda, S. (2018). Three-dimensional structures of bacteriophage neck subunits are shared in Podoviridae. *Siphoviridae and Myoviridae. Genes Cells* 23, 528–536. doi: 10.1111/gtc.12594

Jacquet, S., Miki, T., Noble, R., Peduzzi, P., and Wilhelm, S. (2010). Viruses in aquatic ecosystems: important advancements of the last 20 years and prospects for the future in the field of microbial oceanography and limnology. *Adv. Oceanogr. Limnol.* 1, 97–141. doi: 10.1080/19475721003743843

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kenzaka, T., Tani, K., and Nasu, M. (2010). High-frequency phage-mediated gene transfer in freshwater environments determined at single-cell level. *ISME J.* 4, 648–659. doi: 10.1038/ismej.2009.145

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

Krogh, A. (1934). Physiology of the blue whale. *Nature* 133, 635–637. doi: 10.1038/133635a0

Kumar, S., and Blaxter, M. L. (2010). Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* 11, 571. doi: 10.1186/1471-2164-11-571

Lefort, V., Desper, R., and Gascuel, O. (2015). FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32, 2798–2800. doi: 10.1093/molbev/msv150

Letunic, I., and Bork, P. (2021). Interactive tree Of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, C.-M., Li, D., Sadakane, K., Luo, R., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Logares, R., Bråte, J., Bertilsson, S., Clasen, J. L., Shalchian-Tabrizi, K., and Rengefors, K. (2009). Infrequent marine-freshwater transitions in the microbial world. *Trends Microbiol.* 17, 414–422. doi: 10.1016/j.tim.2009.05.010

Mara, P., Vik, D., Pachiadaki, M. G., Suter, E. A., Poulos, B., Taylor, G. T., et al. (2020). Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *ISME J.* 14, 3079–3092. doi: 10.1038/s41396-020-00739-3

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J* 17, 10–12. doi: 10.14806/ej.17.1.200

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., and Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinfo.* 14, 60. doi: 10.1186/1471-2105-14-60

Meier-Kolthoff, J. P., and Göker, M. (2017). VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* 33, 3396–3404. doi: 10.1093/bioinformatics/btx440

Meier-Kolthoff, J. P., Hahnke, R. L., Petersen, J., Scheuner, C., Michael, V., Fiebig, A., et al. (2014). Complete genome sequence of DSM 30083(T), the type strain (U5/41(T)) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Stand. Genomic Sci.* 9, 2. doi: 10.1186/1944-3277-9-2

Meunier, A., and Jacquet, S. (2015). Do phages impact microbial dynamics, prokaryotic community structure and nutrient dynamics in Lake Bourget? *Biology open* 4, 1528–1537. doi: 10.1242/bio.013003

Mohiuddin, M., and Schellhorn, H. (2015). Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front. Microbiol.* 6:960. doi: 10.3389/fmicb.2015.00960

Moon, K., Jeon, J. H., Kang, I., Park, K. S., Lee, K., Cha, C.-J., et al. (2020). Freshwater viral metagenome reveals novel and functional phage-borne antibiotic resistance genes. *Microbiome* 8, 75. doi: 10.1186/s40168-020-00863-4

Morrison, W. D., Miller, R. V., and Sayler, G. S. (1978). Frequency of F116-mediated transduction of Pseudomonas aeruginosa in a freshwater environment. *Appl. Environ. Microbiol.* 36, 724–730. doi: 10.1128/aem.36.5.724-730.1978

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C., and Pati, A. (2015). Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand. Genomic Sci.* 10, 18. doi: 10.1186/1944-3277-10-18

Murphy, J., Mahony, J., Ainsworth, S., Nauta, A., and van Sinderen, D. (2013). Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl. Environ. Microbiol.* 79, 7547–7555. doi: 10.1128/AEM.02229-13

Paez-Espino, D., Roux, S., Chen, I. M. A., Palaniappan, K., Ratner, A., Chu, K., et al. (2018). IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* 47, D678–D686. doi: 10.1093/nar/gky1127

Palermo, C. N., Shea, D. W., Short, S. M., and Elkins, C. A. (2021). Analysis of different size fractions provides a more complete perspective of viral diversity in a freshwater embayment. *Appl. Environ. Microbiol.* 87, e00197–e00121. doi: 10.1128/AEM.00197-21

Park, E.-J., Kim, K.-H., Abell, G. C. J., Kim, M.-S., Roh, S. W., and Bae, J.-W. (2011). Metagenomic analysis of the viral communities in fermented foods. *Appl. Environ. Microbiol.* 77, 1284–1291. doi: 10.1128/AEM.01859-10

Parsley, L. C., Consuegra, E. J., Thomas, S. J., Bhavsar, J., Land, A. M., Bhuiyan, N. N., et al. (2010). Census of the viral Metagenome within an activated sludge microbial assemblage. *Appl. Environ. Microbiol.* 76, 2673–2677. doi: 10.1128/AEM.02520-09

Paterson, J. S., Smith, R. J., McKerral, J. C., Dann, L. M., Launer, E., Goonan, P., et al. (2019). A hydrocarbon-contaminated aquifer reveals a piggyback-the-persistent viral strategy. *FEMS Microbiol. Ecol.* 95:116. doi: 10.1093/femsec/fiz116

Potapov, S. A., Tikhonova, I. V., Krasnopeev, A. Y., Kabilov, M. R., Tupikin, A. E., Chebunina, N. S., et al. (2019). Metagenomic analysis of Virioplankton from the pelagic zone of Lake Baikal. *Viruses* 11, 991. doi: 10.3390/v11110991

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

R Development Core Team. (2020). "*R: A Language and Environment for Statistical Computing*". (Vienna, Austria: R Foundation for Statistical Computing).

Replicon, J., Frankfater, A., and Miller, R. V. (1995). A continuous culture model to examine factors That affect transduction among Pseudomonas aeruginosa strains in freshwater environments. *Appl. Environ. Microbiol.* 61, 3359–3366. doi: 10.1128/aem.61.9.3359-3366.1995

Ripp, S., Ogunseitan, O. A., and Miller, R. V. (1994). Transduction of a freshwater microbial community by a new Pseudomonas aeruginosa generalized transducing phage, UT1. *Mol. Ecol.* 3, 121–126. doi: 10.1111/j.1365-294x.1994.tb00112.x

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., et al. (2019). Minimum information about an uncultivated virus genome (MIUViG). *Nat. Biotechnol.* 37, 29–37. doi: 10.1038/nbt.4306

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., et al. (2012). Assessing the diversity and specificity of two freshwater viral communities through Metagenomics. *PLoS One* 7:e33641. doi: 10.1371/journal.pone.0033641

Roux, S., Krupovic, M., Debroas, D., Forterre, P., and Enault, F. (2013). Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol.* 3, 130160. doi: 10.1098/rsob.130160

Ruiz-Perez, C. A., Tsementzi, D., Hatt, J. K., Sullivan, M. B., and Konstantinidis, K. T. (2019). Prevalence of viral photosynthesis genes along a freshwater to saltwater transect in Southeast USA. *Environ. Microbiol. Rep.* 11, 672–689. doi: 10.1111/1758-2229.12780

Rusiñol, M., Martínez-Puchol, S., Timoneda, N., Fernández-Cassi, X., Pérez-Cataluña, A., Fernández-Bravo, A., et al. (2020). Metagenomic analysis of viruses, bacteria and protozoa in irrigation water. *Int. J. Hyg. Environ. Health* 224:113440. doi: 10.1016/j.ijheh.2019.113440

Sauret, C., Böttjer, D., Talarmin, A., Guigue, C., Conan, P., Pujo-Pay, M., et al. (2015). Top-Down control of diesel-degrading prokaryotic communities. *Microb. Ecol.* 70, 445–458. doi: 10.1007/s00248-015-0596-5

Saxton, M. A., Naqvi, N. S., Rahman, F., Thompson, C. P., Chambers, R. M., Kaste, J. M., et al. (2016). Site-specific environmental factors control bacterial and viral diversity in stormwater retention ponds. *Aquat. Microb. Ecol.* 77, 23–36. doi: 10.3354/ame01786

Schmieder, R., and Edwards, R. (2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* 6:e17288. doi: 10.1371/journal.pone.0017288

Schmieder, R., and Edwards, R. (2011b). Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., et al. (2020). NCBI taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020:62. doi: 10.1093/database/baaa062

Shen, X., Li, M., Zeng, Y., Hu, X., Tan, Y., Rao, X., et al. (2012). Functional identification of the DNA packaging terminase from Pseudomonas aeruginosa phage PaP3. *Arch. Virol.* 157, 2133–2141. doi: 10.1007/s00705-012-1409-5

Skvortsov, T., de Leeuwe, C., Quinn, J. P., McGrath, J. W., Allen, C. C. R., McElarney, Y., et al. (2016). Metagenomic characterisation of the viral Community of Lough Neagh, the largest freshwater Lake in Ireland. *PLoS One* 11, e0150361. doi: 10.1371/journal.pone.0150361

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., and Chisholm, S. W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. *PLoS Biol.* 4:e234. doi: 10.1371/journal.pbio.0040234

Thingstad, T., and Lignell, R. J. A. M. E. (1997). Theoretical models for the control of bacterial growth rate, abundance. *Diver. Carbon Demand.* 13, 19–27. doi: 10.3354/ame013019

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., et al. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceed. Nat. Aca. Sci.* 108, E757–E764. doi: 10.1073/pnas.1102164108

von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H., and Dutilh, B. E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* 20, 217. doi: 10.1186/s13059-019-1817-x

Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J., and Temperton, B. (2019). Host-hijacking and planktonic piracy: how phages command the microbial high seas. *Virol. J.* 16, 15. doi: 10.1186/s12985-019-1120-1

Wilhelm, S. W., and Suttle, C. A. (1999). Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* 49, 781–788. doi: 10.2307/1313569

Xue, Z. (2022). *ncbitax2lin - convert NCBI taxonomy dump into lineages* [online]. Available at: https://github.com/zyxue/ncbitax2lin [Accessed June, 22, 2022].

Yu, G. (2020). Using ggtree to visualize data on tree-Like structures. *Curr. Protoc. Bioinformatics* 69:e96. doi: 10.1002/cpbi.96

Zhao, S., Guo, Y., Sheng, Q., and Shyr, Y. (2014). Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinfo.* 15, P16. doi: 10.1186/1471-2105-15-S10-P16

Zhu, Y., Stephens, R. M., Meltzer, P. S., and Davis, S. R. (2013). SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinfo.* 14, 19. doi: 10.1186/1471-2105-14-19

Zolfo, M., Pinto, F., Asnicar, F., Manghi, P., Tett, A., Bushman, F. D., et al. (2019). Detecting contamination in viromes using ViromeQC. *Nat. Biotechnol.* 37, 1408–1412. doi: 10.1038/s41587-019-0334-5