Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Full length article

Pathogen dynamics and discovery of novel viruses and enzymes by deep nucleic acid sequencing of wastewater

Emanuel Wyler[a], Chris Lauber[b,c], Artür Manukyan[a], Aylina Deter[a], Claudia Quedenau[a], Luiz Gustavo Teixeira Alves[a], Claudia Wylezich[d], Tatiana Borodina[a], Stefan Seitz[e,f], Janine Altmüller[a,g], Markus Landthaler[a,h,*]

[a] Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany
[b] 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
[c] Cluster of Excellence RESIST (EXC 2155), Hannover Medical School, Hannover, Germany
[d] Institute of Diagnostic Virology, Friedrich-Loeffler-Institut, Federal Research Institute for Animal Health, Greifswald-Insel Riems, Germany
[e] Division of Virus-Associated Carcinogenesis (F170), German Cancer Research Center (DKFZ), Heidelberg, Germany
[f] Department of Infectious Diseases, Molecular Virology, University of Heidelberg, Heidelberg, Germany
[g] Berlin Institute of Health at Charité, Berlin, Germany
[h] Institut für Biologie, Humboldt-Universität zu Berlin, Berlin, Germany

ARTICLE INFO

Handling Editor: Thanh Nguyen

Keywords:
Wastewater
Nucleic acids
Metagenomics
Human pathogens
Novel viruses

ABSTRACT

Wastewater contains an extensive reservoir of genetic information, yet largely unexplored. Here, we analyzed by high-throughput sequencing total nucleic acids extracted from wastewater samples collected during a 17 month-period in Berlin, Germany. By integrating global wastewater datasets and applying a novel computational approach to accurately identify viral strains within sewage RNA-sequencing data, we demonstrated the emergence and global dissemination of a specific astrovirus strain. Astrovirus abundance and sequence variation mirrored temporal and spatial patterns of infection, potentially serving as footprints of specific timeframes and geographical locations. Additionally, we revealed more than 100,000 sequence contigs likely originating from novel viral species, exhibiting distinct profiles in total RNA and DNA datasets and including undescribed bunyaviruses and parvoviruses. Finally, we identified thousands of new CRISPR-associated protein sequences, including Transposase B (TnpB), a class of compact, RNA-guided DNA editing enzymes. Collectively, our findings underscore the potential of high-throughput sequencing of total nucleic acids derived from wastewater for a broad range of applications.

1. Introduction

High-throughput genomic sequencing of environmental samples can be a valuable biomonitoring tool (Cordier et al., 2021). Wastewater has been used for decades to monitor the spread of human pathogens (Kilaru et al., 2023), and was repeatedly shown to be a useful source to detect circulating known as well as novel viruses (Adriaenssens et al., 2018; Bibby and Peccia, 2013; Cantalupo et al., 2011; Guajardo-Leiva et al., 2020; Martinez-Puchol et al., 2021; Perez-Cataluna et al., 2021; Rothman et al., 2021; Tisza et al., 2023; Xagoraraki and O'Brien, 2020). Among the many detected pathogens were astroviruses, which are known for inducing gastroenteritis in humans (Boujon et al., 2017) and have been shown to be present in wastewater samples in significant diversity (Tao et al., 2022; Yang et al., 2021). Also, enteroviruses, particularly polioviruses, are studied in sewage since decades (Anis et al., 2013; Paul et al., 1939). In addition to polioviruses, the genus *Enterovirus* of the family *Picornaviridae* includes rhinoviruses, coxsackieviruses and echoviruses, which can cause a wide range of symptoms, including respiratory illness, meningitis, rash ("hand, foot, and mouth disease"), or paralysis (Harvala et al., 2018). These viruses spread via respiratory droplet, fomites, and fecal-oral transmission.

During the SARS-CoV-2 pandemic, wastewater-based monitoring of

* Corresponding author at: Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany.
E-mail address: markus.landthaler@mdc-berlin.de (M. Landthaler).

https://doi.org/10.1016/j.envint.2024.108875
Received 28 March 2024; Received in revised form 3 July 2024; Accepted 3 July 2024
Available online 8 July 2024

respiratory viruses sparked considerable interest (Toribio-Avedillo et al., 2023). Particularly, targeted sequencing of viral RNA extracted from wastewater attracted considerable interest for near-real time tracking of the evolutionary dynamics of this virus (Diamond et al., 2022; Yousif et al., 2023). In concert, the emergence of ultra-deep sequencing in combination with bioinformatics has enabled the extensive identification of previously unknown viruses, both from public databases and seawater samples, thus expanding considerably our knowledge of viral diversity (Edgar et al., 2022; Gregory et al., 2019; Martinez-Hernandez and Fornas, 2022). Such approaches are currently also used to detect novel genes coding for enzymatic activity from high-throughput sequencing data (Altae-Tran et al., 2023; Paoli et al., 2022; Xiang et al., 2023).

Several aspects are investigated in depth in these studies. This includes enrichment of specific viruses with clinical importance from wastewater RNA (Martinez-Puchol et al., 2021; Rothman et al., 2021; Tisza et al., 2023). Analysis of this enriched data showed specific patterns across time points and neighboring cities for the occurrence of individual viruses (Rothman et al., 2021; Tisza et al., 2023). Novel virus discovery so far was focused mostly on RNA viruses, with the largest yield coming from database and seawater sources (Edgar et al., 2022; Gregory et al., 2019). Similarly, discovery of for example novel DNA endonucleases is currently done from public sequence databases (Altae-Tran et al., 2023; Xiang et al., 2023). What is however lacking, is a comprehensive analysis of the potential of unbiased high-throughput sequencing of both total, i.e. non-enriched, RNA and DNA from wastewater. We therefore aimed to investigate the possibilities of this data type for human pathogen tracking in long time series and comparisons across the entire globe, as well as for the detection of novel viruses and enzymes with biotechnological capacity.

In this study, we describe a longitudinal deep-sequencing effort of nucleic acids extracted from Berlin wastewater, covering a time period of 17 months with 116 RNA sequencings in total, and for the last 3 months 24 paired DNA sequencings. Total RNA revealed unbiased temporal dynamics of subspecies and strains from several virus families, including astroviruses, enteroviruses, noroviruses and adenoviruses. In case of astroviruses, we were able to track the occurrence of single point mutations over the sampling time period. By adding published total wastewater RNA sequencing data from Nagpur/India from 2021 (Stockdale et al., 2023), California/USA from 2020/2021 (Rothman et al., 2021; Rothman et al., 2023), and a worldwide collection from 2016 (Nieuwenhuijse et al., 2020) to our analysis, we show variant and point mutation dynamics over the years and across continents. Furthermore, we significantly extended the knowledge about viral diversity by identifying more than 100 thousand sequence contigs belonging to previously unknown RNA and DNA viruses. Lastly, we identified thousands of new Cas-related protein coding sequences, and describe examples of Transposase B (TnpB), a type of small RNA-guided DNA endonucleases (Karvelis et al., 2021) in detail. In summary, our study shows that total nucleic acid sequencing of wastewater samples provides an extraordinary rich and informative source to monitor known and novel microbes and their variants as well as identify previously unknown protein-coding sequences potentially useful for a range of applications.

2. Results

2.1. RNA and DNA sequencing of wastewater samples

In this study, we aimed for a deep and longitudinal metagenomic profiling of microbes present in wastewater, with a specific focus on viruses. To this end, we collected raw influx samples between March 2021 and July 2022 from a wastewater treatment plant in Berlin/Germany. RNA was extracted over the entire time-course, DNA only from samples from May to July 2022. 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 pepper mild mottle virus (PMMV), tomato brown rugose fruit virus (ToBRFV), SARS-CoV-2, pan-human astrovirus, and qPCR on pan-human adenovirus is shown in Supplementary Table S1. In total, we generated 116 sequencing libraries from 85 RNA samples, and 24 sequencing libraries from 18 paired DNA samples (Supplementary Fig. S1A, Supplementary Table S1 sheets “RNA” and “DNA”).

2.2. Viruses show specific abundance patterns in total wastewater RNA sequencing

For an initial assessment of the diversity of the detected biological entities in the sample, we applied the metagenomics pipeline kaiju to the total RNA sequencing data (Menzel et al., 2016). Bacteria were the most abundant kingdom (Supplementary Fig. S1B), as in previous wastewater metagenomics studies (Numberger et al., 2022; Rothman et al., 2020). Initially, kaiju annotated reads to 67,322 taxonomies. Low count filtering on all samples preserved 16,082 taxonomies. A principal component analysis (Supplementary Fig. S1C) revealed an additional set of 11,984 taxonomies that were only detected in 1–3 samples. These highly variable taxonomies had particularly high read counts in the outlier samples (Supplementary Fig. S1D).

The analysis of the 16,082 taxonomies showed genera from all three super kingdoms and viruses, with some however clearly dominating the sequencing dataset (Supplementary Fig. S1E-G). The virus families *Virgaviridae*, *Siphoviridae*, *Astroviridae*, *Myoviridae*, *Dicistroviridae*, *Podoviridae*, *Microviridae*, and *Picornaviridae* (Fig. S1F) were most abundant as recently defined in the “baseline for global urban virome surveillance in sewage” (Nieuwenhuijse et al., 2020),

In order to verify the sampling time course, we investigated several viruses with known or expected seasonality in the metagenomics data. First, we looked on two human pathogens with relatively abundant RNA levels in wastewater, namely astroviruses and enteroviruses (Fig. 1A, top). As expected, astroviruses peaked during northern hemisphere winter (Boujon et al., 2017; Glass et al., 1996), and enteroviruses in summer (Keeren et al., 2021). Second, we quantified viruses affecting seasonal food (asparagus in spring, watermelons in summer, grapevines in fall) and found them in the expected time frames (Fig. 1A, middle). And thirdly, Daeseongdong virus 2, which infects the common house mosquito (*Culex pipiens*) present in Berlin only during the Northern hemisphere summer, exhibited the expected trend, with peaks during the warmer season (Fig. 1A, bottom).

In summary, the majority of sequenced reads were attributed to a limited set of predominantly bacterial genera, mirroring findings from prior studies on wastewater metagenomics. While the majority of species displayed no discernible temporal trends, certain viruses exhibited distinct and anticipated seasonal patterns.

2.3. Global and temporal distribution of astrovirus strains

Astroviruses were identified as some of the most abundant human pathogens found in our samples. We therefore aimed to explore the spatial and temporal dynamics of astrovirus strains observable in high-throughput sequencing data from total wastewater RNA, by jointly analyzing our data from Berlin/Germany alongside three previously published datasets. From these, we included only samples with sufficient astrovirus reads, which resulted in about 40 samples each from three wastewater treatment plants in the cities of Los Angeles and San Diego in California/USA collected between June 2020 and May 2021 (Rothman et al., 2021; Rothman et al., 2023), three to eight samples each from five sampling locations in the city of Nagpur/India collected in February/March 2021 (Stockdale et al., 2023), and 51 samples from a worldwide sewage “baseline” sample collection from January-March 2016 (Nieuwenhuijse et al., 2020).

Our computational pipeline, tailored for this analysis, seeks to identify viral genomic sequences from the NCBI database, which show

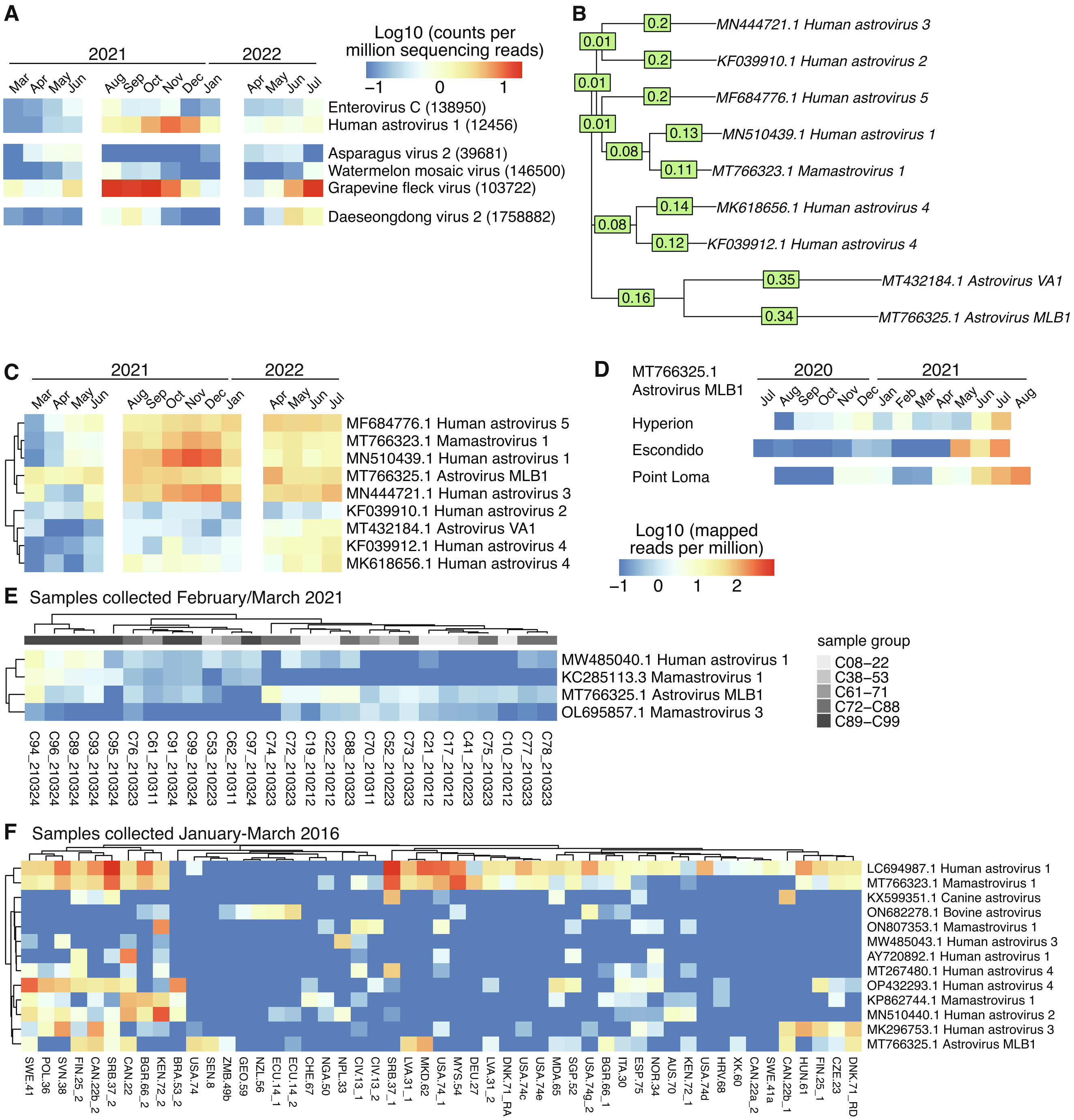

Fig. 1. Spatial and temporal distribution of astrovirus strains. A, Counts from the kaiju metagenomics pipeline aggregated per month for selected human viruses, three viruses affecting seasonal food (asparagus, watermelon, grapevine) and the mosquito virus Daeseongdong virus 2. Reads are shown as log10 transformed counts per million sequencing reads aggregated per month as indicated. Shown are names, and taxonomy IDs in brackets. B, phylogenetic tree of astroviruses shown in C, based on genomic nucleotide sequences. C, upper part, heatmap depicting abundances of the indicated astrovirus strains in the Berlin samples over time. Reads are shown as log10 transformed mapped per million aggregated per month as indicated. The color scale is the same for C-F. D, as for C but for three waste water treatment plants in California/USA collected July 2020 to August 2021 and the only detected astrovirus. E, as for C but for sewage samples from Nagpur/India collected in February/March 2021. Samples in the same group originate from locations close to each other. F, as for C but from various locations worldwide as indicated in the bottom, collected January-March 2016.

the closest match to the actual genomes in the analyzed samples. Thus, the presence of a particular accession number representing a database sequence does not signify presence of that specific viral genome sequence in the data. Rather, it indicates that this sequence is the most probable closest match to the one found in the sample.

Over the entire time course of the Berlin samples, the pipeline identified nine dominant astrovirus strains (Fig. 1B). Around 90 percent of the sequenced RNA fragments, which likely originate from the genus *Mamastrovirus,* mapped to these 9 genomes, indicating that these 9 indeed represent the most abundant circulating viruses (Supplementary Fig. S1H).

Next, we visualized the abundances of these astroviruses over time using normalized read counts aggregated per month in the Berlin sequencing data (Fig. 1C, see Supplementary Fig. S1I for individual sample data). This illustrates distinct peaks of various strains occurring at specific periods during the investigated timeframe. Astrovirus MLB1 was present throughout the analyzed time period, and was the most abundant one during spring 2021, when non-pharmaceutical interventions due to the SARS-CoV-2 pandemic where still in place. Subsequently, human astrovirus 1 peaked around November/December 2021, whereas astroviruses VA1 and 4 emerged only towards the end of the investigation, starting in April 2022. The different astrovirus strains therefore follow a distinct temporal pattern consistent with observed epidemic infection waves (Wang et al., 2023).

The identical computational approach was then independently applied to the remaining three datasets (Fig. 1D-F). Data was aggregated per month for the California samples (Fig. 1D, see Supplementary Fig. S1J for individual samples). Interestingly, the exact same accession number (MT766325) from astrovirus MLB1, as observed in the Berlin samples, emerged as the only identifiable member of the *Mamastrovirus* genus. This occured despite other MLB1 sequences in the database sharing up to 99.5 % nucleic acid identity (Supplementary Fig. S1K). Similarly, astrovirus MLB1/MT766325 exhibited a high prevalence in sewage samples collected during February/March 2021 in India (Fig. 1E). This observation suggests the potential global circulation of the same virus strain, or one with a common origin, between 2020 and 2022.

The dataset from Nagpur/India contains 140 sequencing samples, grouped by geographical location. In contrast to the sample collections from Berlin (Fig. 1C) and California (Fig. 1D), which span 17 and 12 months, respectively, the Nagpur samples and also the following “baseline” collection were collected within a few weeks. We therefore rather looked for geographical instead of temporal associations in the data. Instead of displaying the samples ordered by time, as in Fig. 1C and 1D, we applied unsupervised clustering of the samples (i.e. columns of the heatmap) based on the abundance of the four detected astrovirus strains (Fig. 1E). This clustering indeed to a certain extent recapitulated geographical sampling groups. For example, the combination of all three except mamastrovirus 3 was found in sample group C89-C99, whereas C08-C22 contained mamastrovirus 3, but not KC285113.3/mamastrovirus 1. This indicates that distributions of astrovirus strains in sewage sequencing data can be associated to specific locations.

The global “baseline” sewage sample collection from 2016 revealed a diverse range of astrovirus strains (Fig. 1F). Astrovirus MLB1 MT766325 reappeared, but not as predominant as in 2020/2021. Conversely, the most prevalent sequence in this dataset (LC694987.1/human astrovirus 1) was absent in the more recent datasets, with the closest matches showing only 90.5 % (India, MW485040.1) and 94.2 % (Berlin, MT766323.1) identity.

In summary, these data demonstrate how comprehensive sequencing of total wastewater RNA enables precise and unbiased tracking of variant dynamics over time, particularly for prevalent microbes like human astroviruses. Furthermore, the composition of astrovirus strains in sewage samples can offer distinctive and precise “spatiotemporal barcoding information”.

2.4. *Profiling of individual nucleotide changes within viral genomes in sewage samples*

For astroviruses, sequence coverage in sewage samples was often high enough to reliably detect differences in individual nucleotides. We therefore asked whether we could track not only different strains from total wastewater RNA across time and geographical location, as described above, but also individual nucleotide changes. To this end, we merged, based on time and location, samples with sufficient astrovirus MLB1 MT766325 coverage from the Berlin, India and California datasets. This allowed us to determine the percentage of mutations at 20 positions relative to the reference genome (Fig. 2A). Through unsupervised clustering, we observed distinct mutations specific to locations (e.g. position 5018 only in Berlin, 1079 only in California, 1007 in both California and India etc.) and time (e.g. 3829 only in 2021, 3742 only in 2022).

Furthermore, we investigated the temporal evolution of individual nucleotide changes. Using the most prevalent strain (astrovirus 1, MN510439.1) and total RNA sequencing data from Berlin, we analyzed mutation patterns aggregated monthly. Variability was detected at approximately 100 positions (Supplementary Fig. S2A). Notably, certain mutations appeared or disappeared during specific time periods, such as

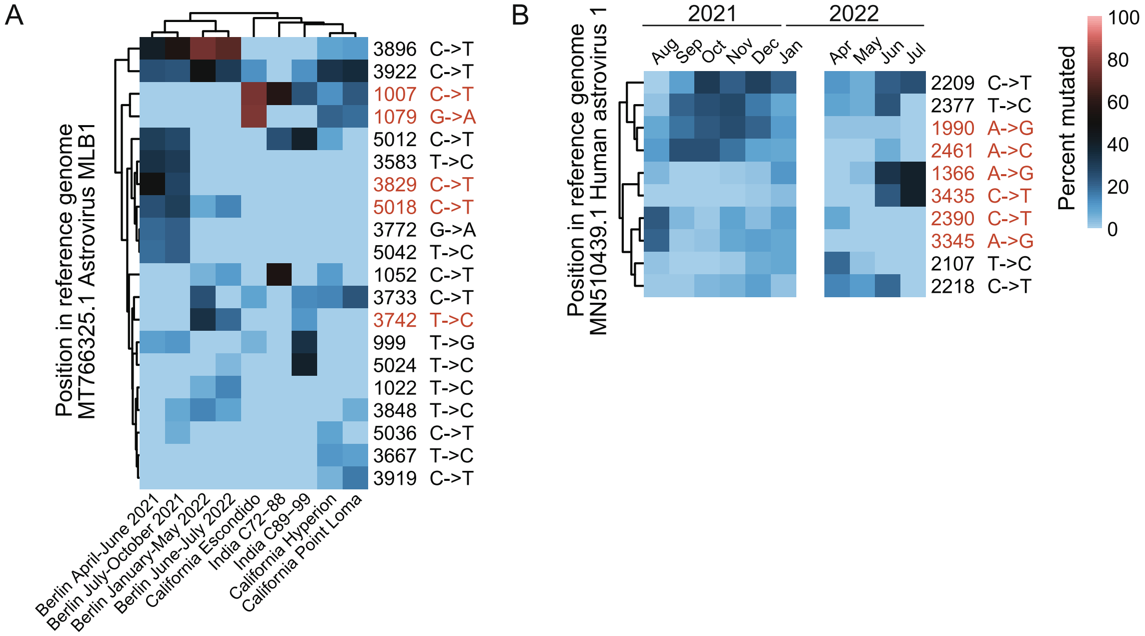

Fig. 2. Profiling of individual nucleotide changes in sewage samples. A, sewage sequencing data was merged by location (India, California) or time period (Berlin) as indicated, and the percent mutation per position indicated on the right was calculated in respect to the reference sequence (MT766325.1 Astrovirus MLB1). The type of mutation is indicated next to the position. The color scale is indicated on the right of the figure. B, as in A, but only samples from Berlin aggregated per month and on MN510439.1 human astrovirus 1. Positions marked in red are referred in the text.

in fall 2021 or June/July 2022 (Fig. 2B), indicative of epidemic wave dynamics.

Our findings underscore the ability to analyze microbial genomic data at various levels, from strains to single nucleotides, depending on abundance and sequencing depth. The clustering of mutation patterns by time and location supports the robustness of the data.

2.5. Temporal dynamics of enteroviruses, noroviruses, and adenoviruses

Next, we pondered whether our analytical framework for detecting virus strains in total RNA sequencing data could extend to virus families with lower RNA abundance in wastewater. We investigated three additional viral genera impacting humans, starting with the notably diverse enteroviruses. We noted the anticipated seasonal pattern (Keeren et al., 2021), with highest signal occurring in northern hemisphere summer months (Fig. 3A illustrates aggregated monthly data, Supplementary Fig. S3A displays data from individual samples). Similarly to the observation with astroviruses, signal levels were overall lower during spring of 2021, coinciding with the continued implementation of significant non-pharmaceutical interventions due to the SARS-CoV-2 pandemic. Nonetheless, relative abundances exhibited distinct patterns, with serotype CVA22 detected throughout the sampling period, whereas CVB5, CVA16, CVA9, and CVA10 were predominantly observed in 2022. Norovirus data demonstrated the anticipated predominance of GII over GI subtypes (Fig. 3B for aggregated monthly data, Supplementary Fig. S3B for individual sample data). Notably, the most prevalent genotypes detected from wastewater corresponded to those identified in clinical samples within the same time period. The most abundant one in wastewater, GII.P16-GII.4, accounted for 37.5 % in clinical samples. GII.P12-GII.3 was second in both waster and clinical cases (12.9 %), followed by GI.P11-GI.6 with 6 % (Jacobsen et al., 2024). The analysis of adenoviruses revealed types 41, 40, and 31 (Fig. 3C for aggregated monthly data, Supplementary Fig. S3C for individual sample data). Again, these were also the most prevalent ones in clinical samples during this time period in Germany (Albert Heim, personal communication). This comprehensive assessment underscores the utility of total wastewater RNA sequencing for monitoring the temporal dynamics of viruses with compared to astroviruses lesser abundant genetic information in wastewater, and how well wastewater data aligns with clinical testing results.

2.6. Enrichment of virus sequences from wastewater RNA

To assess the feasibility of targeted analysis and validate our findings, we employed three sequence enrichment strategies followed by sequencing. Initially, we applied the QIAseq xHYB adventitious agent panel, capable of enriching 132 different pathogenic human viruses. Supplementary Fig. S3D presents the normalized read counts for a subset aggregated monthly, illustrating the detection of respiratory viruses that are typically low in wastewater before enrichment, such as RSV and influenza, or the common cold coronaviruses NL63, 229E, HKU1, and OC43 (Supplementary Fig. S3E). Leveraging data from the German Clinical Virology Network (Adams, 2022), we compared these findings with clinical diagnostics, highlighting percentage test positivity for various respiratory and gastrointestinal viruses (Supplementary Fig. S3E). For example, simultaneous peaks in wastewater and individual testing data were observed for HKU1, NL63, Rotavirus A and RSV (Supplementary Fig. S3F). Importantly, sequencing data following xHyb enrichment revealed a lower abundance of Norovirus GII compared to GI, contrary to total RNA-seq data. This observation underscores the importance of careful enrichment method selection and interpretation, especially for viruses with variable genomes.

The xHYB system enriches double-stranded cDNA post-fragmentation in the sequencing library preparation process. In contrast, ElementZero magIC beads enriches RNA, which was examined for SARS-CoV-2 and Tomato brown rugose fruit virus (ToBRFV) by RT-qPCR and high-throughput sequencing. Notably, PMMV RNA was substantially depleted, while the target viruses were enriched as anticipated (Supplementary Fig. S3G). Intriguingly, the SARS-CoV-2 sequence coverage was restricted to hybridization probe regions, unlike ToBRFV, where coverage spanned the genomic RNA (Supplementary Fig. S3H). This discrepancy may stem from fragmented SARS-CoV-2 RNA in wastewater, limiting signal intensity outside probe regions, while intact ToBRFV RNA facilitated a broader recovery.

Additionally, we assessed the VirBaits 2.0 panel alongside xHYB and magIC systems, featuring oligonucleotide baits for various epizootic and

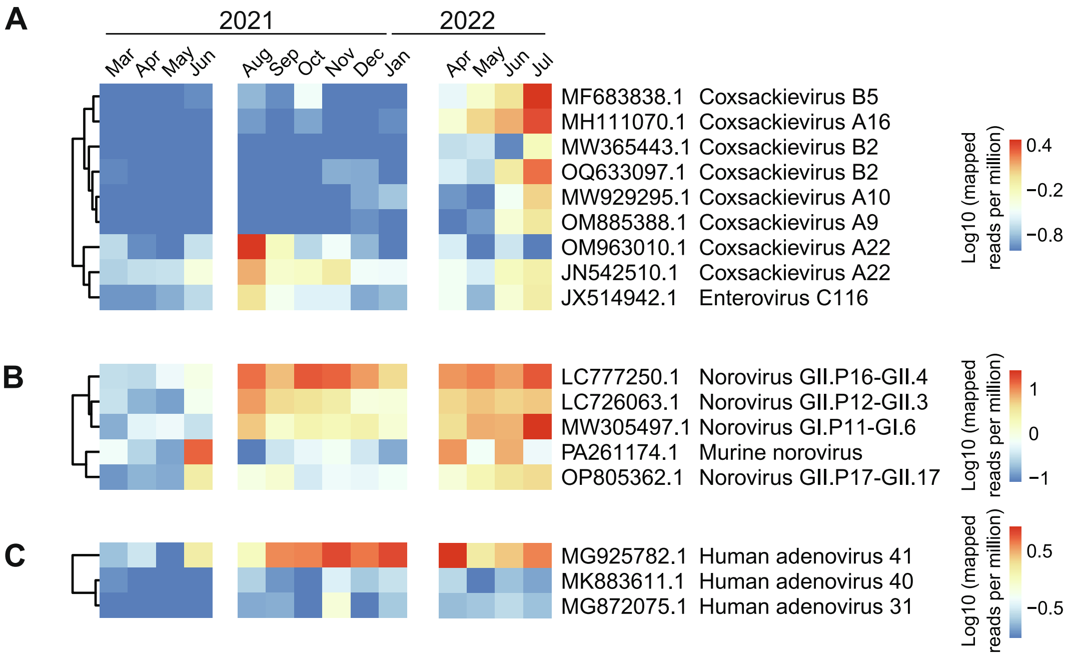

Fig. 3. Temporal species/strain dynamics of enteroviruses, noroviruses and adenoviruses. Heatmap depicting the temporal dynamics of the abundances of serotypes/genotypes of enteroviruses (A), noroviruses (B) and adenoviruses (C) found in the Berlin wastewater samples. Reads are shown as log10 transformed mapped per million, aggregated per month as indicated.

zoonotic viruses. (Wylezich et al., 2021). While the enrichment was robust, this panel proved unsuitable for wastewater samples due to pathogen absence (see Supplementary Note for detailed methods and results).

Overall, these enrichment experiments corroborated our RNA time course data, revealing seasonal trends in respiratory viruses with minimal wastewater RNA levels undetectable by total RNA sequencing. Furthermore, sequencing-based detection of panel-enriched nucleic acids enables simultaneous detection of numerous viruses. However, as demonstrated with noroviruses, targeted methods may introduce biases compared to total nucleic acid sequencing.

2.7. Detection of novel viruses from total wastewater RNA sequencing data

Considering the broad array of known viruses detected by total wastewater RNA sequencing, we set out to search for novel viruses. Employing a Hidden Markov Model (pHMM)-based approach, we conducted a sequence identity search against predicted peptide sequences encoded by ORFs spanning at least 300 nucleotides. This effort led to the identification of 417,972 sequence contigs, assembled using SPAdes (Prjibelski et al., 2020), with significant sequence similarity to a comprehensive set of known DNA and RNA virus sequences, thus confirming their viral origin (Supplementary Table S3, see Methods and data availability statement for details). These viruses span across 49 known and five unclassified virus orders. Notably, the identified viral sequences, quantified in terms of number of non-redundant contigs (Fig. 4A, E) and abundance (Fig. 4B, F), were predominantly members of the orders *Caudovirales* (DNA viruses), *Norzivirales* and *Levivirales* (RNA viruses), primarily infecting microbes and likely representing bacteriophages. With only 49,329 (11.8 %) exceeding 1000 nt (Fig. 4C, G), the considerable sequence fragmentation warrants caution regarding the actual virus count, as a single virus may contribute multiple sequence fragments. For classification of contigs into sub-operational taxonomic units representing virus species (sOTUs), we applied a previously proposed 90 % pairwise amino acid sequence identity threshold (Edgar et al., 2022). The majority of 277,092 of the viral contigs (66.3 %) displayed sequence identity below 90 % compared to known viruses (Fig. 4D, H), suggesting their membership in novel sOTUs potentially originating from hitherto unknown viruses.

To provide a more comprehensive analysis, we selected 107 non-redundant sequences that were classified as *Bunyavirales* or its sister order *Articulavirales*. These sequences exhibited diverse lineages within Bunyavirales, as depicted in the L protein-based phylogeny (Fig. 4I), suggesting the existence of novel virus subfamilies. Additionally, some sequences clustered with the family *Leishbuviridae,* known to infect Leishmania relatives, further highlighting the novelty of the discovered viruses. (Grybchuk et al., 2018) (Fig. 4I). Most of the bunyavirus-like contigs were identified via sequence homology to L proteins of known bunya- and articulaviruses while others matched nucleocapsid proteins or glycoproteins of certain bunyavirus families. To corroborate the authenticity of these finding, we identified 16 bunyavirus-like sequences which we discovered in a large screen of public transcriptome projects from the Sequence Read Archive (SRA) and that shared 80 % or more protein sequence identity to the viruses identified in wastewater.

We re-discovered some similar viral sequences from the wastewater data in the SRA analysis, which provides independent support for the authenticity of the described novel viruses. Finally, an amino acid sequence alignment of motifs A, B and C of the RNA-dependent RNA polymerase showed conserved regions among the viral subfamilies (Fig. 4J, Supplementary Fig. S4A).

Evaluating the identified virus sequences for potential novel taxa above the species level, we analyzed the most abundant RNA virus phylum of *Lenarviricota* (Fig. 4E, F). By joining *Lenarviricota* sequences from three recent large-scale virus discovery studies (Edgar et al., 2022; Neri et al., 2022; Zayed et al., 2022) and the wastewater contigs, we clustered about 90,000 *Lenarviricota* RdRp sequences at 50 % amino acid sequence identity (Supplementary Fig. S4B). We observed a few very small clusters consisting entirely of novel sequences identified from wastewater, while the majority of wastewater sequences grouped, at the chosen sequence identity threshold, with sequences reported before. This underscores the significant diversity of both known and novel virus genomes in wastewater samples, contributing to a better understanding of viral phylogenies and filling existing gaps.

2.8. Detection of novel viruses from total wastewater DNA sequencing data

After observing a considerable number of novel viruses in the total RNA sequencing results, we wondered whether this could also be the case for data from total wastewater DNA. Consequently, we sequenced total wastewater DNA obtained from 24 samples collected between May to July 2022. The metagenomics analysis revealed a considerable diversity across all kingdoms (Supplementary Fig. S5A, Supplementary Table S4). Notably, there was a marked difference in virus detection compared to the RNA data, as only DNA viruses, predominantly phages, were identifiable. Building upon the identification of the adenoviruses in the RNA data (Fig. 3C), we aimed to quantify them in the DNA dataset. Indeed, we detected the same adenovirus types with the same relative abundances (Fig. 5A, see Supplementary Fig. S5B for individual samples). Overall, levels were constant across the sampling time course, in agreement with the qPCR data (Supplementary Table S1, sheet “DNA”). Additionally, and in contrast to the RNA findings, we also identified adeno-associated virus 2 in the DNA dataset (Fig. 5A, see Supplementary Fig. S5B for individual samples).

When applying the same viral detection algorithm used for the RNA dataset, we uncovered a significant number of novel parvoviruses. These viruses, capable of infecting both vertebrate and invertebrate hosts, encompass human pathogens such as parvovirus B19 and human bocavirus 1 (Qiu et al., 2017), with genomes of about 4–6 kb single-stranded DNA. We generated a phylogenetic tree using 152 NCBI RefSeq NS1 sequences, as well as 326 novel parvovirus contigs from our wastewater DNA data (Fig. 5B). A small number of novel contigs clustered together with members of the *Parvovirinae* subfamily, while others were interspersed among the insect-infecting *Densovirinae*. Notably, the largest cluster (top right) however separated clearly and included a sequence originating from hepatitis patients (Xu et al., 2013). Of note however, subsequent analysis revealed that such sequences might originate from lab materials (Asplund et al., 2019; Naccache et al., 2013), suggesting challenges in tracing their origin.

Hence, we investigated the occurrence of both known and novel parvoviruses over time in the RNA and DNA obtained from Berlin, alongside other datasets reanalyzed in the sections above. Of note, the emergence of novel parvovirus sequences stemmed from samples taken within a limited time period in Berlin, resulting in a dataset biased towards specific temporal and geographic parameters. Nevertheless, we identified multiple of the newly defined parvovirus sequences not only in the Berlin RNA data (Supplementary Fig. S5C), but also in the data from California/USA (Supplementary Fig. S5D), Nagpur/India (Supplementary Fig. S5E), and the global Baseline collection (Supplementary Fig. S5F). Confirming expectations, these sequences exhibited their highest frequencies in the Berlin DNA dataset, the source of their derivation (Supplementary Fig. S5G). Among the most prevalent sequences (NODE_913, NODE_1289 and NODE_23291), parvovirus proteins showed 44–77 percent identity with previously known structural/non-structural peptides from insect-associated viruses. Similar to observations for astroviruses, there were distinct groups of sequences that clustered together in different time points and locations. Considering the circulation of these viral sequences in the environment, it is plausible that a considerable number of previously unidentified parvoviruses exist, some with global distribution and others confined regionally.

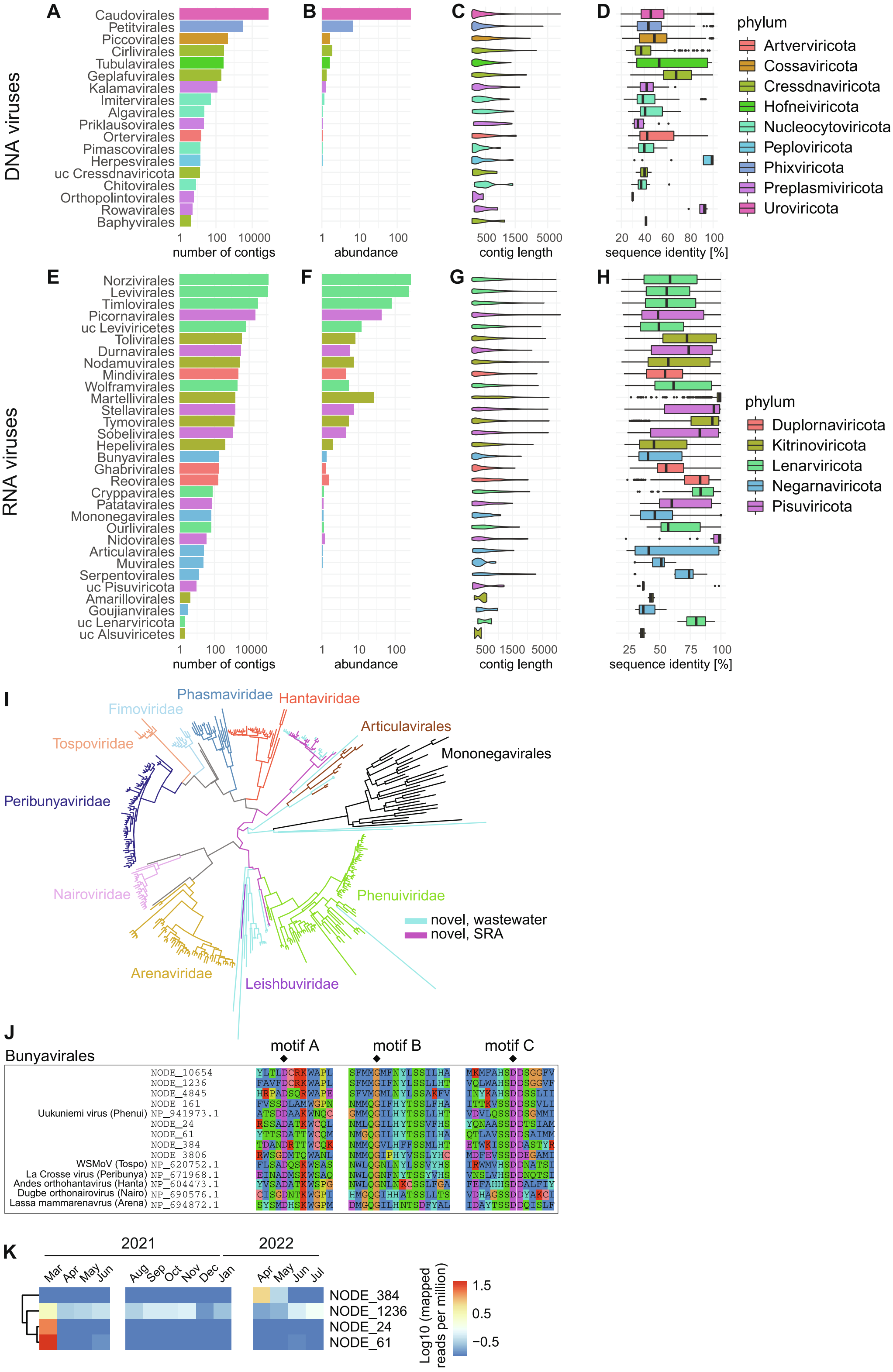

Fig. 4. Novel viral sequences discovered from wastewater. A, E, Number of non-redundant novel DNA viral (top row) and RNA viral (bottom row) contigs identified in all waste water samples together, aggregated by order and colored by phylum. "uc" = unclassified. B, F, Viral abundances. C, G, Lengths of 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. K, Heatmap depicting abundances of the indicated novel bunyaviruses in the Berlin samples over time. Reads are shown as log10 transformed mapped per million (same scale for C-F), aggregated per month as indicated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

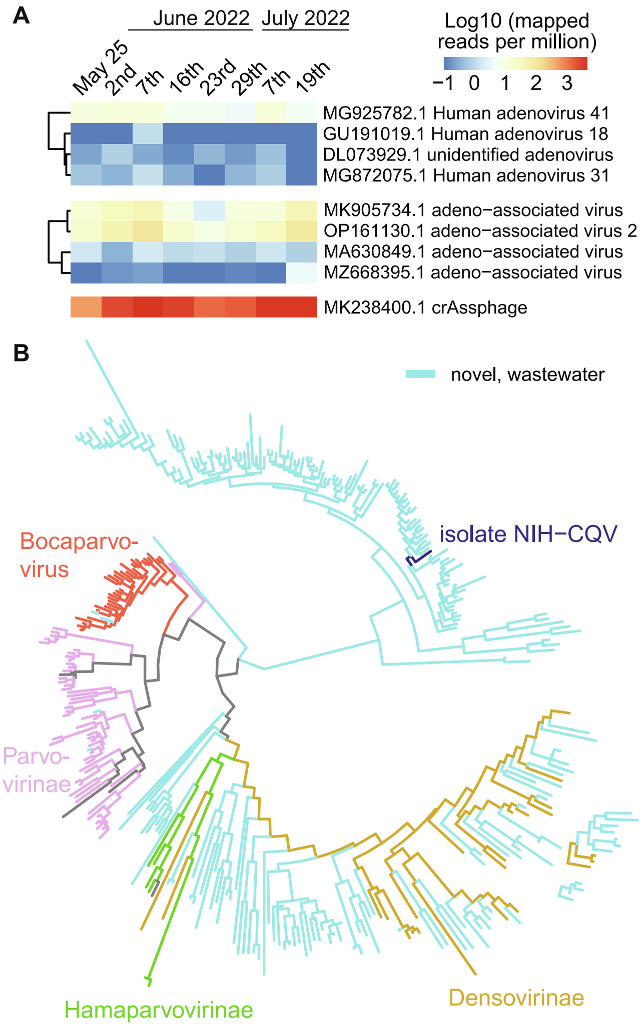

Fig. 5. Virus detection in total waste water DNA sequencing. A, abundance of human adenovirus and adeno-associated virus types, and crAssphage. Reads are shown as log10 transformed mapped per million, aggregated per day as indicated. B, NS1 protein-based Bayesian phylogeny of known and novel parvoviruses. Parvovirus subfamilies and genera are discriminated by branch color. Viruses discovered from wastewater are in cyan. The isolate from Xu et al., 2013 is labeled in dark blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.9. Total wastewater nucleic acid sequencing data contains a diverse set of Cas-related RNA-guided endonucleases

In order to explore the potential for discovering new enzymes from total wastewater RNA and DNA sequencing data, we applied the CRISPRCasTyper algorithm (Russel et al., 2020), which can identify CRISPR-associated proteins on assembled sequence contigs, as an example application. This tool identified 13,531 ORFs with <90 % identity to known proteins from several classes of CRISPR-associated proteins. High-dimensional clustering of these ORF protein sequences showed substantial diversity among the different protein groups (Fig. 6A, Supplementary Table S5).

Notably, among these were about 300 sequences of genes coding for transposase B (TnpB), which small proteins (about 400 amino acids) capable of RNA-guided DNA editing capability (Karvelis et al., 2021). Nearly all sequences originated from the DNA data, underscoring the value of sequencing both RNA and DNA from wastewater. We clustered these protein sequences, focusing on six major clusters (Supplementary Fig. S6A). For each novel TnpB sequence, we determined its identity to the closest relative in the NCBI protein database, with average identities per cluster ranging between 42 % and 75 % (Supplementary Fig. S6A). In addition, we also performed DNA sequence clustering with contigs in the six clusters, which extended more than 200 bp both upstream and downstream of the TnpB gene, (Supplementary Fig. S6B) resulting in the same clusters, highlighting the relatedness of TnpB genes on DNA and amino acid sequence level.

Next, we examined the presence of RuvC nuclease domain and TnpB amino acids motifs (Bao and Jurka 2013; Jiang et al., 2023). While they exihibited high conservation, significant sequence diversity was observed overall (Fig. 6B). Among the six selected clusters, about 80 percent of sequence contigs containing TnpB genes provided sufficient sequence information to assess the presence of a complete transposon (i. e. including an upstream TnpA gene), a configuration observed for 42 of the remaining 123 cases (Fig. 6C). Conversely, other sequence contigs exhibited surrounding ORFs with significant BLAST hits against the NCBI protein database. For instance, we identified a sequence contig with ORFs next to the TnpB gene sharing about 50 % similarity to phage-encoded nucleotide-modifying enzymes (Fig. 6D).

Taken together, sequencing of total nucleic acids from wastewater, particularly DNA, revealed a significant number of TnpB sequences with intact enzymatic amino acid motifs. For some of them, identities to known TnpBs were below 50 %, underscoring their evolutionary distinctiveness from all previously identified proteins.

3. Discussion

This study presents an analysis of total nucleic acid sequencing of wastewater samples from a treatment plant in Berlin, Germany, and explores the potential of a detailed analysis of such data. Overall, we analyzed 116 RNA sequencing libraries over a 17-month period, from March 2021 to July 2022, and 24 DNA sequencing libraries over a period of three months. Wastewater contains a very complex mixture of genetic information, with likely more than 10,000 different entities to be discovered in a single sample. 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 bacteria as well as the high abundance of plant viruses were recapitulated in our data. Given that genetic material from bacteria was most abundant regardless of the 0.2 μm filtering step, their RNA/DNA might be abundantly present outside of intact cells.

Despite the complexity of the data, we attempted to extract meaningful insights, especially for virus families, for which not such a the wealth of data is available as e.g. for SARS-CoV-2 through clinical testing and genome sequencing. To do so, we resorted to investigate temporal developments over our sampling time-course, in relation to known seasonality, and comparisons across geographically diverse locations. Viruses affecting humans, seasonal food or mosquitoes showed the known/expected pattern over the sampling time (Fig. 1A). Specific substrains of astroviruses showed a peak in fall/winter as expected (Boujon et al., 2017; Glass et al., 1996) (Fig. 1C, 2B). Interestingly, despite variation in sample sources and processing methods, identical astrovirus genomes appeared as most prevalent across data from the USA, India, and Europe. In addition, point mutation patterns clustered by time and space (Fig. 2A), suggesting that astroviruses can be used as a model virus to study the distribution of human pathogens over time and continents using total wastewater RNA sequencing. The viral mutation pattern can also be seen as a “geospatiotemporal” barcode for a specific time and location, which may inform population-based analyses of global pathogen spread. We therefore propose that accumulating total nucleic acid sequencing datasets from a wide variety of locations worldwide should be a priority of future research.

Astrovirus RNA can be so prevalent in wastewater that whole genome assembly becomes possible, facilitating detection. However,

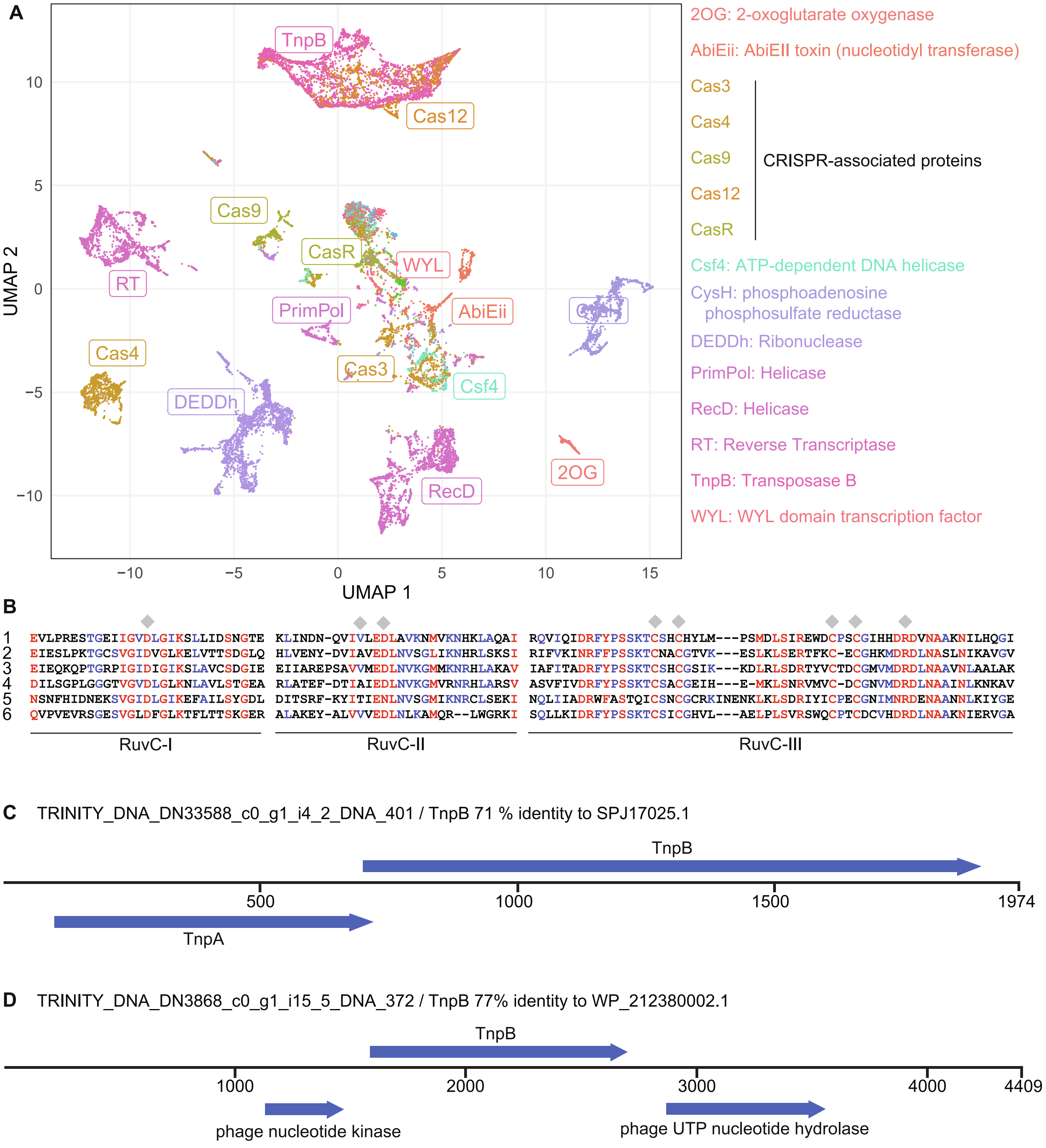

Fig. 6. Enzymes with biotechnological potential in waste water samples. A, Amino acid sequences of Cas like proteins were clustered in high-dimensional space, and projected on two dimensions using UMAP. B, Sequence alignement of novel Transposase B (TnpB) proteins, with one representative from each cluster, defined in Supplementary Fig. S6A, as indicated on the left. Conserved known amino acids of the RuvC nuclease domain and TnpB (S/T-ST-Cys zinc finger-RD) are labeled with diamonds. Highly conserved residues are colored red, medium conserved in blue. C, example of a sequence contig with a TnpA gene next to the TnpB gene. The name of contig as well as the closest relative in the NCBI protein database and the identity to this protein is indicated at the top. D, example of a contig without TnpB gene. Adjacent ORFs with a BLAST hit in the NCBI protein database are labelled with the function of the closest hit. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

this was not the case for many other viruses. Our analysis pipeline, as demonstrated in Fig. 3, enabled us to detect strain detection of enteroviruses, noroviruses and adenoviruses, despite their lower abundance. Currently, epidemiological data for these clinically significant virus families is skewed, as typically only severe cases are tested. Conversely, wastewater offers a comprehensive overview of all circulating strains and variants. Leveraging this data, it may be possible to identify the more pathogenic strains, such as the enterovirus echovirus 30 (Benschop et al., 2021), by normalizing the relative abundances of virus strains in patients with the relative abundances in wastewater.

The RNA viruses under investigation exhibit significant variability in their genomes. Unlike total RNA sequencing, targeted method may possibly introduce biases. This was evident in the norovirus GI/GII ratio, which showed a reversal in the data from the enrichment panel (Supplementary Fig. S3D) compared to total RNA sequencing (Fig. 3). Enrichment approaches therefore are not only restricted to their targets, as e.g. the enteroviruses shown in Fig. 3, not present in the panel used here, are missing in Supplementary Fig. S3D. They can also distort the data depending on the sequence distance between the genomic sequences used for probe/primer design and those circulating in the population at the time of sampling. Interestingly however, RNA-level enrichment only yielded highly fragmented genome coverage from the enveloped virus SARS-CoV-2 (Supplementary Fig. S3H). This finding suggests that this RNA, and possibly RNA from enveloped viruses in general, may be present not as viral particles but bound to proteins or other substances. Additionally, observations such as the increased abundance of mpox and influenza virus DNA/RNA in the particle fractions (Wolfe et al., 2022; Wolfe et al., 2023), underscores the need for systematic investigations into which processing methods are suitable for specific microbe detection.

Whereas wastewater sequencing studies generally investigate only RNA, we simultaneously performed DNA and RNA sequencing. Interestingly, adenovirus, but not adeno-associated virus sequences, were not only in DNA, but also RNA data. Reasons could e.g. be shedding of intact virions or RNA from infected intestinal cells in stool. Further comparisons between total RNA and DNA sequencing data from environmental samples would be needed to better define which microbes can be better detected in one versus the other type of nucleic acids.

Our discovery of possibly thousands to tens of thousands of newfound viruses illustrates how wastewater sequencing can also fill major gaps in our knowledge of the global DNA and RNA virome, provide information on virus evolution as well as aiding future assessment of spillover risks and zoonotic potentials. Several studies have uncovered vast collections of hypothetical novel viromes through the analysis of existing metagenomic data (Edgar et al., 2022; Lee, 2023; Neri et al., 2022; Zayed et al., 2022). Interestingly, our study, as demonstrated with the novel bunyaviruses (Fig. 4I), substantially expanded the diversity of novel viruses beyond what can be found in existing data such as the SRA database. This might be an indication that only a tiny fraction of all viruses has been identified with such massive searches, and/or that locally restricted viromes exist, that can only be resolved with high resolution geographical sampling. The widely different range of viruses found in RNA and DNA data generally underpins the usefulness of separate processing and analysis methods.

For the last part of our investigation, we sought to assess the extent to which our data enables the exploration of potential new enzymes for biotechnological purposes (Paoli et al., 2022). Specifically, we directed our attention to a recently identified class of Cas-related proteins known as TnpB., a recently discovered novel class of RNA-guided DNA endonucleases (Karvelis et al., 2021). Although we did not validate enzymatic activity, the intact RuvC motifs (Fig. 6) and other observations suggest potential functionality of these TnpBs. Notably, some of them have less than 50 % identity to known TnpB sequences, highlighting how total wastewater DNA sequencing can significantly enlarge the search space for potentially useful enzyme sequences. Overall, our study shows the wealth of genetic information accessible through wastewater total nucleic acid sequencing and detailed bioinformatic analysis.

Still, our study is subject to several limitations. We studied only one wastewater treatment plant in one location over time. More wastewater metagenomic data will enable more informative global comparisons. Additionally, our methodology may overlook certain microbes, as evidenced by the absence of mpox virus genomes in our samples compared to other sampling sites in the same time period (Wannigama et al., 2023; Wolfe et al., 2023). Factors such as temperature changes (15 °C to 25 °C) and varying travel time from households to the wastewater plant (one to several hours) may cause variability that is difficult to measure. Our sampling mostly covers each season once and was partly done during a unique pandemic period. And finally, obtaining an accurate and comprehensive description of microbial species content from highly complex, mixed samples is challenging. Firstly, the sensitivity of our sequence-based virus discovery approach depends on how well the natural diversity of viruses is resembled by the sequences in public databases, which are arguably incomplete and biased. Secondly, virus species quantification could be distorted by recombination, segment rearrangement and gene transfer, which can be hard to resolve using short-read sequencing technologies. Future studies, with wider geographical and temporal ranges and improved experimental and sequencing methods such as long-read sequencing, will allow to investigate microbial communities and the information they provide at an unprecedented level of detail and scope.

4. Methods

4.1. Sample collection

The sample collection was done as described previously (Schumann et al., 2022), or briefly 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 h 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.

4.2. Sample processing and RNA isolation

The sample processing was essentially done as described previously (Schumann et al., 2022), with modifications for some samples. After collection, the specimens were kept at four °C, 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 10 kDa cutoff centricon units (Millipore, cat# UFC701008), which were pre-conditioned with 50 mL ultrapure water centrifuged with 3000 g for 15 min at 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.

For the procedures with 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, and 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), as detailed in Supplementary Table S1.

Notes regarding data analysis: first, for some months, none or only one or two samples were successfully processed, therefore these months are omitted in some of the analysis. Second, for the analysis in Figs. 1-3, only the samples with standard procedure (centricon concentration) were used.

4.3. DNA isolation

DNA was isolated from either concentrated wastewater or from the nanotrap beads using the ZymoBIOMICS RNA/DNA combination extraction or the ZymoBIOMICS DNA Miniprep kit, as detailed in Supplementary Table S1, sheet "DNA".

4.4. 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 that a 20 min incubation was performed at 55 °C instead of 10 min. Afterwards, the cDNA was diluted 1:10 with ultrapure water, and of this 3.75 µL diluted cDNA used per 15 µl 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 (Corman et al., 2020; Henke-Gendo et al., 2012; Kitajima et al., 2018; Lu et al., 2020; Untergasser et al., 2012; Willeit et al., 2021). For Adenovirus Taqman qPCRs on DNA samples, 5 µl of the purified DNA was used in a 20 µl reaction with the Luna Universal Probe qPCR Master Mix (NEB cat# M3004) according to the manufacturer's protocol.

4.5. Total RNA sequencing

RNA sequencing libraries 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 protocol. Libraries were then pooled and sequenced on a Novaseq 6000 device with 2x109 bp paired-end sequencing. We generated 116 sequencing libraries, which yielded in total 1.45 billion 2x109 bp read pairs (12 million on average per sample), with about 10–20 % of reads duplicated (Supplementary Fig. S1A).

4.6. Total DNA sequencing

For preparation of sequencing libraries, 3 µl DNA was first amplified using the ResolveDNA Whole Genome Amplification kit (BioSkryb) according to the manufacturer's protocol. Subsequently, sequencing libraries were generated by doing two Nextera reactions with 1 ng pre-amplified DNA each per sample using the Nextera XT DNA Library Preparation Kit (Illumina cat# FC-131–1096) according to the manufacturer's protocol. Libraries were then pooled and sequenced on a Novaseq 6000 device with 2x109 bp paired-end sequencing.

4.7. Metagenomic analysis using kaiju

Total RNA-seq data was analyzed by the kaiju program (Menzel et al., 2016) using sequencing reads (fastq files). For both the application of kaiju pipeline as well as the subspecies analysis, the code is available 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.

4.8. Analysis of viral subspecies abundances and variants

The workflow to determine the accession numbers of the viral sequences that are closest to the viral genomes works as following. Step 1, recover all taxonomy IDs from the kaiju output, belonging to one of the genera *Mamastrovirus* (Fig. 1), *Enterovirus* (Fig. 3A), *Norovirus* (Fig. 3B), *Mastadenovirus* (Fig. 3C), and collect the genomic sequences belonging to these IDs from the NCBI nucleotide database. Step 2, map sequencing reads (fastq files) to all the downloaded sequences which are longer than 5000 nt. If no sequence is longer than 5000 nt, then the 3 longest ones are used. The mapping algorithm (hisat2 with standard parameters) allowed up to about 7 mismatches per 100 nucleotides. In the figures, the sequences are referenced with their NCBI accession IDs. Step 3, for every sample, reads mapping to several accessions are counted for the most abundant one. Step 4, filtering out low abundant accessions (below average or no peaks). Step 5, filtering out accession with high similarity to more abundant accessions. Step 6, map again to the selected set of sequences, and make readcount tables/heatmaps, aggregate by month where applicable. The scripts (bash, R) are available through GitHub (https://github.com/LandthalerLab/wastewater_virome).

Alignments of sequencing reads to viral sequences from the NCBI database were done using hisat2 (Kim et al., 2019), and read counting performed using samtools (Danecek et al., 2021) in bash. Further data processing was done in R.

4.9. Wastewater sequencing data from published datasets

Three published datasets were included in the analysis, from wastewater treatment plants in the cities of Los Angeles and San Diego in California/USA collected between June 2020 and May 2021(Rothman et al., 2021; Rothman et al., 2023), from the city of Nagpur/India collected in February/March 2021 (Stockdale et al., 2023), and 51 samples from a worldwide sewage "baseline" sample collection from January-March 2016 (Nieuwenhuijse et al., 2020). Raw data was obtained through the NCBI SRA/GEO databases.

4.10. Enrichment using the QIAseq xHYB adventitious agent panel

Individual total RNA 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 QIAseq xHYB Adventitious Agent Panel (Qiagen cat# 333355) according to the manufacturer's protocol, and sequenced after the final re-amplification step.

4.11. Enrichment using ElementZero beads

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

4.12. 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. For the DNA sequencing experiments, a joint assembly with all 24 samples was conducted to maximize read coverage breadth of individual viral genomes detected in multiple samples, while due to computational limitations each of the 116 RNA sequencing experiments was assembled separately.

4.13. 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.

4.14. 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 84,420 profiles from VirSorter 2 (Guo et al., 2021), 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: 590,872 viral proteins from NCBI RefSeq (O'Leary et al., 2016), 311,725 RdRp sequences from the Serratus and PalmDB projects (Babaian and Edgar, 2021; Edgar et al., 2022), 49,421 RdRp footprint sequences from the Tara Oceans Virome project (Zayed et al., 2022), 77,510 RdRp sequences from RVMT (Neri et al., 2022) and 15,081 RdRp sequences from RdRp Scan v0.90 (Charon et al., 2022). We considered a viral contig as having originated from a known virus if its pairwise amino acid sequence identity to the closest known virus was 90 % (Edgar et al., 2022) or higher in the DIAMOND blastx search; otherwise, we considered it to represent a novel virus species.

4.15. 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.

4.16. 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 rediscovered 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).

4.17. Detection of CRISPR-associated genes/TnpB sequences

We have separately assembled RNA and DNA sequences using the de novo assembly tool Trinity (Haas et al., 2013). For the DNA data, we did a joint assembly for the 24 samples. For the RNA data, the 116 samples were, by temporal order, grouped into 8 separate assemblies. We have incorporated CRISPRCasTyper (Russel et al., 2020) to search for Cas and other genes that are functionally linked to CRISPR-Cas systems across peptide sequences encoded by ORFs generated from both DNA and RNA assemblies. We have filtered out sequences that are shorter than 50 and those with E-value larger than 0.01. In order to create a subset of representative sequences, we have used cd-hit command from CD-HIT (Fu et al., 2012) to cluster ORFs and selected only the longest sequence (i.e. representative ORFs) from each cluster. BLASTP (Camacho et al., 2009) was used to compare representative ORFs to the non-redundant (NR) protein database from NCBI (Hahsler, 2019). We have filtered out BLASTP results for sequences shorter than 30 and used 0.01 for E-value cutoff, and we sorted hits for highest-lowest percent identity for each ORF We have used ProtTrans (Elnaggar et al., 2022) to find token-based embeddings of representative ORFs which is followed by dimensionality reduction with PCA (# of PCs = 30) and UMAP (McInnes, 2018). To analyze the similarity across protein sequences, ProtTrans incorporates Natural language processing (NLP) models that learn semantic relationships across amino acids in protein sequences by representing amino acids as "words" or "tokens", and sequences as "sentences". By using the vector representations of the last hidden state from the ProtTrans NLP model, we convert novel ORFs to points in 1024-dimensional embedding space, where each dimension represents a latent feature (i.e. an observable common structure) of protein sequences (Pokharel et al., 2022).

CRediT authorship contribution statement

Emanuel Wyler: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chris Lauber:** Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Artür Manukyan:** Software, Methodology, Investigation, Formal analysis, Data curation. **Aylina Deter:** Investigation. **Claudia Quedenau:** Investigation. **Luiz Gustavo Teixeira Alves:** Investigation. **Claudia Wylezich:** Investigation. **Tatiana Borodina:** Supervision. **Stefan Seitz:** Supervision, Software, Methodology. **Janine Altmüller:** Supervision. **Markus Landthaler:** Writing – original draft, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All raw sequencing data as well as processed data (kaiju output, assembled sequence contigs fasta files, xHYB count tables) are available via NCBI GEO, accession number GSE228220, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc = GSE228220

Acknowledgements

We are very grateful to Fredrik Zietzschmann, Katharina Flatau, Regina Gnirss and Uta Böckelmann from Berliner Wasserbetriebe for providing samples and continuous support. We thank Sindy Böttcher and Julian Kreibich (enteroviruses), Sandra Niendorf (noroviruses), Albert Heim (adenoviruses), as well as René Kallies, Hans-Christoph Selinka, Dylan Barbera, Anne Schlenther, Karol Rogowski, Martin Meixner, Tautvydas Karvelis and Virginijus Šikšnys for support and helpful discussions. We are particularly grateful for the clinical virology data provided by Norbert Schmeisser and Rolf Kaiser. 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).

Code availability

Code is available through GitHub at https://github.com/lauberlab/wastewater_virome_paper (novel virus detection) and https://github.com/LandthalerLab/wastewater_virome (everything else).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2024.108875.

References

Adams, A.G., B., Kaiser, R., Prifert, C., Schmeisser, N. Respiratory Virus Network. 2022.

Adriaenssens, E.M., Farkas, K., Harrison, C., Jones, D.L., Allison, H.E., McCarthy, A.J. 2018. Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. mSystems. 3.

Altae-Tran, H., Shmakov, S.A., Makarova, K.S., Wolf, Y.I., Kannan, S., Zhang, F., Koonin, E.V., 2023. Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. Proc. Natl. Acad. Sci. USA 120 e2308224120.

Anis, E., Kopel, E., Singer, S.R., Kaliner, E., Moerman, L., Moran-Gilad, J., Sofer, D., Manor, Y., Shulman, L.M., Mendelson, E., Gdalevich, M., Lev, B., Gamzu, R., Grotto, I., 2013. Insidious reintroduction of wild poliovirus into Israel, 2013. Euro. Surveill. 18.

Asplund, M., Kjartansdottir, K.R., Mollerup, S., Vinner, L., Fridholm, H., Herrera, J.A.R., Friis-Nielsen, J., Hansen, T.A., Jensen, R.H., Nielsen, I.B., Richter, S.R., Rey-Iglesia, A., Matey-Hernandez, M.L., Alquezar-Planas, D.E., Olsen, P.V.S., Sicheritz-Ponten, T., Willerslev, E., Lund, O., Brunak, S., Mourier, T., Nielsen, L.P., Izarzugaza, J.M.G., Hansen, A.J., 2019. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clin. Microbiol. Infect 25, 1277–1285.

Babaian, A., Edgar, R.C. 2021. Ribovirus classification by a polymerase barcode sequence. bioRxiv. 2021.2003.2002.433648.

Bao, W., Jurka, J., 2013. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. Mob. DNA 4, 12.

Benschop, K.S.M., Broberg, E.K., Hodcroft, E., Schmitz, D., Albert, J., Baicus, A., Bailly, J. L., Baldvinsdottir, G., Berginc, N., Blomqvist, S., Bottcher, S., Brytting, M., Bujaki, E., Cabrerizo, M., Celma, C., Cinek, O., Claas, E.C.J., Cremer, J., Dean, J., Dembinski, J. L., Demchyshyna, I., Diedrich, S., Dudman, S., Dunning, J., Dyrdak, R., Emmanouil, M., Farkas, A., De Gascun, C., Fournier, G., Georgieva, I., Gonzalez-Sanz, R., van Hooydonk-Elving, J., Jaaskelainen, A.J., Jancauskaite, R., Keeren, K., Fischer, T.K., Krokstad, S., Nikolaeva-Glomb, L., Novakova, L., Midgley, S.E., Mirand, A., Molenkamp, R., Morley, U., Mossong, J., Muralyte, S., Murk, J.L., Nguyen, T., Nordbo, S.A., Osterback, R., Pas, S., Pellegrinelli, L., Pogka, V., Prochazka, B., Rainetova, P., Van Ranst, M., Roorda, L., Schuffenecker, I., Schuurman, R., Stoyanova, A., Templeton, K., Verweij, J.J., Voulgari-Kokota, A., Vuorinen, T., Wollants, E., Wolthers, K.C., Zakikhany, K., Neher, R., Harvala, H., Simmonds, P., 2021. Molecular Epidemiology and Evolutionary Trajectory of Emerging Echovirus 30. Europe. Emerg. Infect. Dis 27, 1616–1626.

Bibby, K., Peccia, J., 2013. Identification of viral pathogen diversity in sewage sludge by metagenome analysis. Environ. Sci. Technol 47, 1945–1951.

Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C., Hochreiter, S., 2015. msa: an R package for multiple sequence alignment. Bioinformatics 31, 3997–3999.

Boujon, C.L., Koch, M.C., 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., Drost, H.G., 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat. Methods 18, 366–368.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformat. 10, 421.

Cantalupo, P.G.; Calgua, B., Zhao, G., Hundesa, A., Wier, A.D., Katz, J.P., Grabe, M., Hendrix, R.W., Girones, R., Wang, D., Pipas, J.M. 2011. Raw sewage harbors diverse viral populations. mBio;2.

Charon, J., Buchmann, J.P., Sadiq, S., Holmes, E.C. 2022. RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. Virus Evolution; 8:veac082.

Cordier, T., Alonso-Saez, L., Apotheloz-Perret-Gentil, L., Aylagas, E., Bohan, D.A., Bouchez, A., Chariton, A., Creer, S., Fruhe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., Lanzen, A., 2021. Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. Mol. Ecol 30, 2937–2958.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., Bleicker, T., Brunink, S., Schneider, J., Schmidt, M.L., Mulders, D.G., Haagmans, B.L., van der Veer, B., van den Brink, S., Wijsman, L., Goderski, G., Romette, J.L., Ellis, J., Zambon, M., Peiris, M., Goossens, H., Reusken, C., Koopmans, M.P., Drosten, C., 2020. Detection of novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro. Surveill 25.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H. 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., Rouchka, E.C., Smith, T., Stevenson, B.S., Susswein, Z., Vogel, J.R., Wolfe, M.K., Stadler, L.B., Scarpino, S.V., 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., Banfield, J.F., de la Pena, M., Korobeynikov, A., Chikhi, R., Babaian, A., 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature 602, 142–147.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., Rost, B., 2022. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE. Trans. Pattern. Anal. Mach. Intell 44, 7112–7127.

Fernandez-Cassi, X., Timoneda, N., Martinez-Puchol, S., Rusinol, M., Rodriguez-Manzano, J., Figuerola, N., Bofill-Mas, S., Abril, J.F., 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.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Glass, R.I., Noel, J., Mitchell, D., Herrmann, J.E., Blacklow, N.R., Pickering, L.K., Dennehy, P., Ruiz-Palacios, G., de Guerrero, M.L., Monroe, S.S., 1996. The changing epidemiology of astrovirus-associated gastroenteritis: a review. Arch. Virol. Suppl 12, 287–300.

Gregory, A.C., Zayed, A.A., Conceicao-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Dominguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., Poulain, J., Tremblay, J.E., Vik, D., Tara Oceans, C., Babin, M., Bowler, C., Culley, A.I., de Vargas, C., Dutilh, B.E., Iudicone, D., Karp-Boss, L., Roux, S., Sunagawa, S., Wincker, P., Sullivan, M.B., 2019. Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell 177 (1109–1123), e1114.

Grybchuk, D., Akopyants, N.S., Kostygov, A.Y., Konovalovas, A., Lye, L.F., Dobson, D.E., Zangger, H., Fasel, N., Butenko, A., Frolov, A.O., Votypka, J., d'Avila-Levy, C.M., Kulich, P., Moravcova, J., Plevka, P., Rogozin, I.B., Serva, S., Lukes, J., Beverley, S. M., Yurchenko, V., 2018. Viral discovery and diversity in trypanosomatid protozoa with a focus on relatives of the human parasite Leishmania. Proc. Natl. Acad. Sci. USA 115, E506–E515.

Gu, Z., Eils, R., Schlesner, M., 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849.

Guajardo-Leiva, S., Chnaiderman, J., Gaggero, A., 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., Roux, S., 2021. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome 9, 37.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., De, R.A., 2013. novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc 8, 1494–1512.

Hahsler, M., Nagar, A. 2019. rBLAST: R Interface for the Basic Local Alignment Search Tool.

Harvala, H., Broberg, E., Benschop, K., Berginc, N., Ladhani, S., Susi, P., Christiansen, C., McKenna, J., Allen, D., Makiello, P., McAllister, G., Carmen, M., Zakikhany, K., Dyrdak, R., Nielsen, X., Madsen, T., Paul, J., Moore, C., von Eije, K., Piralla, A., Carlier, M., Vanoverschelde, L., Poelman, R., Anton, A., Lopez-Labrador, F.X., Pellegrinelli, L., Keeren, K., Maier, M., Cassidy, H., Derdas, S., Savolainen-Kopra, C., Diedrich, S., Nordbo, S., Buesa, J., Bailly, J.L., Baldanti, F., MacAdam, A., Mirand, A., Dudman, S., Schuffenecker, I., Kadambari, S., Neyts, J., Griffiths, M.J., Richter, J., Margaretto, C., Govind, S., Morley, U., Adams, O., Krokstad, S., Dean, J., Pons-Salort, M., Prochazka, B., Cabrerizo, M., Majumdar, M., Nebbia, G., Wiewel, M., Cottrell, S., Coyle, P., Martin, J., Moore, C., Midgley, S., Horby, P., Wolthers, K., Simmonds, P., Niesters, H., Fischer, T.K., 2018. Recommendations for enterovirus diagnostics and characterisation within and beyond Europe. J. Clin. Virol 101, 11–17.

Hauser, M., Steinegger, M., Soding, J., 2016. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics 32, 1323–1330.

Henke-Gendo, C., Ganzenmueller, T., Kluba, J., Harste, G., Raggub, L., Heim, A., 2012. Improved quantitative PCR protocols for adenovirus and CMV with an internal inhibition control system and automated nucleic acid isolation. J. Med. Virol 84, 890–896.

Jacobsen, S., Faber, M., Altmann, B., Mas Marques, A., Bock, C.T., Niendorf, S., 2024. Impact of the COVID-19 pandemic on norovirus circulation in Germany. Int. J. Med. Microbiol 314, 151600.

Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., Banziger, C., Devaux, A.J., Stachler, E., Caduff, L., Cariti, F., Corzon, A.T., Fuhrmann, L., Chen, C., Jablonski, K.P., Nadeau, S., Feldkamp, M., Beisel, C., Aquino, C., Stadler, T., Ort, C., Kohn, T., Julian, T.R., Beerenwinkel, N., 2022. Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. Nat. Microbiol 7, 1151–1160.

Jiang, K., Lim, J., Sgrizzi, S., Trinh, M., Kayabolen, A., Yutin, N., Bao, W., Kato, K., Koonin, E.V., Gootenberg, J.S., Abudayyeh, O.O., 2023. Programmable RNA-guided DNA endonucleases are widespread in eukaryotes and their viruses. Sci. Adv 9, eadk0171.

Karvelis, T., Druteika, G., Bigelyte, G., Budre, K., Zedaveinyte, R., Silanskas, A., Kazlauskas, D., Venclovas, C., Siksnys, V., 2021. Transposon-associated TnpB is a programmable RNA-guided DNA endonuclease. Nature 599, 692–696.

Katoh, K., 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., Diedrich, S., 2021. Enterovirus Surveillance (EVSurv) in Germany. Microorganisms 9.

Kilaru, P., Hill, D., Anderson, K., Collins, M.B., Green, H., Kmush, B.L., Larsen, D.A., 2023. Wastewater Surveillance for Infectious Disease: A Systematic Review. Am. J. Epidemiol 192, 305–322.

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol 37, 907–915.

Kitajima, M., Sassi, H.P., Torrey, J.R., 2018. Pepper mild mottle virus as a water quality indicator. Npj. Clean. Water 1, 19.

Kolde, R. pheatmap: Pretty Heatmaps. 2019.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Lauber, C., 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., Seitz, S., 2021. Deep mining of the Sequence Read Archive reveals bipartite coronavirus genomes and inter-family Spike glycoprotein recombination. bioRxiv 2021, 2020, 2010.465146.

Lee, B.D., Neri, U., Roux, S., Wolf, Y.I., Camargo, A.P., Krupovic, M., Simmonds, P., Kyrpides, N., Gophna, U., Dolja, V.V., Koonin, E.V., 2023. Mining metatranscriptomes reveals a vast world of viroid-like circular RNAs. Cell 186 (646–661), e644.

Lu, X., Wang, L., Sakthivel, S.K., Whitaker, B., Murray, J., Kamili, S., Lynch, B., Malapati, L., Burke, S.A., Harcourt, J., Tamin, A., Thornburg, N.J., Villanueva, J.M., Lindstrom, S., 2020. US CDC Real-Time Reverse Transcription PCR Panel for Detection of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg. Infect. Dis 26, 1654–1665.

Martinez-Hernandez, F., Fornas, O., 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., Bofill-Mas, S., 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.

McInnes, L., Healy, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction edˆeds.

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

Naccache, S.N., Greninger, A.L., Lee, D., Coffey, L.L., Phan, T., Rein-Weston, A., Aronsohn, A., Hackett Jr., J., Delwart, E.L., Chiu, C.Y., 2013. The perils of pathogen discovery: origin of a novel parvovirus-like hybrid genome traced to nucleic acid extraction spin columns. J. Virol 87, 11966–11977.

Neri, U., Wolf, Y.I., Roux, S., Camargo, A.P., Lee, B., Kazlauskas, D., Chen, I.M., Ivanova, N., Allen, L.Z., Paez-Espino, D., Bryant, D.A., 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., Munk, P., Venkatakrishnan, S., Aarestrup, F.M., Cotten, M., Koopmans, M.P., 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., 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., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic. Acids. Res 44, D733–D745.

Paoli, L., Ruscheweyh, H.J., Forneris, C.C., Hubrich, F., Kautsar, S., Bhushan, A., Lotti, A., Clayssen, Q., Salazar, G., Milanese, A., Carlstrom, C.I., Papadopoulou, C., Gehrig, D., Karasikov, M., Mustafa, H., Larralde, M., Carroll, L.M., Sanchez, P., Zayed, A.A., Cronin, D.R., Acinas, S.G., Bork, P., Bowler, C., Delmont, T.O., Gasol, J. M., Gossert, A.D., Kahles, A., Sullivan, M.B., Wincker, P., Zeller, G., Robinson, S.L., Piel, J., Sunagawa, S., 2022. Biosynthetic potential of the global ocean microbiome. Nature 607, 111–118.

Paul, J.R., Trask, J.D., Culotta, C.S., 1939. Poliomyelitic Virus in Sewage. Science 90, 258–259.

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

Pokharel, S., Pratyush, P., Heinzinger, M., Newman, R.H., Kc, D.B., 2022. Improving protein succinylation sites prediction using embeddings from protein language model. Sci. Rep 12, 16933.

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

Qiu, J., Soderlund-Venermo, M., Young, N.S., 2017. Human Parvoviruses. Clin. Microbiol. Rev 30, 43–113.

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

Rice, P., Longden, I., 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., 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 III, J., Adams, E.D., Steele, J.A., Zimmer-Faust, A. G., Langlois, K., Wanless, D., Griffith, M., Mao, L., Chokry, J., 2021. RNA Viromics of Southern California Wastewater and Detection of SARS-CoV-2 Single-Nucleotide Variants. Appl. Environ. Microbiol 87, e0144821.

Rothman, J.A., Saghir, A., Chung, S.A., Boyajian, N., Dinh, T., Kim, J., Oval, J., Sharavanan, V., York, C., Zimmer-Faust, A.G., Langlois, K., Steele, J.A., Griffith, J.F., Whiteson, K.L., 2023. Longitudinal metatranscriptomic sequencing of Southern California wastewater representing 16 million people from August 2020–21 reveals widespread transcription of antibiotic resistance genes. Water. Res 229, 119421.

Russel, J., Pinilla-Redondo, R., Mayo-Munoz, D., Shah, S.A., Sorensen, S.J., 2020. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. CRISPR. J 3, 462–469.

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., Woyke, T., 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., Freimuth, J., Meixner, M., Grau, J.H., Liere, K., Hackenbeck, T., Zietzschmann, F., Gnirss, R., Bockelmann, U., Uyar, B., Franke, V., Barke, N., Altmuller, J., Rajewsky, N., Landthaler, M., Akalin, A., 2022. SARS-CoV-2 infection dynamics revealed by wastewater sequencing analysis and deconvolution. Sci. Total. Environ 853, 158931.

Stockdale, S.R., Blanchard, A.M., Nayak, A., Husain, A., Nashine, R., Dudani, H., McClure, C.P., Tarr, A.W., Nag, A., Meena, E., Sinha, V., Shrivastava, S.K., Hill, C., Singer, A.C., Gomes, R.L., Acheampong, E., Chidambaram, S.B., Bhatnagar, T., Vetrivel, U., Arora, S., Kashyap, R.S., Monaghan, T.M., 2023. RNA-Seq of untreated wastewater to assess COVID-19 and emerging and endemic viruses for public health surveillance. Lancet. Reg. Health. Southeast. Asia 14, 100205.

Suchard, M.A., Lemey, P., Baele, G., Ayres, D.L., Drummond, A.J., 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., Cui, N., 2022. Detection of multiple human astroviruses in sewage by next generation sequencing. Water. Res 218, 118523.

Tisza, M., Javornik Cregeen, S., Avadhanula, V., Zhang, P., Ayvaz, T., Feliz, K., Hoffman, K.L., Clark, J.R., Terwilliger, A., Ross, M.C., Cormier, J., Moreno, H., Wang, L., Payne, K., Henke, D., Troisi, C., Wu, F., Rios, J., Deegan, J., Hansen, B., Balliew, J., Gitter, A., Zhang, K., Li, R., Bauer, C.X., Mena, K.D., Piedra, P.A., Petrosino, J.F., Boerwinkle, E., Maresso, A.W., 2023. Wastewater sequencing reveals community and variant dynamics of the collective human virome. Nat. Commun. 14, 6878.

Toribio-Avedillo, D., Gomez-Gomez, C., Sala-Comorera, L., Rodriguez-Rubio, L., Carcereny, A., Garcia-Pedemonte, D., Pinto, R.M., Guix, S., Galofre, B., Bosch, A., Merino, S., Muniesa, M. 2023. Monitoring influenza and respiratory syncytial virus in wastewater. Beyond COVID-19. Sci. Total Environ. 892:164495.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., Rozen, S. G., 2012. Primer3–new capabilities and interfaces. Nucleic. Acids. Res 40, e115.

Wang, H., Churqui, M.P., Tunovic, T., Enache, L., Johansson, A., Lindh, M., Lagging, M., Nystrom, K., Norder, H., 2023. Measures against COVID-19 affected the spread of human enteric viruses in a Swedish community, as found when monitoring wastewater. Sci. Total. Environ 895, 165012.

Wannigama, D.L., Amarasiri, M., Hongsing, P., Hurst, C., Modchang, C., Chadsuthi, S., Anupong, S., Phattharapornjaroen, P., S, M.A., Fernandez, S., Huang, A.T., Kueakulpattana, N., Tanasatitchai, C., Vatanaprasan, P., Saethang, T., Luk-In, S., Storer, R.J., Ounjai, P., Ragupathi, N.K.D., Kanthawee, P., Sano, D., Furukawa, T., Sei, K., Leelahavanichkul, A., Kanjanabuch, T., Hirankarn, N., Higgins, P.G., Kicic, A., Chatsuwan, T., McLellan, A.D., Abe, S. 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., François, R., Henry, L., Müller, K., 2022b. dplyr: A Grammar of Data Manipulation.

Willeit, P., Krause, R., Lamprecht, B., Berghold, A., Hanson, B., Stelzl, E., Stoiber, H., Zuber, J., Heinen, R., Kohler, A., Bernhard, D., Borena, W., Doppler, C., von Laer, D., Schmidt, H., Proll, J., Steinmetz, I., Wagner, M., 2021. Prevalence of RT-qPCR-detected SARS-CoV-2 infection at schools: First results from the Austrian School-SARS-CoV-2 prospective cohort study. Lancet. Reg. Health. Eur 5, 100086.

Wolfe, M.K., Duong, D., Bakker, K.M., Ammerman, M., Mortenson, L., Hughes, B., Arts, P., Lauring, A.S., Fitzsimmons, W.J., Bendall, E., Hwang, C.E., Martin, E.T., White, B.J., Boehm, A.B., Wigginton, K.R., 2022. Wastewater-Based Detection of Two Influenza Outbreaks. Environ. Sci. Technol. Lett. 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., Boehm, A.B., 2023. Use of Wastewater for Mpox Outbreak Surveillance in California. N. Engl. J. Med 388, 570–572.

Wylezich, C., Calvelage, S., Schlottau, K., Ziegler, U., Pohlmann, A., Hoper, D., Beer, M., 2021. Next-generation diagnostics: virus capture facilitates a sensitive viral diagnosis for epizootic and zoonotic pathogens including SARS-CoV-2. Microbiome 9, 51.

Xagoraraki, I., O'Brien, E., 2020. Wastewater-Based Epidemiology for Early Detection of Viral Outbreaks. In: O'Bannon, D.J. (Ed.), Women in Water Quality: Investigations by Prominent Female Engineers. Springer International Publishing, Cham.

Xiang, G., Li, Y., Sun, J., Huo, Y., Cao, S., Cao, Y., Guo, Y., Yang, L., Cai, Y., Zhang, Y.E., Wang, H., 2023. Evolutionary mining and functional characterization of TnpB nucleases identify efficient miniature genome editors. Nat. Biotechnol.

Xu, B., Zhi, N., Hu, G., Wan, Z., Zheng, X., Liu, X., Wong, S., Kajigaya, S., Zhao, K., Mao, Q., Young, N.S., 2013. Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. Proc. Natl. Acad. Sci. USA 110, 10264–10269.

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

Yousif, M., Rachida, S., Taukobong, S., Ndlovu, N., Iwu-Jaja, C., Howard, W., Moonsamy, S., Mhlambi, N., Gwala, S., Levy, J.I., Andersen, K.G., Scheepers, C., von Gottberg, A., Wolter, N., Bhiman, J.N., Amoako, D.G., Ismail, A., Suchard, M., McCarthy, K., 2023. SARS-CoV-2 genomic surveillance in wastewater as a model for monitoring evolution of endemic viruses. Nat. Commun 14, 6325.

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., Cronin, D., Solden, L., Delage, E., Alberti, A., Aury, J.M., Carradec, Q., da Silva, C., Labadie, K., Poulain, J., Ruscheweyh, H.J., Salazar, G., Shatoff, E., Tara Oceans Coordinatorsdouble, d., Bundschuh, R., Fredrick, K., Kubatko, L.S., Chaffron, S., Culley, A.I., Sunagawa, S., Kuhn, J.H., Wincker, P., Sullivan, M.B., Acinas, S.G., Babin, M., Bork, P., Boss, E., Bowler, C., Cochrane, G., de Vargas, C., Gorsky, G., Guidi, L., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Kandels, S., Karp-Boss, L., Karsenti, E., Not, F., Ogata, H., Poulton, N., Pesant, S., Sardet, C., Speich, S., Stemmann, L., Sullivan, M. B., Sungawa, S., Wincker, P., 2022. Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. Science 376, 156–162.