Comprehensive profiling of wastewater viromes by genomic sequencing

Emanuel Wyler (1)#, Chris Lauber (2), Artür Manukyan (1), Aylina Deter (1), Claudia Quedenau (1), Luiz Gustavo Teixeira Alves (1), Stefan Seitz (3, 4), Janine Altmüller (1, 5), Markus Landthaler (1, 6)#

1) Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany
2) Institute for Experimental Virology, TWINCORE Centre for Experimental and Clinical Infection Research, a Joint Venture between the Hannover Medical School (MHH) and the Helmholtz Centre for Infection Research (HZI), Hannover, Germany
3) Division of Virus-Associated Carcinogenesis (F170), German Cancer Research Center (DKFZ), Heidelberg, Germany
4) Department of Infectious Diseases, Molecular Virology, University of Heidelberg, Heidelberg, Germany
5) Berlin Institute of Health at Charité, Berlin, Germany
6) Institute for Biology, Humboldt-Universität zu Berlin, Berlin, Germany

corresponding authors

Corresponding authors: emanuel Wyler, Emanuel.wyler@mdc-berlin.de and Markus Landthaler, markus.landthaler@mdc-berlin.de

Abstract

Genomic material in wastewater provides a rich source of data for detection and surveillance of microbes. Used for decades to monitor poliovirus and other pathogens, the SARS-CoV-2 pandemic and the falling costs of high-throughput sequencing have substantially boosted the interest in and the usage of wastewater monitoring. We have longitudinally collected over 100 samples from a wastewater treatment plant in Berlin/Germany, from March 2021 to July 2022, in order to investigate three aspects. First, we conducted a full metagenomic analysis and exemplified the depth of the data by temporal tracking strains and to a certain extent also variants of human astroviruses and enteroviruses. Second, targeting respiratory pathogens, a broad enrichment panel enabled us to detect waves of RSV, influenza, or common cold coronaviruses in high agreement with clinical data. Third, by applying a profile Hidden Markov Model-based search for novel viruses, we identified more than 100 thousand novel transcript assemblies likely not belonging to known virus species, thus substantially expanding our knowledge of virus diversity. Taken together, we present a longitudinal and deep investigation of the viral genomic information in wastewater that underlines the value of sewage surveillance for both public health purposes and planetary virome research.

Introduction

Environmental surveillance of wastewater samples combined with high-throughput genomic sequencing can be a particularly valuable tool to monitor the vast diversity of microbes and anti-microbial resistance genes (Leifels et al., 2022; Santiago-Rodriguez, 2022). Accumulating genomic data from environmental samples can serve three purposes. First, it can be used as a sentinel system to monitor human pathogens and inform public health decisions (Diamond et al., 2022). Wastewater was and is essential for assessing Polio outbreaks (Anis et al., 2013; Chowdhary and Dhole, 2008; Ryerson et al., 2022). However, a range of other viruses can be detected, including gastroenteritis and hepatitis viruses, and influenza (Heijnen and Medema, 2011; Hellmer et al., 2014). In the SARS-CoV-2 pandemic, wastewater monitoring was established worldwide to monitor the virus in almost real time (Diamond et al., 2022).

The second use is to characterize microbial community ecosystems, which can be indicators for e.g. quality and function of freshwater systems (Numberger et al., 2022). And third, the advent of ultra-deep sequencing and computational methods, so far from public databases and seawater samples, has enabled the detection of large number of novel viruses, thus expanding considerably our knowledge of viral diversity (Edgar et al., 2022; Gregory et al., 2019; Martinez-Hernandez et al., 2022).

For monitoring human viruses in wastewater, various aspects need to be considered. This includes how much and in which form genomic material is being shed. Whereas in stool samples, for example, substantial amounts of infectious enteroviruses are found (Blomqvist and Roivainen, 2016), it is debated whether e.g. SARS-CoV-2 RNA in feces is present as part of intact viral particles or – what may be more prevalent – as fragments (Cerrada-Romero et al., 2022; Guo et al., 2021b). Facing the general higher stability of non-enveloped viruses (picornaviridae such as enteroviruses, caliciviruses or astroviruses) when compared to enveloped viruses (influenza, coronaviruses, RSV etc.) (Firquet et al., 2015), this can lead to substantial differences in the detection potential in addition to the incidence rates in the population. Overall, human pathogens are only a minor part of the totality of microbes in wastewater (Cantalupo et al., 2011; Wu et al., 2019), making their detection challenging. It is also important to note that distinct viral nucleic acids can be present in different fractions of wastewater samples. For example, both mpox and influenza genomic nucleic acids have been found to be associated with solids (Mercier et al., 2022; Wolfe et al., 2022b).

Previous studies have shown that wastewater is a useful source to detect circulating known as well as novel viruses (Adriaenssens et al., 2018; Bibby and Peccia, 2013; Cantalupo et al., 2011; Fernandez-Cassi et al., 2018; Guajardo-Leiva et al., 2020; Martinez-Puchol et al., 2021; Perez-Cataluna et al., 2021; Rothman et al., 2021). These studies explored several aspects, including enrichment of specific viruses (Martinez-Puchol et al., 2021; Rothman et al., 2021). Whereas most studies analyzed a small number of samples, a recent investigation collected 85 samples over a period of five months (Rothman et al., 2021), showcasing the relevance of longitudinal samplings.

In our work – although we all superkingdoms as well as antimicrobial resistance genes in the data, we focused our analysis on viruses –, we extend this body of work in several directions. First, we covered a time period of one and a half years with 116 samples in total, including a time period with reduced hygiene measurements compared to the height of the SARS-CoV-2 pandemic. The longitudinal sampling allows the observation of seasonal recurrence and monitoring of clinically relevant viruses such as respiratory syncytial virus (RSV). Second, very deep sequencing reveals subspecies of several viruses (exemplified with mamastroviruses and

enteroviruses), and, for astroviruses, also enables to track virus variant development over the sampling time period by assessing single point mutations. Third, the enrichment method applied here shows high concordance between clinical sampling and wastewater data for several respiratory viruses. And fourth, we substantially extend knowledge about viral diversity by identifying more than 100 thousand contigs belonging to previously unknown RNA and DNA viruses. In summary, our study shows that wastewater is an extraordinary rich source of nucleic acids to track known and novel viruses.

Methods

Sample collection

The sample collection was done as described previously (Schumann et al., 2022), or as following. Samples were collected (except for the first date, see supplementary table S1) from a single wastewater treatment plant in Berlin, Germany, as two hours composite samples (8–10 pm and 10–12 pm) at the primary influent collector, except for the indicated 24 hours composite samples on July 23, 2022. Berlin wastewater treatment plant effluents usually contain 500–1500 mg/L chemical oxygen demand, 200–600 mg/L suspended solids, 40–80 mg/L ammonium-N, 2–8 mg/L orthophosphate-P, 1500–2000 µS/cm electrical conductivity.

Sample processing and RNA isolation

The sample processing was essentially done as described previously (Schumann et al., 2022), with modifications for some samples. Samples were kept at four degrees, until processed about 12 h after collection. Processing was done along a published protocol (Jahn et al., 2022). First, the raw sample was filtered through 2 µm glass fiber and 0.2 µM PVDF filters (Millipore, cat# AP2007500 and S2GVU02RE). For the standard procedure, 60 ml filtrate were subsequently concentrated on a 10 kDa cutoff centricon unit (Millipore, cat# UFC701008), that was previously pre-conditioned with 50 mL ultrapure water centrifuged with 3000 g for 15 minutes at 4 °C, for 30 minutes at 3000 g/4 °C. The concentrate (about 300-450 µl) was mixed 1:3 with Trizol LS (ThermoFisher cat# 10296-010), and the RNA extracted using the DirectZol RNA kit (Zymo cat# R2052), including DNase treatment, and eluted in 50 µL ultrapure water according to the manufacturer's instruction. When using nanotrap beads (Ceres Nanosciences, cat# 44202), 10 ml filtered wastewater were mixed with 100 µl Enhancement Reagent 2, followed by addition of 150 µl beads, following extraction according to the manufacturer's protocol. RNA was subsequently extracted using either Trizol/DirectZol or the ZymoBIOMICS RNA/DNA combination extraction (Zymo, cat# ZYM-R2002).

Reverse transcription and quantitative PCRs

For the RT-qPCRs, 16 µl RNA were mixed with 4 µl LunaScript master mix (NEB cat# E3010L), according to the manufacturer's instructions, except for a 20 min incubation at 55 °C instead of 10 min. Afterwards, the cDNA was diluted 1:10 with ultrapure water, and 3.75 µL diluted cDNA used per qPCR reaction, using a SYBR green master mix (ThermoFisher cat# 43-643-46), with 250 nM final concentration of the primers listed in Supplementary Table S2.

Enrichment using ElementZero beads

SARS-CoV-2 and TBRV RNA, respectively, were enriched from total RNA using the SARS-CoV-2 / TBRV MagIC beads (ElementZero Biolabs) according to the manufacturer's protocol.

Total RNA sequencing

RNA sequencing were prepared using the SMARTer® Stranded Total RNA-Seq Kit v3 - Pico Input Mammalian (Takara cat# 634485) with 5 µl RNA from either the total RNA or the ElementZero eluate as starting material according to the manufacturer's protocoll. Libraries were then pooled and sequenced on a Novaseq 6000 device with 2x109 bp paired-end sequencing.

Enrichment using the xHYB adventitious agent panel

Sequencing libraries were pooled according to Supplementary Table S1 into three pools, with an approximate equal amount of starting material for every sample. The pools were then processed individually using the Qiagen xHYB adventitious agent panel (Qiagen cat# 333355) according to the manufacturer's protocol, and sequenced after the final re-ampflication step.

Metagenomic analysis using kaiju

Total RNA-seq data was analyzed by the kaiju program (Menzel et al., 2016). For both the application of kaiju pipeline as well as the subspecies analysis, the code will be detailed in the github repository. Specifically, the cd-hit-dup command from CD-HIT (Fu et al., 2012) was used to filter duplicated reads out, and kaiju command was incorporated to assign taxonomies to remaining reads from each 116 samples. Custom R scripts were used to summarize duplicated and assigned reads. Initial low count filtering was conducted by removing annotations with a maximum count of 10 across all samples. We have used Hellinger transformation (Nieuwenhuijse et al., 2020) to normalize the count table. Principal components analysis was performed on the normalized counts, and top PC1 and PC2 loadings were used to detect annotations that are most correlated with PC1 and PC2. Additional low count filtering removed annotations with mean count lower than 10 which revealed an additional set of annotations that are associated with outlying samples (i.e. outlying annotations). The following R packages were used for data processing and visualization: dplyr (Wickham, 2022b), ggplot2 (Wickham, 2022a), Complex heatmaps (Gu et al., 2016), pheatmap (Kolde, 2019), msa (Bodenhofer et al., 2015), reshape2.

Analysis of viral abundances and variants

Alignements were done using hisat2 (Kim et al., 2019), and read counting performed using samtools (Danecek et al., 2021).

Sequence contig assembly

Adapter sequences and low-quality bases were trimmed from the raw sequencing reads using fastp v0.23.2 (Danecek et al., 2021) with parameters '-q 20 --dedup'. The trimmed reads were assembled into scaffolds in paired-end mode using SPAdes v3.15.4 (Prjibelski et al., 2020) with default parameters. Each of the 116 RNA sequencing experiments was assembled separately.

Taxonomic classification of sequences

We extracted peptide sequences encoded by open reading frames (ORFs) of at least 300 nucleotides in length from the scaffolds using getorf from the EMBOSS package v6.6.0.0 (Rice et al., 2000). The peptide sequences were classified at the taxonomic ranks superkingdom, kingdom, phylum, subphylum, class, order, suborder, family, subfamily, genus, subgenus and species using the MMseqs2 taxonomy module v5f8735872e189991a743f7ed03e7c9d1f7a78855 (Hauser et al., 2016). We used the full nr database, downloaded in March 2022, for this analysis. Sequences classified as Bacteria, Eukaryota or Archaea at the superkingdom rank and the scaffold sequences they originated from were not considered further.

Discovery of viral sequences

We applied the following multi-stage process to identify and annotate viral sequences in the set of assembled scaffolds. We run a profile Hidden Markov Model (pHMM)-based sequence homology search against predicted peptide sequences encoded by ORFs of at least 300 nucleotides in length using hmmsearch from the HMMER v3.1b1 package (Eddy, 2011) in default mode. We used the following sets of pHMMs: the combined set of 84420 profiles from VirSorter 2 (Guo et al., 2021a), a set of 74 lineage major capsid protein (MCP) profiles of nucleocytoplasmic large DNA viruses (NCLDVs) (Schulz et al., 2020), five RNA-dependent RNA polymerase (RdRp) profiles of putative novel RNA virus phyla from the Tara Oceans Virome project (Zayed et al., 2022), 8390 profiles from the RNA Virus in MetaTranscriptomes (RVMT) project (Neri et al., 2022), 68 RdRp profiles from RdRp Scan v0.90 (Charon et al., 2022), and several in-house pHMMs of DNA and RNA virus proteins.

From the hits obtained during the pHMM searches we kept those sequences that were not classified as Eukaryota, Bacteria or Archaea at the superkingdom level by MMseqs2. To remove sequence redundancy, we clustered the scaffolds at 95% nucleotide sequence identity using MMseqs2 easy-linclust with parameter '--min-seq-id 0.95 -c 0.65 --cluster-mode 2'. To assess the genetic distance of these non-redundant, putatively viral sequences we run a DIAMOND blastx v2.0.13.151 (Buchfink et al., 2021) search with parameters '--ultra-sensitive –masking 0 -k 1 -f 6' against the following set of known viral sequences: 590872 viral proteins from NCBI RefSeq (O'Leary et al., 2016), 311725 RdRp sequences from the Serratus and PalmDB projects (Babaian and Edgar, 2021; Edgar et al., 2022), 49421 RdRp footprint sequences from the Tara Oceans Virome project (Zayed et al., 2022), 77510 RdRp sequences from RVMT (Neri et al., 2022) and 15081 RdRp sequences from RdRp Scan v0.90 (Charon et al., 2022).

Estimation of viral abundance

The trimmed sequencing reads were aligned in paired-end mode to the viral scaffolds using Bowtie 2 v2.3.4.1 (Langmead and Salzberg, 2012) with parameters '--no-unal -L 20 -N 1'. Samtools v1.10 (Danecek et al., 2021) was used to sort and index the resulting SAM/BAM files and to count the number of aligned reads per scaffold. Virus abundance was calculated as the total number of reads aligning to a scaffold across all sequencing libraries divided by the scaffold length. In comparisons between sequencing libraries, abundance was calculated as the number of reads of a particular library aligning to the viral scaffold divided by scaffold length divided by the total number of reads in the library. Abundance of a certain virus taxon (for

instance a virus family or order) was calculated as the sum of the abundance values of all scaffolds classified as belonging to this taxon.

Phylogenetic analysis of novel *Bunyavirales* sequences

We selected all scaffolds classified as *Bunyavirales* or the sister order *Articulavirales*. To reconfirm correct classification of these sequences, we conducted a pHMM search against profiles that we constructed based on 24 order-level RdRp alignments obtained from the Serratus project (Edgar et al., 2022) as well as in-house glycoprotein (GP) and nucleocapsid protein (NP) profiles of all 13 recognized virus families of the order *Bunyavirales*. We only kept sequences that showed the lowest E-value against either *Bunyavirales* or *Articulavirales* profiles in this search.

Putative RdRp protein sequences were aligned using MAFFT v7.310 (Katoh and Standley, 2013) with parameters '--localpair --maxiterate 1000 --reorder' followed by manual curation. For phylogenetic tree reconstruction, *Mononegavirales* RdRp sequences available at NCBI RefSeq, clustered at 20% amino acid sequence identity using MMseqs2, were added as an outgroup. We also added *Bunyavirales* and *Articulavirales* reference proteins, clustered at 90% amino acid sequence identity. In addition, we included bunya- und articulavirus sequences that we discovered in an independent screen of eukaryotic transcriptome projects in the Sequence Read Archive (SRA); details of the method are described here (Lauber et al., 2021) and a general introduction to this type of data-driven virus discovery approach is reviewed here (Lauber and Seitz, 2022). We only considered 16 out of more than 8000 potential bunya- or articulavirus-like contigs from our SRA search that covered a sufficient length of the RdRp and which showed at least 80% protein sequence identity to one of the bunya- or articulavirus sequences retrieved from the wastewater data. The fact that we re-discovered some of the viral sequences from the wastewater analysis in the SRA analysis provides independent support for the authenticity of the described novel viruses.

We reconstructed a Bayesian phylogenetic tree using BEAST v1.8.0 (Suchard et al., 2018) with the LG+G4+I substitution model, a relaxed molecular clock model with lognormally distributed rates and a Yule speciation tree prior. Two chains were run for 5 million generations and their convergence was verified using Tracer (Rambaut et al., 2018) after removing the first 10% of sampled trees as burn-in. The maximum clade credibility tree was visualized in R using ggtree v3.4.0 (Yu, 2020).

Results

Sample collection and processing

We aimed to perform a deep and longitudinal profiling of microbes and particularly viruses in wastewater. To this end, we collected raw influx samples between March 2021 and July 2022 from a wastewater treatment plant in Berlin/Germany, using a procedure that depletes intact bacteria by including a 0.2 µm filtration step (Fig. 1A). Of note, SARS-CoV-2 characterization in some of these samples has been described previously (Schumann et al., 2022). A full overview of the samples including Ct values from RT-qPCR assays on various viruses is shown in Supplementary Table S1. Following RNA isolation, we generated 116 high-throughput sequencing libraries, which yielded in total 1.45 billion 2x109 bp read pairs (12 million on average per sample), with about 10-30% of reads duplicated (Fig. S1A).

Most of the samples were processed using the standard procedure (Fig. 1A). In addition, some samples were processed using the Ceres nanotrap beads (see Supplementary Table S1 and method section for details). Except if indicated, for the following analysis only sequencing data from the samples with the standard procedure is used. Note that for some months, none or only one or two samples were successfully processed, henceforth they are omitted in some of the analysis shown below.

A wide range of eukaryotes, bacteria, and viruses can be detected, however only viruses show specific patterns

For an initial assessment of the diversity of the detected organisms in the sample, we applied the metagenomics pipeline kaiju (Menzel et al., 2016). As in previous wastewater metagenomics studies (Rothman et al., 2020), bacteria were more abundant despite their depletion by size-exclusion (Fig. S1B). Initially, kaiju annotated reads to 67322 taxonomies. Low count filtering on all samples preserved 16082 taxonomies. A principal components analysis (Fig. S1C) using these taxonomies revealed an additional set of 11984 taxonomies (Methods) that are detected only in very few (1-3) samples. These highly variable taxonomies had particularly high read counts in the outlier samples (Fig. S1D).

Analyzing the 16082 taxonomies, we found genera from all three superkingdoms and viruses, with some however clearly dominating the dataset (Fig. 1, Fig. S1E). Of note, a recently published metagenomic study of wastewater treatment plant influxes also in Berlin/Germany from an earlier time period, has also found Acinetobacter among the most abundant bacterial species (Numberger et al., 2022).

Furthermore, when comparing our dataset with a recently defined "baseline for global urban virome surveillance in sewage" (Nieuwenhuijse et al., 2020), mostly the same virus families were the most abundant ones, such as *Virgaviridae*, *Siphoviridae*, *Astroviridae*, *Myoviridae*, *Dicistroviridae*, *Podoviridae*, *Microviridae*, and *Picornaviridae* (Fig. S1F).

In order to investigate individual virus species, we quantified relative levels of viruses from seasonal food, a human pathogen previously identified to be abundant in wastewater, and a virus infecting the common house mosquito, present in Berlin only during the Northern hemisphere summer. For the grapevine fleck virus, and for the Daeseondong virus 2 (host *C. pipiens*), we saw the expected patterns with peaks only in late summer/fall. Mamastrovirus 1, a virus causing gastroenteritis in humans (Boujon et al., 2017), also showed a distinct pattern. Overall, most of the sequencing reads belonged to a limited number of genera without apparent temporal patterns that recapitulates previous waste water metagenomic studies. Single virus species however showed specific patterns, which we also took as an indication that the overall sampling and data analysis was appropriate.

Strain distribution and variant emergence of astroviruses

As observed in the data, astroviruses were among the most abundant human pathogens detected in our samples. Previous research has shown a substantial variety of astroviruses in wastewater samples (Tao et al., 2022; Yang et al., 2021). We therefore aimed to quantify the temporal distribution of different astrovirus subtypes within the investigated timeframe, and to determine the changes in the mutation patterns. As a starting point, we used the taxonomy IDs within the *Astroviridae* family detected by the kaiju metagenomics pipeline (Fig. 2A). For every taxonomy, we selected those sequences (i.e. accession numbers) that matched the collected

sequencing data best, as determined by the number of mapped sequencing reads. These sequences were then grouped manually according to their phylogenetic tree (Figure S2A), and one or two sequences per group were selected. Subsequently, point mutations/InDels in the the sequences were then corrected based on the pooled sequencing data to yield the final set, which was again displayed as a phylogenetic tree (Fig. 2B).

We quantified the number of sequences found for the respective accession group per month (Fig. 2C). This analysis showed that the different strains peaked at distinct periods within the investigated time frame. Whereas MLB1 was present throughout the analyzed period, Astrovirus 1 peaked around November/December 2021. Astroviruses VA1 and 4 on the other side emerged only towards the end of the investigation, starting in April 2022.

Since we could obtain full genome coverages over many months for many astroviruses, we were able to quantify mutations per month. As an example, we found variabilities at around 100 positions throughout the Astrovirus 1 (MN510439.1) genome (Fig. S2B). Clustering by variability showed again distinct temporal profiles. A subset of mutated positions is shown in Figure 4D, with e.g. a group of mutations disappearing in fall 2021 (positions 3345, 3390), being only present in fall 2021 (positions 1990, 2461), or emerging in June/July 2022 (positions 1366, 3435).

In addition to profiling known astroviruses, we set out to discover novel ones as well. To do so, all RNA sequencing samples were individually assembled into transcripts using SPAdes (Prjibelski et al., 2020). Next, we searched for assembled transcripts that showed significant sequence similarity to members of the *Astroviridae* family. In total, we found five that could potentially represent novel viruses. Two of them were outside the genera of *Avastrovirus* and *Mamastrovirus*, rather pointing to non-vertebrate hosts (Fig. S2C, upper part). For the three other ones, partial genomes were identified that were grouped within clusters of astroviruses infecting mammals, indicating that some of them could represent novel human pathogens.

Temporal dynamics of enteroviruses

Next, we analyzed the highly diverse genus of enteroviruses in the sequencing data. This genus contains rhinoviruses, coxsackieviruses, echoviruses, and polioviruses, which can cause a wide range of symptoms, including respiratory illness, meningitis, rash ("hand, foot, and mouth disease"), or paralysis (Harvala et al., 2018). The most important route of transmission is likely via fomites. We followed the same analysis path as for the astroviruses, however we did not apply the genome correction step due to overall low coverage. In general, we observed the expected seasonality (Keeren et al., 2021) with highest signals in northern hemisphere summer months (Fig. 4E). Signal levels were overall lower in the summer of 2021, which could be due hygiene measurements during the SARS-CoV-2 pandemic. Relative abundances were nevertheless distinct, with e.g. subtypes A19 and C113 being mainly present in 2021, and A9, A16, B2 and B5 mainly in 2022.

Enrichment of specific virus sequences from wastewater

In the previous sections, we focused on viruses that could be readily detected in the total RNA sequencing. However, the signal from many clinically relevant, particularly respiratory viruses was very low or absent. We therefore resorted to a commercial enrichment system (xHYB adventitious agent panel) that can be used to enrich 280 genomes/genome segments from 132 different viruses (see Fig. S3A for a list of viruses). The sequencing libraries generated from

total RNA were merged into three pools, as indicated in Table S1. The three pools were then subjected to enrichment, re-amplification, and sequencing. Normalized read counts for all viruses are shown in Fig. S3A, and for a subset with minimal thresholds in Fig. 3. Next to e.g. astroviruses detailed before, this analysis now also captures respiratory viruses, for which the amount of detectable material in the wastewater is low or absent before enrichment (Fig. S3B). This includes respiratory syncytial virus (RSV), influenza, or the common cold coronaviruses NL63, 229E, HKU1, and OC43 (Fig. 3A). We used data from the German Clinical Virology Network (Adams, 2022) in order to relate these observations to clinical diagnostics. Test positivity is shown for a range of respiratory and gastrointestinal viruses (Fig. 3B). A side-by-side comparison for incidence waves of RSV (October/November 2021), Rotavirus A (April-June 2022) or HKU1 (April to June 2022) and NL63 (April 2021 and April 2022), shows the good correlation of data from wastewater and individual testing (Fig. 3C).

The xHYB system enriches on the level of dsDNA, i.e. after the fragmentation steps that are part of the sequencing library preparation protocol. We therefore also tested a procedure that enriches on the level of RNA, starting from the total wastewater RNA. For that purpose, we applied the ElementZero system separately for SARS-CoV-2 and tomato brown mosaic virus, and analyzed the enrichment both using RT-qPCR and high-throughput sequencing. The RT-qPCR showed strong depletion of PMMV RNA in all enrichment reactions (about 1000-fold loss, i.e. 10 PCR cycles), and about two- to four-fold loss of the RNA in question, indicating a very strong enrichment (Fig. S3C, upper part). This was corroborated in the sequencing data (Fig. S3C, lower part). Interestingly, when investigating the coverage profiles, we found that SARS-CoV-2 showed coverage “islands” in the region covered by hybridization probes, whereas for TBRV the coverage was as equally distributed as for the total RNA, i.e. the input of the ElementZero enrichment (Fig. S3D). An explanation could be that the SARS-CoV-2 RNA is already fragmented in the wastewater, since barely any signal outside of the probe regions was recovered. In contrast, the TBRV RNA would be intact, as the recovery was independent of the distance from the probes.

Detection of novel viruses from wastewater

Following characterization of known viruses, we set out to search for novel viruses that contain known amino acid sequence motifs, such as the one for the RNA-dependent RNA polymerase. We identified in total 417,972 contigs with sequence similarity to a comprehensive set of known DNA and RNA virus sequences (see Methods for details). Viruses from 49 known und five unclassified (uc) virus orders were discovered. The identified viral sequences were dominated, in terms of number of non-redundant contigs (Fig. 4A, E) and abundance (Fig. 4B, F), by members of the orders *Caudovirales* (DNA viruses), *Norzivirales* and *Levivirales* (RNA viruses), which infect microbes and likely represent bacteriophages. Many of the contigs were short, with 49,329 (11.8%) being longer than 1000 nt (Fig. 4C, G). Due to the considerable degree of sequence fragmentation we cannot exclude that the actual number of viruses is lower than the number of reported viral contigs as a virus might contribute with several sequence fragments. The majority of 277,092 of the viral contigs (66.3%) showed a protein sequence identity to the closest known virus of less than 90% (Fig. 4D, H), indicating that these sequences originated from novel viruses.

Among the viral sequences, 107 were novel, non-redundant sequences that were classified as *Bunyavirales* or its sister order *Articulavirales*. These included sequences that prototype distant lineages within the *Bunyavirales* in the L protein-based phylogeny (Fig. 4I), suggesting that

some of the newly discovered viruses form novel virus families. The majority of the contigs were identified via sequence homology to L proteins of known bunya- and articulaviruses while others gave hits against nucleocapsid proteins or glycoproteins of certain bunyavirus families. To provide additional and independent evidence that the discovered viruses are genuine, we retrieved 16 bunyavirus-like sequences that shared 80% or more protein sequence identity to the viruses identified from wastewater and which we discovered in a large screen of public transcriptome projects from the Sequence Read Archive (SRA) (Fig. 4L). The fact that we re-discovered some of the viral sequences from the wastewater analysis in the SRA analysis provides independent support for the authenticity of the described novel viruses. An amino acid sequence alignment of motifs A, B and C of the RNA dependent RNA polymerase shows conserved regions among the subfamilies to which the novel viruses belong (Fig. 4J). Finally, we probed the amount of five of the most abundant novel virus contigs across the entire sampling time course. Three of the five were detected in only one sample each, one over the entire time course, and the fifth was restricted to the period from April to May 2022 (Fig. S4A).

Discussion

In this study, we present a deep longitudinal profiling of the wastewater virome from a treatment plant in Berlin, Germany. Overall, we analyzed samples covering a period of 17 months, from March 2021 to July 2022. Of note, at the beginning of the time series, there were still substantial measures for mitigation of viral spread in place due to the SARS-CoV-2 pandemic, which very likely also affected a range of pathogens. From May 2021 onwards, there was a mask mandate in place, however no closures or shutdown measures. The data therefore may differ from a situation without such measures, as before the pandemic.

Nevertheless, we could track a wide variety of microbes, including a broad range of human viruses. Previous wastewater metagenomics studies (Adriaenssens et al., 2018; Bibby and Peccia, 2013; Cantalupo et al., 2011; Fernandez-Cassi et al., 2018; Guajardo-Leiva et al., 2020; Martinez-Puchol et al., 2021; Perez-Cataluna et al., 2021; Rothman et al., 2020; Rothman et al., 2021) have assessed the scope of taxa to be found in this sample type. These findings, such as the dominance of bacterial sequences as well as the high abundance of plant viruses were recapitulated in our data. In extension of this work, we present several findings. First, the relatively broad sampling period allowed the detection of non-human seasonal viruses such as those in seasonal food (grapevines, watermelons) as well as in mosquitos, which are in Berlin present only during Northern hemisphere summer (Fig. 1C). Such findings provide some degree of insurance for the seasonality in our data set, but also hint to the breadth of the information that wastewater can contain, to allow monitoring of entire ecosystems, and not merely specific pathogens. Second, our detailed and temporal investigations of astrovirus and enterovirus subspecies, as well as the temporal variant dynamics exemplified in mamastrovirus 1 (Fig. 2) shows the depth of information which can be recovered by wastewater monitoring. Third, although several clinically relevant respiratory viruses are – not surprisingly, in contrast to gastroenteritis viruses – much less abundant, they can be recovered in high-throughput sequencing using targeted enrichment approaches (Fig. 3). Comparison with clinical testing obtained from individual patients shows high agreement (Fig. 3C), underlining that wastewater can, with appropriate methodology, possibly be used to track every circulating human pathogen. An interesting observation was that enrichment on the RNA level recovered only highly fragmented genome coverage from the enveloped virus SARS-CoV-2. This finding indicates that at least this RNA, and maybe RNA from enveloped viruses in general, is present not as part of

viral particles but rather bound to protein proteins or other matter (Fig. S3). Along with observations such as the higher abundance of mpox and influenza DNA/RNA in the particle fractions (Wolfe et al., 2022a; Wolfe et al., 2022b), this underscores the need for systematic investigations into which processing methods are suitable for the detection of a specific microbe. And fourth, our discovery of possibly tens of thousands novel viruses shows that wastewater can also fill major gaps in our knowledge of the planetary virome and thus can inform future assessment of zoonotic potentials.

Our study has several limitations. We have only investigated one wastewater treatment plant at one location, albeit longitudinally. Together with increasing deep wastewater metagenomic data, global comparisons will become possible. Furthermore, our processing method may capture a wide array of microbes, however it certainly misses others. For example, despite relatively high local incidences, we have not been able to detect mpox genomes in our samples in contrast to other sites (Wannigama et al., 2023; Wolfe et al., 2022b). Also, temperature differences (in Berlin, wastewater is in the range between 15 °C in winter and 25 °C) as well as the flow time from the households to the treatment plant (in the range of one to several hours) can introduce variability that is difficult to track. And finally, although our sampling series is to our knowledge the longest published so far, it still captures seasons only once or twice, and in parts was done during an exceptional period of pandemic mitigation measures. With still decreasing costs for high-throughput sequencing along with refined experimental and computational methodology however, future studies will be able to capture microbial communities at a so far unfathomable level of detail.

Acknowledgements

We are very grateful to Fredrik Zietzschmann, Katharina Flatau, Regina Gnirss and Uta Böckelmann from the municipal water authority (Berliner Wasserbetriebe) for providing samples and continuous support. Figure 1A was generously created and provided by Thomas Zahn, Peter Pennitz and Martin Witzenrath. We thank Sindy Böttcher and Julian Kreibich for support particularly with enteroviruses, as well as René Kallies, Antonis Chatzinotas, Hans-Christoph Selinka and Martin Meixner for support and helpful discussions. EW, CL, GTA, SS, and ML are supported by the Project "Virological and immunological determinants of COVID-19 pathogenesis – lessons to get prepared for future pandemics (KA1-Co-02 'COVIPA')", a grant from the Helmholtz Association Initiative and Networking Fund. CL acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2155 - project number 390874280. CL is a member of the European Virus Bioinformatics Center (EVBC).

Author contributions

EW, AD, CQ and GTA performed experiments. EW, CL, AM, SS and ML analyzed data and prepared figures. EW wrote the manuscript, with contributions from CL. TB, JA and SS supervised parts of the projects, ML supervised the entire project. All authors approved the manuscript.

Data and code availability

All raw sequencing data will be made available via NCBI GEO, and code on github.com, shortly. Meanwhile, please contact emanuel.wyler@mdc-berlin.de for access.

Competing Interest Statement

The authors have no competing interests to declare.

References

Adams, A.G., B.; Kaiser, R.; Prifert, C.; Schmeisser, N. (2022). Respiratory Virus Network.
Adriaenssens, E.M., Farkas, K., Harrison, C., Jones, D.L., Allison, H.E., and McCarthy, A.J. (2018). Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. mSystems *3*.
Anis, E., Kopel, E., Singer, S.R., Kaliner, E., Moerman, L., Moran-Gilad, J., Sofer, D., Manor, Y., Shulman, L.M., Mendelson, E.*, et al.* (2013). Insidious reintroduction of wild poliovirus into Israel, 2013. Euro Surveill *18*.
Babaian, A., and Edgar, R.C. (2021). Ribovirus classification by a polymerase barcode sequence. bioRxiv, 2021.2003.2002.433648.
Bibby, K., and Peccia, J. (2013). Identification of viral pathogen diversity in sewage sludge by metagenome analysis. Environ Sci Technol *47*, 1945-1951.
Blomqvist, S., and Roivainen, M. (2016). Isolation and Characterization of Enteroviruses from Clinical Samples. Methods Mol Biol *1387*, 19-28.
Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C., and Hochreiter, S. (2015). msa: an R package for multiple sequence alignment. Bioinformatics *31*, 3997-3999.
Boujon, C.L., Koch, M.C., and Seuberlich, T. (2017). The Expanding Field of Mammalian Astroviruses: Opportunities and Challenges in Clinical Virology. Adv Virus Res *99*, 109-137.
Buchfink, B., Reuter, K., and Drost, H.G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods *18*, 366-368.
Cantalupo, P.G., Calgua, B., Zhao, G., Hundesa, A., Wier, A.D., Katz, J.P., Grabe, M., Hendrix, R.W., Girones, R., Wang, D.*, et al.* (2011). Raw sewage harbors diverse viral populations. mBio *2*.
Cerrada-Romero, C., Berastegui-Cabrera, J., Camacho-Martinez, P., Goikoetxea-Aguirre, J., Perez-Palacios, P., Santibanez, S., Jose Blanco-Vidal, M., Valiente, A., Alba, J., Rodriguez-Alvarez, R.*, et al.* (2022). Excretion and viability of SARS-CoV-2 in feces and its association with the clinical outcome of COVID-19. Sci Rep *12*, 7397.
Charon, J., Buchmann, J.P., Sadiq, S., and Holmes, E.C. (2022). RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. Virus Evolution *8*, veac082.
Chowdhary, R., and Dhole, T.N. (2008). Interrupting wild poliovirus transmission using oral poliovirus vaccine: environmental surveillance in high-risks area of India. J Med Virol *80*, 1477-1488.
Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M.*, et al.* (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*.
Diamond, M.B., Keshaviah, A., Bento, A.I., Conroy-Ben, O., Driver, E.M., Ensor, K.B., Halden, R.U., Hopkins, L.P., Kuhn, K.G., Moe, C.L.*, et al.* (2022). Wastewater surveillance of pathogens can inform public health responses. Nat Med *28*, 1992-1995.
Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.
Edgar, R.C., Taylor, J., Lin, V., Altman, T., Barbera, P., Meleshko, D., Lohr, D., Novakovsky, G., Buchfink, B., Al-Shayeb, B.*, et al.* (2022). Petabase-scale sequence alignment catalyses viral discovery. Nature *602*, 142-147.

Fernandez-Cassi, X., Timoneda, N., Martinez-Puchol, S., Rusinol, M., Rodriguez-Manzano, J., Figuerola, N., Bofill-Mas, S., Abril, J.F., and Girones, R. (2018). Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. Sci Total Environ *618*, 870-880.
Firquet, S., Beaujard, S., Lobert, P.E., Sane, F., Caloone, D., Izard, D., and Hober, D. (2015). Survival of Enveloped and Non-Enveloped Viruses on Inanimate Surfaces. Microbes Environ *30*, 140-144.
Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150-3152.
Gregory, A.C., Zayed, A.A., Conceicao-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C.*, et al.* (2019). Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell *177*, 1109-1123 e1114.
Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847-2849.
Guajardo-Leiva, S., Chnaiderman, J., Gaggero, A., and Diez, B. (2020). Metagenomic Insights into the Sewage RNA Virosphere of a Large City. Viruses *12*.
Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitua, M.C., Vik, D., Sullivan, M.B.*, et al.* (2021a). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 37.
Guo, M., Tao, W., Flavell, R.A., and Zhu, S. (2021b). Potential intestinal infection and faecal-oral transmission of SARS-CoV-2. Nat Rev Gastroenterol Hepatol *18*, 269-283.
Harvala, H., Broberg, E., Benschop, K., Berginc, N., Ladhani, S., Susi, P., Christiansen, C., McKenna, J., Allen, D., Makiello, P.*, et al.* (2018). Recommendations for enterovirus diagnostics and characterisation within and beyond Europe. J Clin Virol *101*, 11-17.
Hauser, M., Steinegger, M., and Soding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics *32*, 1323-1330.
Heijnen, L., and Medema, G. (2011). Surveillance of influenza A and the pandemic influenza A (H1N1) 2009 in sewage and surface water in the Netherlands. J Water Health *9*, 434-442.
Hellmer, M., Paxeus, N., Magnius, L., Enache, L., Arnholm, B., Johansson, A., Bergstrom, T., and Norder, H. (2014). Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. Appl Environ Microbiol *80*, 6771-6781.
Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., Banziger, C., Devaux, A.J., Stachler, E., Caduff, L.*, et al.* (2022). Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. Nat Microbiol *7*, 1151-1160.
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.
Keeren, K., Bottcher, S., and Diedrich, S. (2021). Enterovirus Surveillance (EVSurv) in Germany. Microorganisms *9*.
Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907-915.
Kolde, R. (2019). pheatmap: Pretty Heatmaps.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.
Lauber, C., and Seitz, S. (2022). Opportunities and Challenges of Data-Driven Virus Discovery. Biomolecules *12*.
Lauber, C., Vaas, J., Klingler, F., Mutz, P., Gorbalenya, A.E., Bartenschlager, R., and Seitz, S. (2021). Deep mining of the Sequence Read Archive reveals bipartite coronavirus genomes and inter-family Spike glycoprotein recombination. bioRxiv, 2021.2010.2020.465146.
Leifels, M., Khalilur Rahman, O., Sam, I.C., Cheng, D., Chua, F.J.D., Nainani, D., Kim, S.Y., Ng, W.J., Kwok, W.C., Sirikanchana, K.*, et al.* (2022). The one health perspective to improve environmental surveillance of zoonotic viruses: lessons from COVID-19 and outlook beyond. ISME Commun *2*, 107.
Martinez-Hernandez, F., Fornas, O., and Martinez-Garcia, M. (2022). Into the Dark: Exploring the Deep Ocean with Single-Virus Genomics. Viruses *14*.

Martinez-Puchol, S., Itarte, M., Rusinol, M., Fores, E., Mejias-Molina, C., Andres, C., Anton, A., Quer, J., Abril, J.F., Girones, R., *et al.* (2021). Exploring the diversity of coronavirus in sewage during COVID-19 pandemic: Don't miss the forest for the trees. Sci Total Environ *800*, 149562.

Menzel, P., Ng, K.L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun *7*, 11257.

Mercier, E., D'Aoust, P.M., Thakali, O., Hegazy, N., Jia, J.J., Zhang, Z., Eid, W., Plaza-Diaz, J., Kabir, M.P., Fang, W., *et al.* (2022). Municipal and neighbourhood level wastewater surveillance and subtyping of an influenza virus outbreak. Sci Rep *12*, 15777.

Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D., Chen, I.M., Ivanova, N., Zeigler Allen, L., Paez-Espino, D., *et al.* (2022). Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell *185*, 4023-4037 e4018.

Nieuwenhuijse, D.F., Oude Munnink, B.B., Phan, M.V.T., Global Sewage Surveillance project, c., Munk, P., Venkatakrishnan, S., Aarestrup, F.M., Cotten, M., and Koopmans, M.P.G. (2020). Setting a baseline for global urban virome surveillance in sewage. Sci Rep *10*, 13748.

Numberger, D., Zoccarato, L., Woodhouse, J., Ganzert, L., Sauer, S., Marquez, J.R.G., Domisch, S., Grossart, H.P., and Greenwood, A.D. (2022). Urbanization promotes specific bacteria in freshwater microbiomes including potential pathogens. Sci Total Environ *845*, 157321.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-745.

Perez-Cataluna, A., Cuevas-Ferrando, E., Randazzo, W., and Sanchez, G. (2021). Bias of library preparation for virome characterization in untreated and treated wastewaters. Sci Total Environ *767*, 144589.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics *70*, e102.

Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst Biol *67*, 901-904.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet *16*, 276-277.

Rothman, J.A., Loveless, T.B., Griffith, M.L., Steele, J.A., Griffith, J.F., and Whiteson, K.L. (2020). Metagenomics of Wastewater Influent from Southern California Wastewater Treatment Facilities in the Era of COVID-19. Microbiol Resour Announc *9*.

Rothman, J.A., Loveless, T.B., Kapcia, J., 3rd, Adams, E.D., Steele, J.A., Zimmer-Faust, A.G., Langlois, K., Wanless, D., Griffith, M., Mao, L., *et al.* (2021). RNA Viromics of Southern California Wastewater and Detection of SARS-CoV-2 Single-Nucleotide Variants. Appl Environ Microbiol *87*, e0144821.

Ryerson, A.B., Lang, D., Alazawi, M.A., Neyra, M., Hill, D.T., St George, K., Fuschino, M., Lutterloh, E., Backenson, B., Rulli, S., *et al.* (2022). Wastewater Testing and Detection of Poliovirus Type 2 Genetically Linked to Virus Isolated from a Paralytic Polio Case - New York, March 9-October 11, 2022. MMWR Morb Mortal Wkly Rep *71*, 1418-1424.

Santiago-Rodriguez, T.M. (2022). The Detection of SARS-CoV-2 in the Environment: Lessons from Wastewater. In Water.

Schulz, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D.A., Denef, V.J., McMahon, K.D., Konstantinidis, K.T., Eloe-Fadrosh, E.A., Kyrpides, N.C., *et al.* (2020). Giant virus diversity and host interactions through global metagenomics. Nature *578*, 432-436.

Schumann, V.F., de Castro Cuadrat, R.R., Wyler, E., Wurmus, R., Deter, A., Quedenau, C., Dohmen, J., Faxel, M., Borodina, T., Blume, A., *et al.* (2022). SARS-CoV-2 infection dynamics revealed by wastewater sequencing analysis and deconvolution. Sci Total Environ *853*, 158931.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol *4*, vey016.

Tao, Z., Lin, X., Liu, Y., Ji, F., Wang, S., Xiong, P., Zhang, L., Xu, Q., Xu, A., and Cui, N. (2022). Detection of multiple human astroviruses in sewage by next generation sequencing. Water Res *218*, 118523.

Wannigama, D.L., Amarasiri, M., Hongsing, P., Hurst, C., Modchang, C., Chadsuthi, S., Anupong, S., Phattharapornjaroen, P., S, M.A., Fernandez, S.*, et al.* (2023). Multiple traces of monkeypox detected in non-sewered wastewater with sparse sampling from a densely populated metropolitan area in Asia. Sci Total Environ *858*, 159816.

Wickham, H. (2022a). ggplot2: Elegant Graphics for Data Analysis.

Wickham, H.F.R.H.L.M.K. (2022b). dplyr: A Grammar of Data Manipulation.

Wolfe, M.K., Duong, D., Bakker, K.M., Ammerman, M., Mortenson, L., Hughes, B., Arts, P., Lauring, A.S., Fitzsimmons, W.J., Bendall, E.*, et al.* (2022a). Wastewater-Based Detection of Two Influenza Outbreaks. Environmental Science & Technology Letters *9*, 687-692.

Wolfe, M.K., Yu, A.T., Duong, D., Rane, M.S., Hughes, B., Chan-Herur, V., Donnelly, M., Chai, S., White, B.J., Vugia, D.J.*, et al.* (2022b). Wastewater Surveillance for Monkeypox Virus in Nine California Communities. medRxiv, 2022.2009.2006.22279312.

Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., Zhang, Q., Brown, M.R., Li, Z., Van Nostrand, J.D.*, et al.* (2019). Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nat Microbiol *4*, 1183-1195.

Yang, Q., Rivailler, P., Zhu, S., Yan, D., Xie, N., Tang, H., Zhang, Y., and Xu, W. (2021). Detection of multiple viruses potentially infecting humans in sewage water from Xinjiang Uygur Autonomous Region, China. Sci Total Environ *754*, 142322.

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. Curr Protoc Bioinformatics *69*, e96.

Zayed, A.A., Wainaina, J.M., Dominguez-Huerta, G., Pelletier, E., Guo, J., Mohssen, M., Tian, F., Pratama, A.A., Bolduc, B., Zablocki, O.*, et al.* (2022). Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. Science *376*, 156-162.

Figure 1

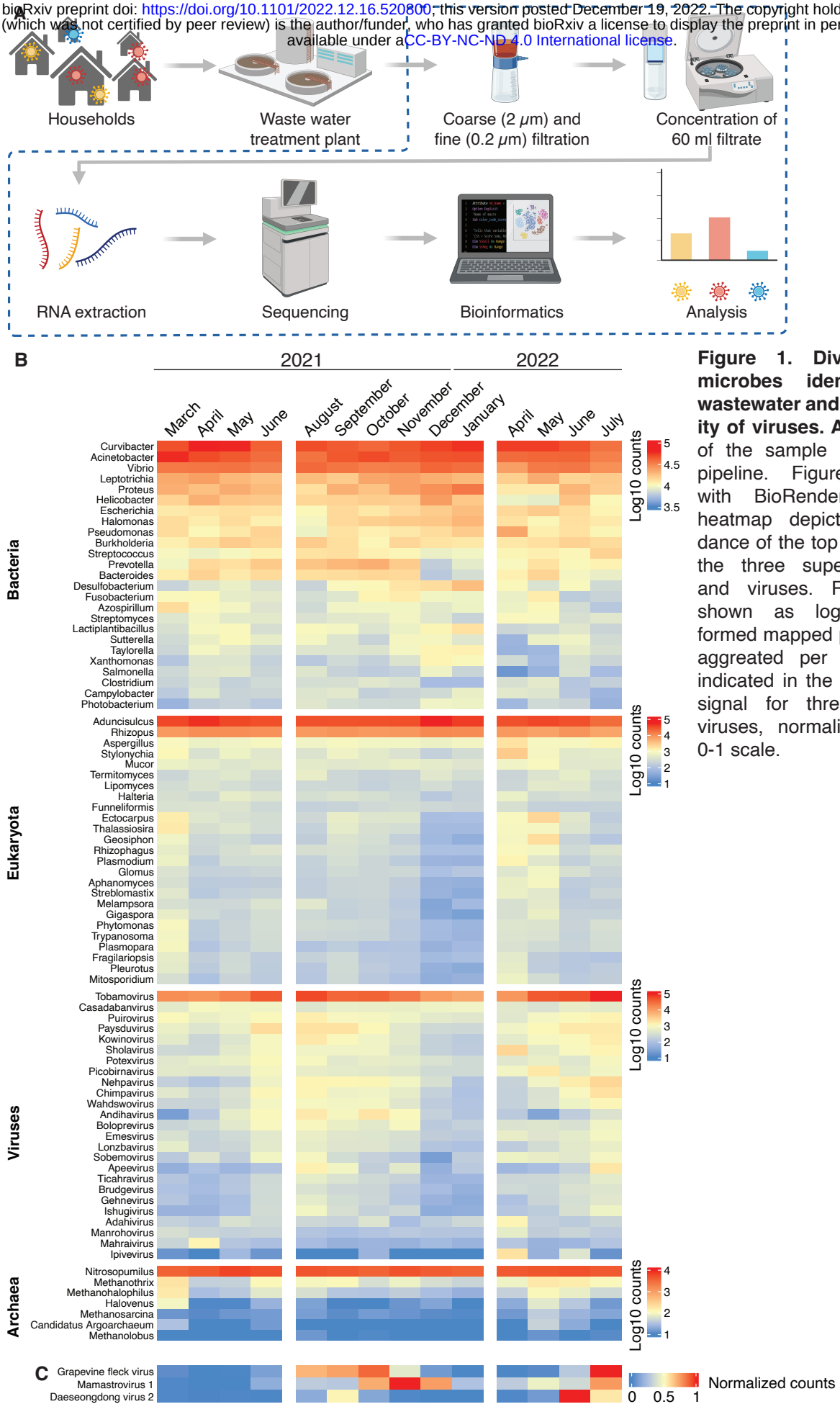

Figure 1. Diversity of microbes identified in wastewater and seasonality of viruses. A, overview of the sample processing pipeline. Figure created with BioRender.com. **B,** heatmap depicting abundance of the top genera for the three superkingdoms and viruses. Reads are shown as log10 transformed mapped per million, aggreated per month as indicated in the bottom. **C,** signal for three specific viruses, normalized on a 0-1 scale.

Figure 2

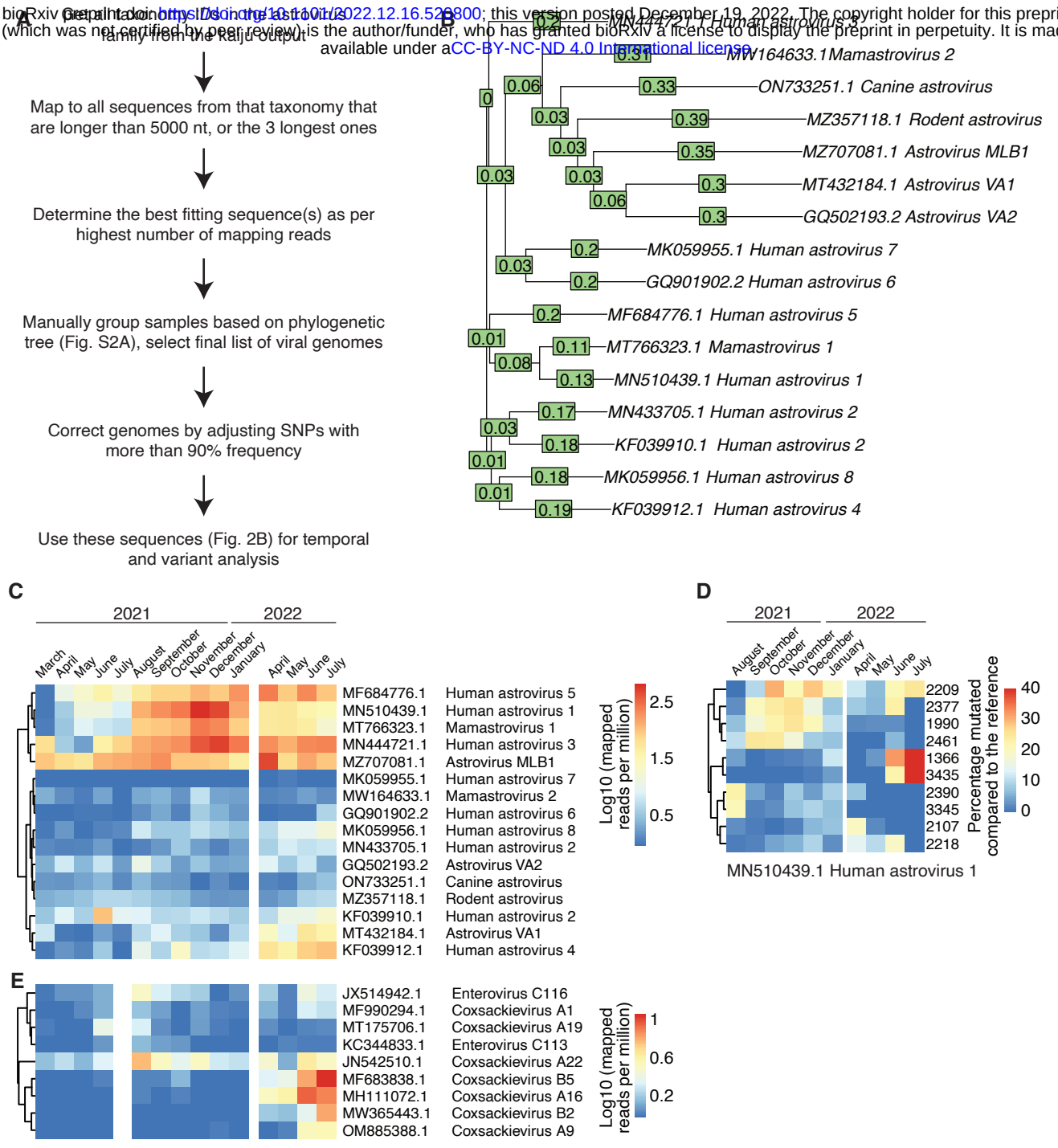

Figure 2. Temporal subspecies/variant dynamics of astroviruses and enteroviruses. A, analysis scheme for human astroviruses. **B,** phylogenetic tree of the astroviruses analyzed in detail. **C,** heatmap depicting abundance of the astroviruses over time. Reads are shown as log10 transformed mapped per million, aggreated per month as indicated in the bottom. **D,** for selected position in the human astrovirus 1 genome (based on accession number MN510439.1), the frequency of a mutated residue is shown as a heatmap. **E,** as for C, but for enteroviruses.

Figure 3

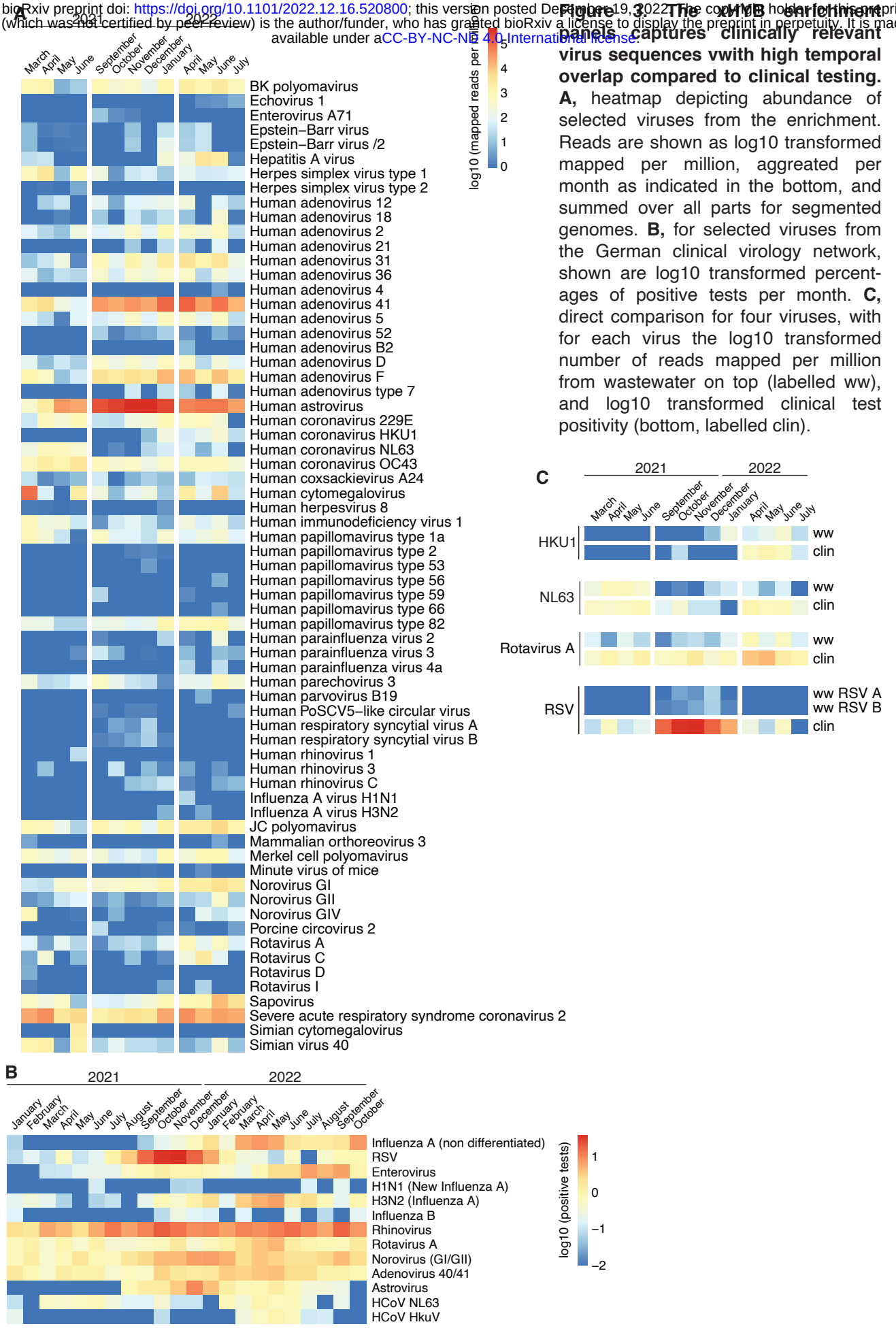

Figure 3. The xHYB enrichment panels captures clinically relevant virus sequences vwith high temporal overlap compared to clinical testing. **A,** heatmap depicting abundance of selected viruses from the enrichment. Reads are shown as log10 transformed mapped per million, aggreated per month as indicated in the bottom, and summed over all parts for segmented genomes. **B,** for selected viruses from the German clinical virology network, shown are log10 transformed percentages of positive tests per month. **C,** direct comparison for four viruses, with for each virus the log10 transformed number of reads mapped per million from wastewater on top (labelled ww), and log10 transformed clinical test positivity (bottom, labelled clin).

Figure 4

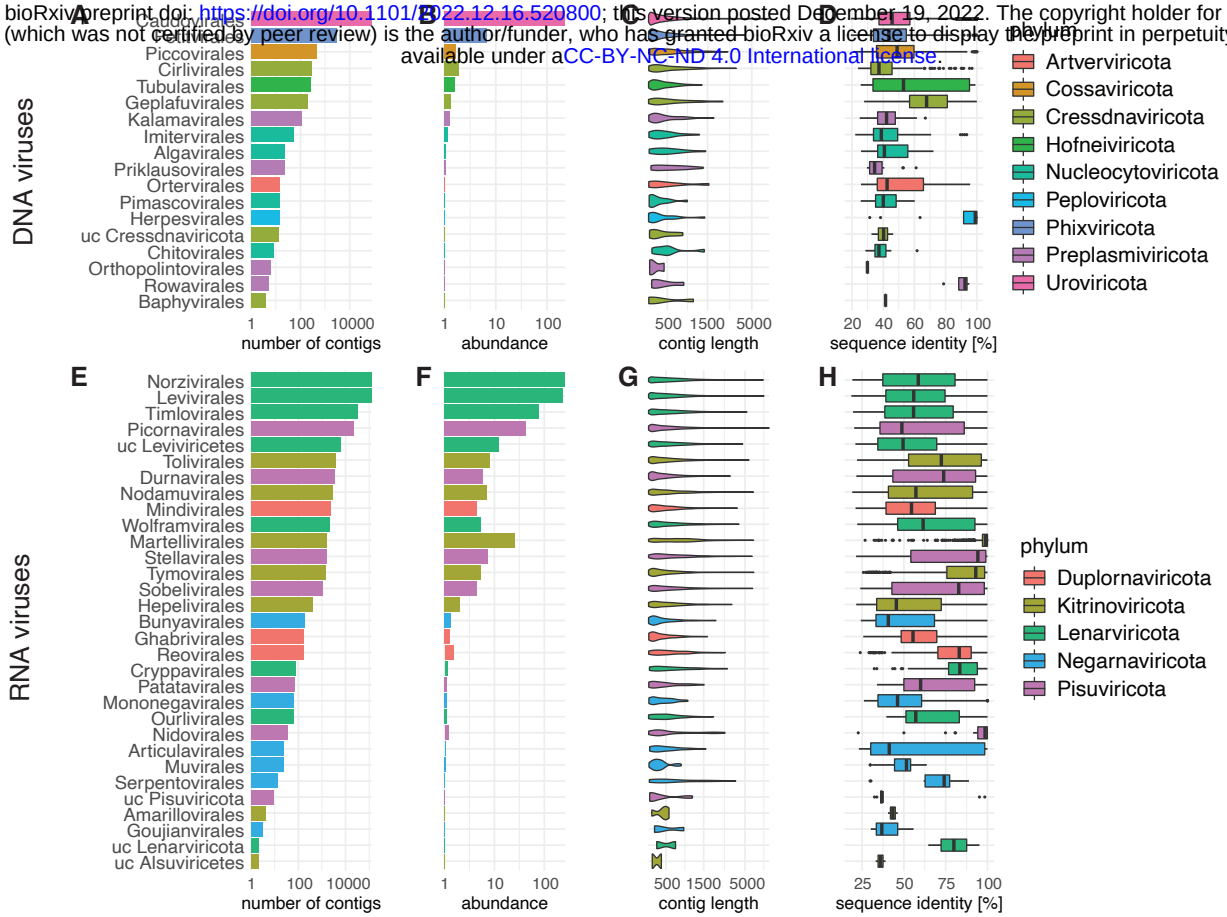

Figure 4. Novel viral sequences discovered from wastewater. A, E, Number of non-redundant novel DNA viral (top row) and RNA viral (bottom row) contigs identified in all waste water samples together, aggregated by order and colored by phylum. **B, F,** Viral abundances. **C, G,** Lengths of the contigs within the order labelled on the left. **D, H**, Protein sequence identities to the closest known reference virus for the orders labelled on the left. **I,** L protein-based Bayesian phylogeny of known and novel bunya- and articulaviruses. Bunyavirus families, the Articulavirales order and the outgroup Mononegavirales order (black) are discriminated by branch color. Viruses discovered from wastewater and the SRA screen are in cyan and purple, respectively. **J,** Multiple sequence alignment of the region around most conserved motifs A, B and C of the viral RdRp for a selected set of novel and reference viruses. Two catalytic residues of motifs A and C and one residue of motif B involved in nucleotide selection are marked by diamond symbols.