



OPEN ACCESS

EDITED BY

Tiong Gim Aw,
Tulane University, United States

REVIEWED BY

Kathryn Langenfeld,
Stanford University, United States
Walter Betancourt,
University of Arizona, United States

*CORRESPONDENCE

Françoise S. Le Guyader
✉ soizick.le.guyader@ifremer.fr

Received 08 February 2023

ACCEPTED 03 April 2023

PUBLISHED 25 April 2023

CITATION

Schaeffer J, Desdouits M, Besnard A and Le Guyader FS (2023) Looking into sewage: how far can metagenomics help to detect human enteric viruses?

Front. Microbiol. 14:1161674.

doi: 10.3389/fmicb.2023.1161674

COPYRIGHT

© 2023 Schaeffer, Desdouits, Besnard and Le Guyader. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Looking into sewage: how far can metagenomics help to detect human enteric viruses?

Julien Schaeffer, Marion Desdouits, Alban Besnard and Françoise S. Le Guyader*

Ifremer, Laboratoire de Microbiologie, U. Microbiologie Aliment Santé et Environnement, Nantes, France

The impact of human sewage on environmental and food contamination constitutes an important safety issue. Indeed, human sewage reflects the microbiome of the local population, and a variety of human viruses can be detected in wastewater samples. Being able to describe the diversity of viruses present in sewage will provide information on the health of the surrounding population health and will help to prevent further transmission. Metagenomic developments, allowing the description of all the different genomes present in a sample, are very promising tools for virome analysis. However, looking for human enteric viruses with short RNA genomes which are present at low concentrations is challenging. In this study we demonstrate the benefits of performing technical replicates to improve viral identification by increasing contig length, and the set-up of quality criteria to increase confidence in results. Our approach was able to effectively identify some virus sequences and successfully describe the viral diversity. The method yielded full genomes either for norovirus, enterovirus and rotavirus, even if, for these segmented genomes, combining genes remain a difficult issue. Developing reliable viromic methods is important as wastewater sample analysis provides an important tool to prevent further virus transmission by raising alerts in case of viral outbreaks or emergence.

KEYWORDS

sewage, human viruses, method, metagenomics, public health, molecular epidemiology

Introduction

Contaminated food and water are a major pathway for the transmission of infectious diseases, this poses a global health issue, with viruses being one of the most frequently reported infectious agents. Considering the host specificity of most human enteric viruses, despite rare exception such as for hepatitis E virus, this contamination mainly occurs through contact with human sewage that have not been treated properly. People infected with enteric viruses, pathogenic or non-pathogenic, that are derived from the human gut, excreted in the stools and vomitus. Early evidence of the importance of sewage contamination with human viruses emerged in the late 1930's, raising the question of possible transmission of poliovirus through contaminated water (Melnick, 1947; Metcalf et al., 1995). Since then, wastewater has been considered as a major source of viruses in environmental waterways. It is now acknowledged that understanding the viral composition of sewage can be informative with respect to the health of the surrounding population, as exemplified during the on-going SARS-CoV2 pandemic. Indeed, the

analysis of sewage samples may provide information on the dynamics of the virus circulation at local or regional scales, measure lockdown effect and provide information on the number of infected individuals (Wurtzer et al., 2020). Using an optimized sequencing method, the spread of virus variants may also be detected in these sewage samples, making possible the early detection of shifts in emerging strains (Barbé et al., 2022).

Up to now, the major restriction of sewage analysis is that only previously described viruses are searched for and therefore detected. In 1995, a review described studies of sewage analysis up to the development of molecular biology and the possibility to detect a large range of viruses, even though many of these do not grow or grow poorly in cell culture (Metcalf et al., 1995). Since then, even if some questions on the infectivity of particles detected at the genomic level are not completely solved, many studies have demonstrated the power of molecular detection to monitor outbreaks and the usefulness of these approaches (Zheng et al., 2013; Barril et al., 2015; Miura et al., 2016).

Metagenomic developments propose a new step for the analysis of environmental samples. Next-generation sequencing (NGS) or high-throughput deep sequencing allow the massive, parallel sequencing of DNA making possible, which in theory can describe all micro-organisms present in one sample. Such NGS method has been shown to be very promising for sewage analysis as it described the diversity of virus strains circulating in the population (McCall et al., 2020; Nieuwenhuijse et al., 2020; Yang et al., 2021). In an era with increasing demographic population and an increased demand for clean water including wastewater recycling, knowing what is in our sewage will help to prevent further disease (Aarestrup et al., 2021). Consistent with this, some food such as bivalve molluscan shellfish may be contaminated by sewage, leading to consumer diseases or economic losses by food destruction (Savini et al., 2021). In these settings, human enteric viruses are frequently implicated in foodborne outbreaks (Torok et al., 2019; Savini et al., 2021). However, characterization of human enteric viruses that have short RNA genomes is challenging due to the high levels of nucleic acids from cellular organisms, the low concentrations of some viruses, and the high number of unclassified sequences (Greninger, 2018; Cobbin et al., 2021; Mazur et al., 2022). A few studies have reported the use of metagenomics to describe viral diversity in water or sewage samples, with encouraging results in the identification of different types of human virus (Kim et al., 2017; Nieuwenhuijse and Koopmans, 2017; Hendriksen et al., 2019; Martinez-Puchol et al., 2020; Adriaenssens et al., 2021). However, most of the time only amplicon-based metagenomic approaches have solved the issue of sensitivity to describe the genetic diversity of one virus genus (Suffredini et al., 2018; Fumian et al., 2019; Mabasa et al., 2022). In previous works, we found that sample pre-treatment and capture-based enrichment during the library preparation stage allowed the recovery of higher numbers of viral reads and thus the identification of longer viral contigs (Strubbia et al., 2019; Bonny et al., 2021).

Here the previously developed method was applied for the analysis of archived wastewater samples with the aim to evaluate the benefits of performing technical replicates to improve viral identification, with a focus on virus that may cause food contamination.

Materials and methods

Samples collection

Twelve archival samples collected from six wastewater sewage treatment plants (WWTP), all located in northwestern France within 200 km of each other, were selected (Figure 1). All WWTP received sewage from a comparable number of inhabitants (8,000 to 58,000), except one which served less than two thousand inhabitants (D; Table 1). One sewage treatment plant (B) was located on a small island, 10 km away from the continent.

All samples but one (2006-D) were 24-h composite raw sewage samples, and had been stored at $-20 \pm 1^\circ\text{C}$ in 1-liter aliquots before the metagenomics study (storage length varying from 14 to 2 years).

Screening method

The 12 samples were concentrated from 40 ml raw sewage by adding 10 ml of 50% polyethylene glycol (PEG 6,000; Sigma-Aldrich, St-Quentin France; Sima et al., 2011). After incubation and centrifugation, the PEG pellet was suspended in 1 ml of deionized distilled water (DDW) with a vortex mixer, and nucleic acids were extracted with an automatic easyMAG extractor (bioMérieux, Lyon, France) and the NucliSENS kit (bioMérieux). Norovirus detection was performed as described below.

Sample preparation for metagenomic analysis

All 12 samples were analyzed under four technical replicates (Figure 2). The 48 extractions were performed by series of six samples rather than by biological replicates. After sample thawing, each of the 48 replicates (12×4) of 40 ml was incubated for 40 min with sodium pyrophosphate (10 mM final concentration) under gentle agitation at room temperature before three cycles of sonication for 1 min at maximum power in a cup-horn adaptor (Bandelin, HD 2200), followed by 1 min on ice (Bisseux et al., 2018; Strubbia et al., 2019). After centrifugation for 20 min at $8,000 \times g$, supernatants were recovered, the pH adjusted to 7 (using HCl), and then 10 ml of 50% of PEG 6,000 were added. After overnight incubation at 4°C under gentle agitation, the mixture was centrifuged for 1.5 h at $13,500 \times g$, the pellet was suspended in 2 ml of 0.05 M glycine buffer (pH 9).

Purification steps and nucleic acid extraction for metagenomics analysis

The re-suspended pellets were filtrated using a cascade of 5, 1.2, and $0.45 \mu\text{m}$ filter pores (Minisart NML 17594, NML17593, PES16533) to eliminate cells and the filtrate treated with 2000 units of OmniCleave Endonuclease (Lucigen, Wisconsin, USA) for 1 h at 37°C to eliminate free DNA and RNA. Nucleic acids (NA) were extracted by adding 10 ml of a guanidium-based lysis buffer (bioMérieux) and $140 \mu\text{l}$ of paramagnetic beads (NucliSENS kit, bioMérieux; Strubbia et al., 2019). After washing steps using the eGENE-UP[®] apparatus (bioMérieux)

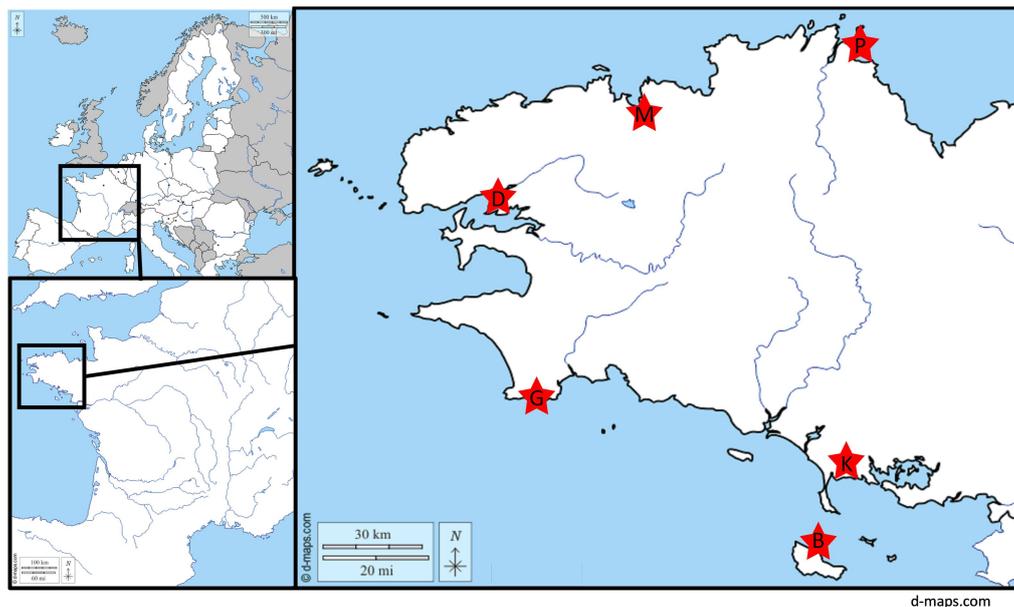


FIGURE 1
Localization of the 6 wastewater treatment plants collected for this study.

TABLE 1 Sample description and norovirus concentrations.

Sample	Sampling date	Inhabitants*	Average Daily flow	NoV conc.**	
				Screening	cDNA control
2006-D	Jan. 10, 2006	1,700	496	2.45×10^5	3.32×10^5
2006-G	Jan. 24, 2006	26,000	889	4.34×10^5	6.78×10^5
2006-M	Feb. 28, 2006	58,000	6,663	2.85×10^5	2.90×10^5
2006-G2	Jul. 23, 2006	26,000	791	2.78×10^5	2.88×10^5
2007-P	Jul. 23, 2007	22,000	2,101	2.13×10^4	6.51×10^4
2009-P	Dec. 29, 2009	22,000	4,545	2.26×10^4	5.03×10^4
2012-K	Oct. 17, 2012	14,950	1,853	1.85×10^3	5.52×10^3
2014-K2	Feb. 05, 2014	21,500	1,876	2.02×10^4	5.12×10^4
2014-K1	Mar. 12, 2014	21,500	1,876	2.25×10^4	6.87×10^4
2014-K	Mar. 19, 2014	21,500	1,876	1.75×10^5	7.21×10^5
2016-B [†]	Jun. 02, 2016	8,000	1,152	2.20×10^5	8.91×10^5
2016-B2 [†]	Jun. 06, 2016	8,000	1,152	7.16×10^4	6.02×10^4

*Treatment capacity (<https://www.assainissement.developpementdurable.gouv.fr/PortailAC/data.php>).

**Norovirus geometric mean concentration expressed in RNAc/L, the one-step corresponds to concentrations measured after nucleic acid extractions, and the cDNA control to concentrations measured on the cDNA synthesized for the library preparation.

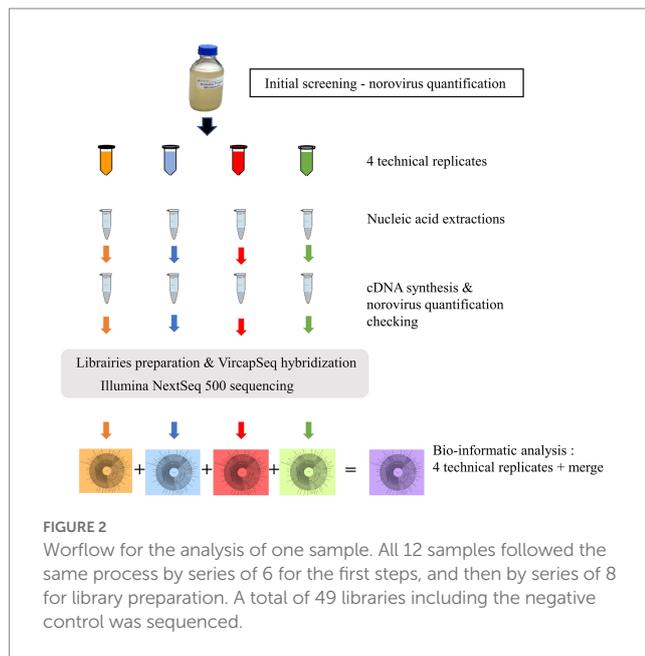
[†]Sewage treatment plant located on a small island, 10 km away from the continent.

following manufacturer's instructions, extracted NA were treated with 25 U TURBO™ DNase (Ambion, Thermo Fisher Scientific, France) for 30 min at 37°C. An additional RNA purification was performed to remove DNase and PCR inhibitors (RNA Clean & Concentrator™ -5 kit, Zymo Research, Irvine, USA).

Norovirus quantification

Norovirus quantification was performed for the two main genogroups I and II (GI and GII) using primers and probes

targeting the conserved region at the beginning of ORF2 (Kageyama et al., 2003; Loisy et al., 2005; da Silva et al., 2007; Svraka et al., 2007). The rRT-PCR was carried out with the Ultrasense One-step quantitative RT-PCR system (Life Technologies, France) using 5 µl of extracted NA per well (final volume of 25 µl) in duplicate on a Mx3000P QPCR System (Agilent Technologies, France). Standard curves based on double strand DNA (gBlocks, IDT USA) corresponding to nucleotides 4,484 to 5,668 of the GI.1 Norwalk virus (Genbank M87661) or nucleotides 4,217 to 5,355 of the GII.4 Houston virus (Genbank EU310927), were included in each run. Concentrations range from 8.1 and 3.4 RNA copies/µl to 8.1×10^4



and 3.4×10^4 RNA copies/ μl for norovirus GI.1 and GII.4, respectively.

Two negative amplification controls (sterile, RNase-free water) were included in each amplification series and filter tips and dedicated rooms were used to prevent false-positives. Standard curves included in each run were analyzed and retained for NA quantification if amplification efficiencies were comprised between 85 and 110%. For the screening, concentrations were calculated per sample by adding GI concentrations to GII. When samples were prepared for metagenomics, concentrations obtained for each replicate (4 per sample), were calculated separately, then GI and GII concentrations were added and the four values obtained were used to calculate the geometric mean concentration (GMC) per sample.

Library preparation and sequencing

Libraries were prepared by series of eight. All the 48 NA extracts were transcribed into cDNA using the Superscript II (Invitrogen, Saint-Aubin, France) and random hexamers (New England Biolabs (NEB), USA). Completion of this step was checked by PCR quantification for norovirus GI and GII as described above, except that enzyme mix was replaced by Platinum taq (ThermoFisher Scientific) and the RT step was removed from the thermal profile.

After double strand DNA synthesis using second strand reaction buffer from NEBNext Ultra RNA Library prep, physical fragmentation of the DNA was performed using sonication (Ultrasonicator Covaris M220, duty factor 5%, peak power: 75, cycles per dust: 200 for 195 s, Woburn, MA) and libraries were prepared using the kit KAPA Hyper Prep (Roche), according to the manufacturer's instructions. Viral sequences were enriched using VirCapSeq-Vert probe panel (Roche; Briese et al., 2015; Strubbia et al., 2020). A negative control, in the form of sterile RNase-free water, was treated in parallel from cDNA synthesis onwards. All 49 libraries were sequenced using Illumina NextSeq 500 technologies to generate 2×150 bp reads.

Bioinformatic analysis

Bioinformatic analysis was performed using a Nextflow pipeline as previously described (Bonny et al., 2021). For the first part of the analysis, reads from each library as well as reads merged from all four libraries per sample were considered (5 analysis for each sample; Figure 2). Briefly, Fastq were trimmed using fastp with a quality threshold at 25. Clean reads were deduplicated (CD-hit) and mapped to remove bacterial RNA reads (Silva RNA database) and PCR duplicates. *De novo* assembly was then performed using metaSPAdes with kmer length 21, 33, 55, 77, 99 (Nurk et al., 2017). Contigs longer than 500 bp were filtered and identified using BLASTn compared to *nt* database (download 2022-05-08) with an *e*-value of 10^{-5} , and *nr* database (download 2021-02-06) using diamond with an *e*-value of 10^{-3} . When both approaches gave results, the BLASTn match was kept. To evaluate the coverage of each contig, post-process reads were mapped using Bowtie2 (v2.3.0; Langmead and Salzberg, 2012) on the metaSPAdes contigs. Multi-mapped reads were removed as a source of potential overestimation of the abundance. Coverage was calculated using the Lander-Waterman equation $C = \frac{N \times (R)}{L}$, where *C* is the coverage, *N* is the number of reads, *R* the length of the reads (150 bases) and *L* the length of the considered contig (Lander and Waterman, 1988). After this step, only the "merged" dataset was considered for each sample.

Taxonomic identification was done using Entrez direct tool, the taxid allowed to extract information at a defined taxonomic level. Reads per millions (rpm) were calculated using the number of reads per family and the total number of reads after trimming and deduplication. The heatmap was done using Graphpad prism v 9.0.0 (GraphPad Software, sanDiego, CA, US).

Genotyping of viral sequences was performed using several web-based tools, and only contigs represented more than 100 reads were considered. For norovirus, sapovirus or hepatitis A & E virus sequences, the identification was done using the respective online Typing Tools and for rotavirus the Rotac tool (Maes et al., 2009; Kroneman et al., 2011). Other contigs identified as enterovirus, mamastrovirus and picobirnavirus sequences were characterized using Genome Detective (Vilsker et al., 2019).

Phylogenetic analysis were performed by aligned sequences using Muscle or ClustalW2, followed by IQ-TREE algorithm to choose the best fit model according to the BIC (Bayesian Information Criterion) and to construct phylogenetic trees with a bootstrap of 1,000.

Results

Samples were collected from sewage treatment plants from relatively small communities over a period of 10 years (Table 1). Eight samples were collected during winter months (October to March) and four during summer (June and July). Despite these different seasons, following the initial screening, all samples were positive for norovirus GI and/or GII and there were no differences in total norovirus concentration between winter and summer samples (Table 1).

For metagenomics, the 12 samples were treated and extracted 4 times each (Figure 2).

After the nucleic acid extractions of the four replicates from each of the 12 samples, 29 were positive for norovirus GI and 40 for norovirus GII. After cDNA synthesis, 22 extracts were positives for

norovirus GI and 45 for norovirus GII (data not shown). One replicate of sample 2009-P did not give a GI or GII Ct value and none of the reads were identified as norovirus. Norovirus concentrations, expressed as GI+GII, obtained after cDNA synthesis were similar or half-a log higher compared to concentrations obtained for the RNA (Table 1). All four replicates were kept for the library preparation to ensure that at least three replicates would be obtained.

Among the 818 million of reads obtained after the Illumina sequencing, 44 libraries gave comparable read numbers ranging from 1,565,594 to 85,453,164 reads, and four libraries yielded a low number of reads (17,134 to 736,618 reads; Table 2). More than 35,000 reads were obtained from the negative control but after the different cleaning steps and identification steps no viral read was identified.

Some samples (2006-M, 2006-G2, 2014-K2, and 2016-B2) presented a large variability between replicates when considering raw reads, clean reads or viral reads, while some others (2006-D, 2006-G, 2007-P, and 2014-K) presented a similar number of reads for all replicates (Figure 3). The coefficient of variation calculated to evaluate viral reads variability between replicates showed over 100% for five samples, and only two samples displayed a coefficient of variation $\leq 50\%$ (Table 2). In addition to this between analysis replicate, data from the 4 replicates were merged and analyzed together for each sample. When considering merge analysis, the variability between samples in terms of raw reads decreased (from 13,721,286 reads obtained for sample 2006-D to 140,560,834 reads for sample 2016-B) compared to individual libraries (from 17,134 reads obtained for replicate 1 sample 2006-M to 85,435,164 reads for replicate 3 sample 2006-G2; Table 2). The percentage of viral reads varied from 45% (sample 2009-P) to 93% (sample 2006-M) after merged replicates analysis (Table 2).

To further decide which approach allows better viral identification, the length of contigs obtained for three genera of enteric viruses (mammastrovirus, enterovirus and norovirus), were compared considering contigs obtained for each separate replicate (blue) and with merged replicates (orange; Figure 4). For mammastrovirus, only samples 2006-M and 2012-K presented a longer contig with one replicate analysis compared to the merged replicate, however these long contigs had a low coverage (Figure 4B). For all the other samples, the merged-replicates analysis provided longer contigs with increased coverages, for example for samples 2007-P, 2012-K, 2014-K1, 2014-K, and 2016-B2. For enterovirus, the merged-replicates analysis increased the contig length for all samples except two. A clear improvement was observed for sample 2016-B for which the merged analysis allowed the assembly of a complete genome that remained fragmented in the separate replicates, and for sample 2016-B2 by increasing the coverage of the different contigs (Figure 4A). Similar observations were obtained for norovirus, with complete genomes identified using the merged-replicates analysis and a good coverage as observed for sample 2006-M (Figure 4C). The long contig obtained with only one replicate for sample 2016-B2 was based on few reads and was not confirmed with the merged-replicate analysis.

Therefore, only the merged-replicate datasets were used for subsequent analyses.

Viral family identification

An initial taxonomic identification was conducted directly on the clean reads of merged replicates. Among the 76,076,257 viral reads,

the large majority of reads were identified as sequences related to viral families infecting humans such as *Astroviridae*, *Caliciviridae*, *Hepeviridae*, *Picobirnaviridae*, *Picornaviridae*, and *Reoviridae* (Figure 5B). Reads belonging to the *Astroviridae* and *Picobirnaviridae* families were detected in all samples, representing 51% of total reads (Figures 5A,B). A large proportion of reads (17%) were identified as belonging to the *Reoviridae* family, but these were not detected in all samples. The same observation can be made for *Caliciviridae* or *Picornaviridae* reads. For eight samples, viral reads belonging to the six families cited above represented the majority of reads with less than 10% of reads identified as other viruses. Conversely, for the four other samples (2006-G, 2007-P, 2009-P, and 2016-B2), the proportions of 'other viruses' (in which we included different families such as unclassified ribovirus, *Nodaviridae*, *Tombusviridae*, *Virgaviridae*, or *Marmaviridae*) ranged from 17 to 63% (Figure 5C). Reads belonging to the *Hepeviridae* family were identified only in two samples (2014-K and 2016-B).

After assembly, contigs were identified using different tools (see method section) and 1,375 contigs were identified as belonging to nine selected viral families comprising known human pathogens (Figure 6): seven ssRNA virus families and two dsRNA viruses with segmented genomes (rotavirus and picobirnavirus). For more precise identification, we focused on contigs longer than 1,000 bases, except for the segmented viruses such as rotavirus (sequences representing at least 97% compared to reference segment length) or picobirnavirus (sequences representing at least 90% compared to reference segment length). For mammastrovirus, considering the high number of long contigs, only complete genomes were analyzed thoroughly (Supplemental Table S1). For picobirnavirus, numerous contigs were obtained with a high number of complete segments (1 and 2), mainly related to human or dog strains.

Long contigs or complete genomes were obtained for different viral genera with a clear relationship to read numbers (Figure 6). However, for norovirus or orthohepevirus, even with less than 10^4 reads, complete genomes were assembled. For kobuvirus, all the different samples provided a low number of reads and contigs obtained were all under 1 kb.

Within the *Caliciviridae* family, when considering contigs longer than 1 kb obtained with more than 100 reads, 32 contigs were identified as norovirus and three as sapovirus. A large range of norovirus read numbers per sample was observed from 264 reads (sample 2009-P) to 1,309,920 reads (2006-G2; Figure 6; Supplemental Table S1). Only sample 2012-K presented both norovirus contigs and sapovirus contigs. In this sample, two sapovirus contigs were identified as a GII (ON807341, ON807342), and the third one as a GII.1 (ON807343). All the other contigs were identified as norovirus, with sequences identified either in the ORF1 or ORF2 typing regions (Table 3).

No clear relationship between norovirus concentrations measured by RT-qPCR, and the abundance of norovirus reads, was observed. For example, sample 2012-K that presented the lowest norovirus concentration and 1,360 rpm, whereas samples 2014-K1 or 2016-B2 provided lower rpm numbers with norovirus concentrations 10 times higher. Two samples (2006-D and 2007-P) for which none of the cDNA replicates were positive for norovirus GI by PCR provided no norovirus GI reads, while some GII reads were identified. Similarly, no norovirus reads were obtained from the replicate of sample 2009-P that did not provided any Ct value. Norovirus sequences were

TABLE 2 Overview of read numbers obtained for the 48 libraries and the merge value for the 12 samples.

Name	Rep.*	Raw reads	Clean reads	Identified reads	Viral reads	Coef. of var. %**	Viral reads %***
2006-D	1	4,506,340	688,804	387,983	361,518	40	
	2	2,402,894	393,000	266,782	249,453		
	3	3,606,002	393,464	289,073	288,018		
	4	3,206,050	312,984	120,321	119,597		
	Merge	13,721,286	1,788,252	1,323,336	1,272,409		71
2006-G	1	9,541,300	2,260,746	1,488,848	990,656	88	
	2	11,217,902	1,530,728	1,274,720	1,261,356		
	3	2,966,010	356,374	209,670	208,750		
	4	6,423,356	598,924	340,585	335,904		
	Merge	30,148,568	4,743,756	3,699,942	3,167,343		67
2006-M	1	17,134	5,236	2,582	2,582	106	
	2	47,644,656	4,573,184	4,187,147	4,173,108		
	3	14,262,232	1,191,024	930,015	928,085		
	4	76,825,830	7,576,116	7,053,620	7,026,021		
	Merge	138,749,852	13,343,538	12,436,823	12,386,553		93
2006-G2	1	29,093,288	4,081,130	3,522,608	3,041,567	123	
	2	8,137,376	680,206	562,770	562,451		
	3	85,435,164	8,064,160	7,529,846	7,509,918		
	4	1,565,594	132,970	16,431	14,371		
	Merge	124,231,422	12,953,180	11,718,516	11,211,954		87
2007-P	1	5,774,206	92,5,076	464,453	410,478	50	
	2	9,458,542	919,580	620,486	611,586		
	3	8,181,372	862,832	676,385	673,825		
	4	3,455,326	338,988	162,431	158,685		
	Merge	26,869,446	3,044,114	2,149,086	2,061,661		68
2009-P	1	3,772,094	698,092	374,629	213,068	67	
	2	4,836,042	587,498	407,026	300,583		
	3	372,442	37,518	22,264	17,321		
	4	5,450,408	448,972	209,349	180,879		
	Merge	14,430,986	1,770,770	1,107,117	795,808		45
2012-K	1	14,176,218	4,587,968	1,060,613	558,029	82	
	2	9,356,070	2,572,568	2,110,438	1,684,264		
	3	3,554,214	450,022	197,616	137,633		
	4	36,261,558	2,545,060	2,086,422	2,038,462		
	Merge	63,348,060	10,150,928	5,871,134	4,810,896		47
2014-K2	1	1,856,764	558,526	256,651	98,446	109	
	2	3,872,212	588,342	324,504	213,169		
	3	52,270,136	3,846,026	3,371,485	3,300,670		
	4	736,618	263,830	20,151	2,553		
	Merge	58,735,730	5,256,534	4,041,606	3,686,030		70
2014-K1	1	12,176,666	2,336,528	1,514,188	984,878	87	
	2	16,887,440	1,544,840	1,115,033	779,709		
	3	3,298,316	298,646	162,226	150,098		
	4	3,114,878	239,354	128,259	118,964		
	Merge	35,477,300	4,418,862	3,153,740	2,248,319		51
2014-K	1	9,137,872	1,418,210	877,541	789,752	177	
	2	12,253,584	1,132,042	858,364	850,963		
	3	5,421,844	447,440	222,963	218,749		
	4	39,869,020	4,487,688	3,574,163	3,444,980		
	Merge	66,682,320	7,485,152	5,957,168	5,720,168		76

(Continued)

TABLE 2 (Continued)

Name	Rep.*	Raw reads	Clean reads	Identified reads	Viral reads	Coef. of var. %**	Viral reads %***
2016-B	1	10,936,716	2,801,174	1,762,383	1,440,351	81	
	2	68,870,930	8,652,778	7,858,150	7,735,805		
	3	39,108,178	3,875,918	3,124,556	3,042,426		
	4	21,645,010	2,476,722	2,031,273	2,012,830		
	Merge	140,560,834	17,795,292	15,290,037	14,708,048		
2016-B2	1	24,749,188	5,210,512	3,080,758	2,350,567	117	
	2	48,291,662	13,195,394	11,603,668	9,160,827		
	3	715,534	104,526	76,346	75,782		
	4	31,200,266	3,267,994	2,112,003	2,029,485		
	Merge	104,956,650	21,772,204	17,266,476	14,007,068		
Control		35,160	11,458	414	0		

*Data are provided for each replicate (rep.) and for the merge value of all four replicates for each sample (grey line).

**Coefficient of variation is calculated as the ratio of the standard deviation σ to the mean μ , $c_v = \frac{\sigma}{\mu}$.

***Percentage of viral reads calculated based on viral reads divided by clean reads, and multiplied by 100.

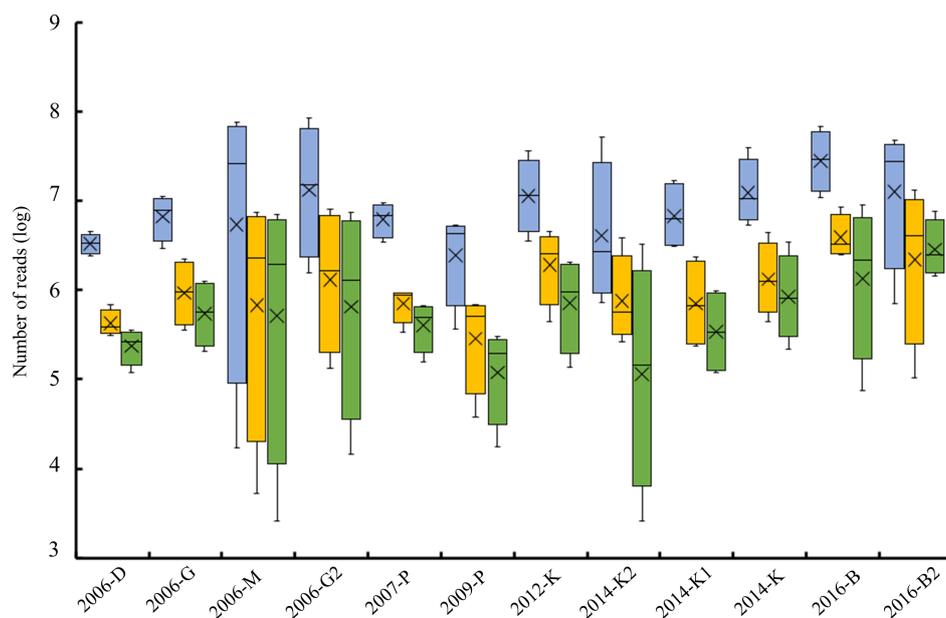


FIGURE 3

Schematic representation of reads obtained for the four replicates analyzed per sample. Boxplots represent the number of raw reads (blue box), after trimming and filtering (orange box), and viral reads (green box) for each sample. Boxplots show the minimum, 25th percentile, average (cross), median (line), 75th percentile and maximum number of reads.

genotyped using the Typing tool and the Noronet database, for the ORF1 and ORF2 regions separately (Table 3). Overall, we identified four different GI P-types, five GII P-types, four GI genotypes and six GII genotypes. Four complete genomes assigned as GII.6[P7] (ON706286, ON706287, ON706288, and ON706289) were identified in four samples collected in 2006, 2007 and 2014. Two samples collected in 2006 and 2016 provided almost complete genomes of two GII.6[P7] strains (ON706290 and ON706291). For some samples, despite a high number of reads, the contigs obtained did not allow a more precise identification than the genogroup assignation, because of the contig position on the genome outside of the typing regions in

ORF1 or 2. Samples collected serially in the same WWTP show a high variability in the identified circulating norovirus strains. For instance, in 2014 in WWTP K, most of contigs were identified as GII.6 and GII.[P7], which were not detected in the sample 2014-K1 collected one week before. However, this sample provided very low number of reads.

Among the *Hepeviridae* family, four contigs were identified as orthohepevirus from samples collected in site B and K (Figure 6). One contig was short (<1 kb), but the three other contigs were complete genomes (ON807340, ON807339, ON807338) identified as hepatitis E virus, genotype 3f (blast identification 98%, Supplemental Table S1). Sample 2016-B2 gave a complete genome (7,254 bases) assembled

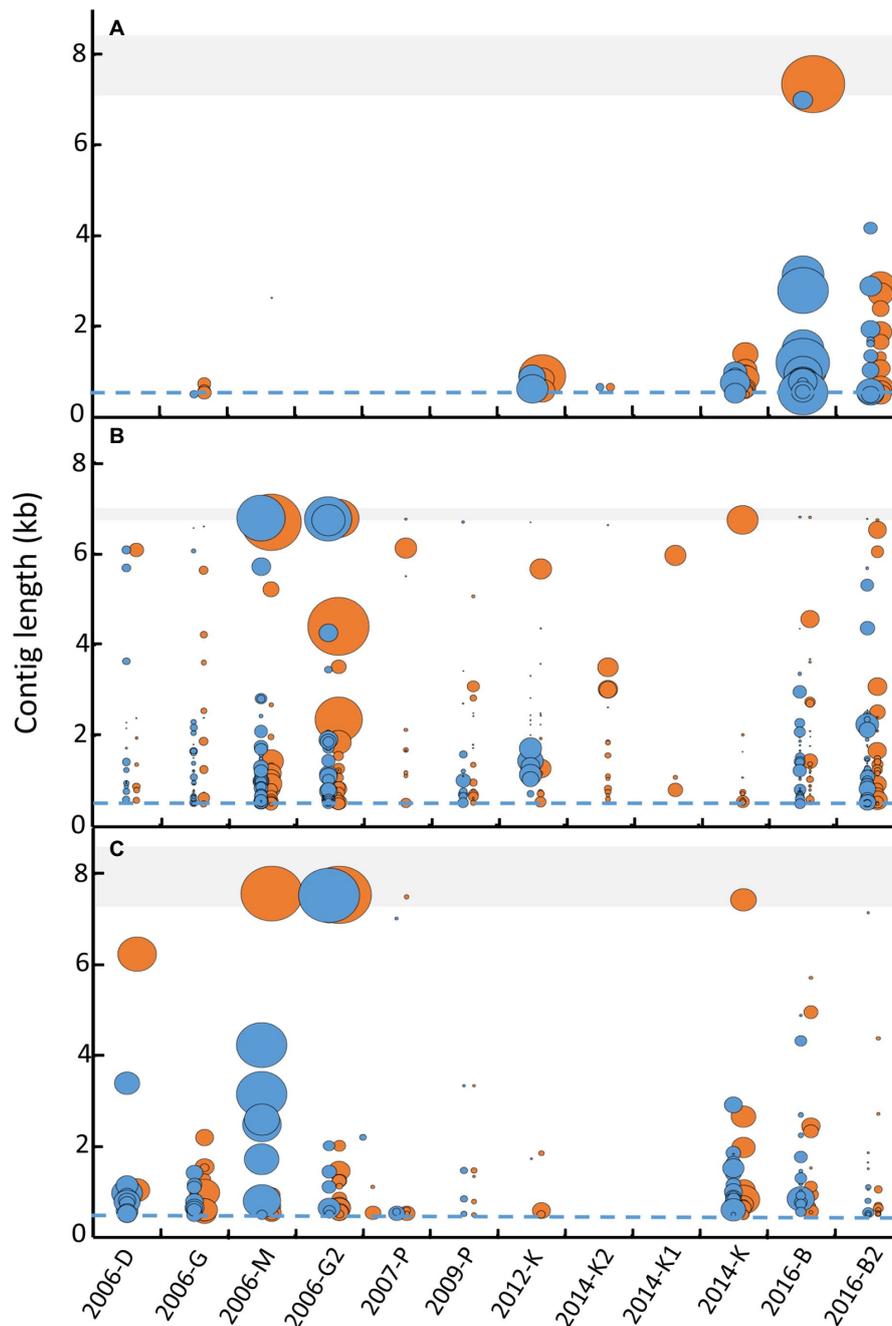
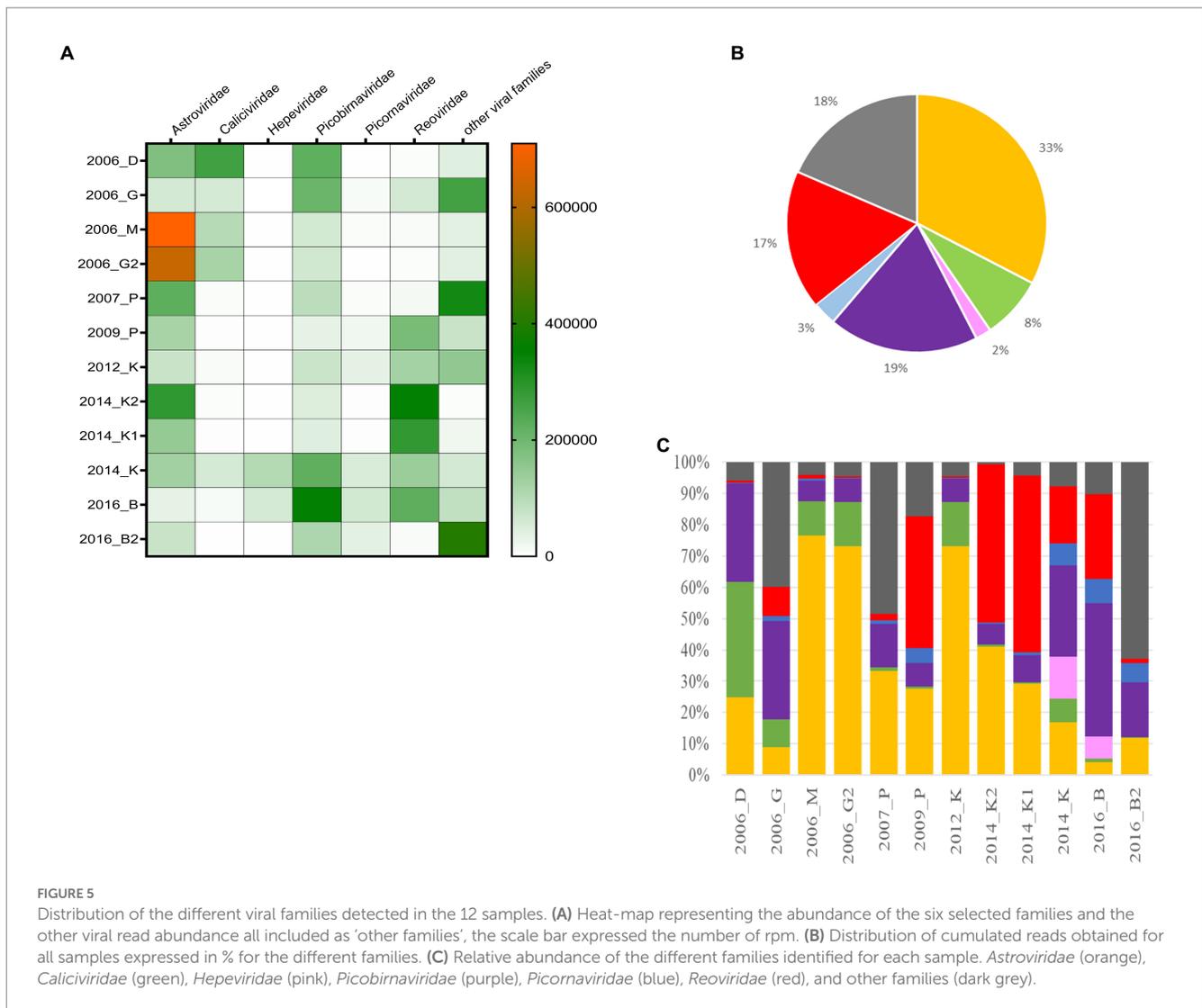


FIGURE 4

Comparison of contig lengths identified after separate library analysis or merged analysis for three of ssRNA virus genus. Contigs longer than 500 bases (dashed blue lines) obtained for each library (blue dots) and the merge analysis (orange dots) were reported for the 12 samples for enterovirus (A), mammastrovirus (B), and norovirus (C). For each viral family, the shaded area represents the expected size of complete genome (7.2 to 8.5 kb for enterovirus, 6.8 to 7 kb for mammastrovirus and 7.3 to 8.3 kb for norovirus). The size of the dots represents the coverage obtained for each contig (calculated by the number of bases obtained from all reads divided by the length of the contig) and represented as a ratio between coverage of the contig considered divided by the highest coverage obtained for the genus.

from a relatively low number of reads (2,991 reads), corresponding to a mean coverage of 123 reads. The two other complete genomes were obtained with much higher numbers of reads (737,250 reads for sample 2014-K and 1,012,569 reads for sample 2016-K) with mean coverages of 32,162 and 41,573 reads, respectively. These three contigs were identical (Figure 7) and similar to strains circulating in France 2011 and 2012.

Among the *Picornaviridae* family, 40 contigs were identified. Of these contigs, 16 were assigned as Beihai or Wenzhou picorna-like viruses, nine as salivirus A, two as parechovirus A, two as hunnivirus, one as palmarnavirus and the other contigs remained unassigned. A full genome of hepatitis A virus was obtained from a sample collected in WWTP B, the only site in which hepatitis A virus contigs longer than 1 kb were obtained. The two other contigs obtained from this



sampling site B were identical to the full genome, identified as a genotype IA (Supplemental Table S1).

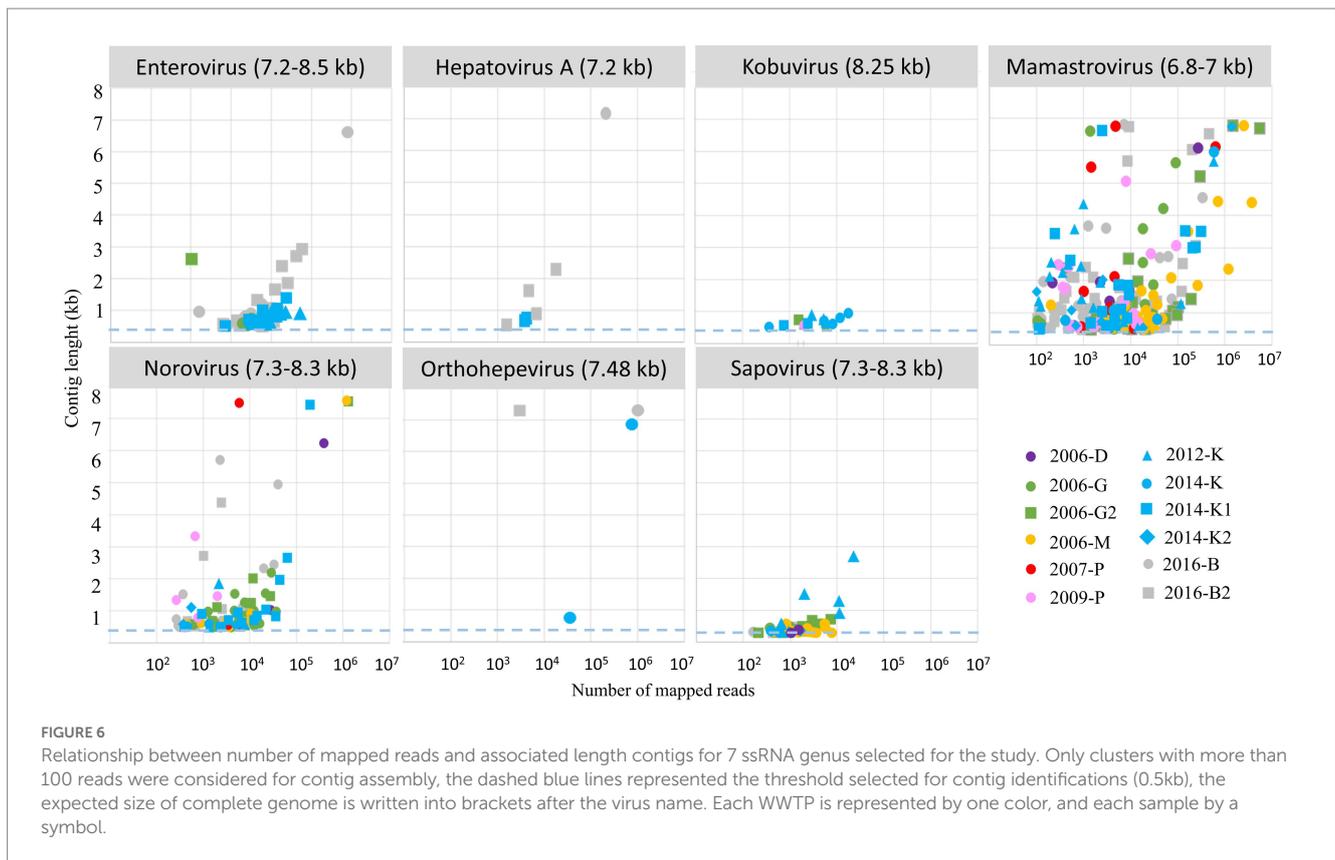
This WWTP B showed also the highest diversity for enterovirus sequences (Table 4). Ten enterovirus contigs were detected in seven samples collected in three WWTP (G, K and B), and were most frequently identified as enterovirus type A and B. One complete genome (6,619 nucleotides) of a coxsackievirus A9 strain was identified in the first sample collected in June in WWTP B. A week later in the same WWTP, up to 5 different contigs were identified, including a coxsackievirus A9 sequence presenting 65% homology over 647 bases with the previous one. In two samples collected in 2016 in WWTP B, two contigs were identified as the enterovirus type EV-A114, an enterovirus that was described in childrens stool samples in India (Deshpande et al., 2016).

For the *Astroviridae* family, 128 of contigs longer than 1 kb were assigned as belonging to mamastrovirus genus. A large majority, 63% (80 contigs) of the contigs, were identified as mamastrovirus 1 (HAstV-1), using the genome detective web tool. This genus was represented in all samples. Feline astrovirus 2 (FAstV-2) was the second most identified genus (17%) with 22 contigs. Other contigs were identified as mamastrovirus 2 (HAstV-2) 1 contig, mamastrovirus

5 (HAstV5) 1 contig and mamastrovirus 9 (HAstV-9) 3 contigs. Twenty one contigs were close to reference sequences that were not been approved as a species by the ICTV, but belong to the mamastrovirus genus: 16 contigs of Human astrovirus MLB2 (HAstV-MLB2), 4 contigs of the Culex Bastrovirus-like virus and 1 contig of Canine astrovirus (CaAstV-1). Twelve long contigs (more than 6kb) were found in all samples except sample 2009-P.

Among the complete genomes obtained in the different samples, six clustered to HAstV-6, and six to HAstV-1. Phylogenetic analysis of the complete RdRP protein sequence confirmed identification of the complete HAstV genomes compared to previously available sequences in GenBank (Figure 8).

Picobirnaviridae reads were identified in all samples (Figure 5A), and up to 793 contigs (longer than 500 bases) were identified as picobirnavirus. The number of contigs varied largely depending on the samples, from 6 contigs in sample 2014-K2 to 259 contigs in sample 2016-B2. Picobirnavirus genome is constituted of two segments and 13% of contigs matched to segment 1 (encoding hypothetical proteins) with four sequences being almost complete (around 2,525 bases). All the other contigs (87%) were identified as segment 2 (RNA-dependant RNA polymerase coding gene), with 70



contigs being complete sequences (1745 bases). Among these complete or nearly complete segments, only 11 showed sequence similarity with human picobirnavirus (NC_007027.1). The other contigs showed sequence similarity with otarine picobirnavirus (NC_034160.1; 46 contigs), porcine picobirnavirus (NC_029802.1; 12 contigs), roe deer Picobirnavirus (NC_040753.1; 4 contigs) and 1 Chicken picobirnavirus (NC_040439.1; 1 contig).

Rotavirus genes were detected in all samples with 26 complete genes segments being sequenced in several samples (Table 5). Only one complete VP1 gene was obtained, as well as one VP7, while five, VP2, eight VP3, two VP4, two VP6 were obtained. Regarding non-structural proteins, only two NSP3 genes, one NSP1 and one NSP2 were identified as full sequences. Twenty-seven nearly complete genes were identified. Contigs coding for VP7 and VP4 allowed the identification of RVA G1P[8], for which most of the other genomics segments were identified in all samples (Table 5). Moreover, in samples 2016-B and 2014-K1, all genomics segments previously identified as belonging to this strain could be sequenced, likely representing full RVA G1P[8] Wa-like genomes (Table 5). Evidence for co-circulating RVA strains were observed in several samples. For example, four samples displayed three different complete VP7 gene segments (G1, G3 and G9), while 6 other samples contained either G1 and G3 or G1 and G9 complete VP7 gene segments. In most samples, the VP4 genes were identified as P[8] except for sample 2016-B2 for which a P[4] and a P[1] VP4 genes were also identified. In other samples, two different VP1 sequences, or up to 3 VP2 or NSP3, could be detected. Segments that may correspond to the constellation of the Human DS-1 like strain were detected in three samples while segments from the AU-1-like constellation were identified in 9 samples (Table 5, red and orange,

respectively). NSP4 (only one complete gene) and NSP5 segments were rarely detected, and the few sequences obtained were identified as belonging to the constellation of Human Wa-like strain. A NSP3 sequence identified as T6 was identified once (Table 5, purple).

Discussion

As mentioned above, the development of metagenomic approaches opens new perspectives in environmental virology through their capacity to sequence all the genomes present in a sample. Strong evidence of the potential of metagenomics in environmental virology have been provided for example by the TARA expedition which brought new data on the ocean viromes or by the analysis of sewage viromes (Nieuwenhuijse et al., 2020; Sunagawa et al., 2020). Using metagenomic analysis, a study was able to suggest a conceptual model of virus circulation considering human and livestock inputs throughout the fresh-marine continuum of a river catchment (Adriaenssens et al., 2021). A good complement to such approach would be to increase the identification of viral sequences that may impact human health. Indeed, in virome studies, most of recovered contigs belong to phages, or animal viruses making human virus sequence identification difficult as some animal viruses are closely related to Human strains (Guajardo-Leiva et al., 2020; Martinez-Puchol et al., 2020; McCall et al., 2020). Human enteric viruses usually have single-strand RNA genomes, relatively short (around 7 kb), some being segmented as for rotavirus, making precise identification complex. Moreover, these RNA molecules are fragile and may

TABLE 3 Heatmap profile showing the relative abundance the different norovirus genotype sequences detected in the 12 samples.

	ORF1										ORF2								
	GI.[P1]	GI.[P8]	GI.[P9]	GI.[P13]	GI.[P7]	GI.[P17]	GI.[P21]	GI.[P25]	GI.[P31]	GI.1	GI.3	GI.8	GI.9	GI.3	GI.6	GI.7	GI.12	GI.14	GI.17
2006-D	0	0	0	0	230,961	0	0	0	0	0	0	0	0	0	215,421	0	0	0	0
2006-G	0	0	1,257	1,104	19,173	0	315	0	0	0	0	0	351	778	13,198	1,332	1,098	4,246	0
2006-M*	67	0	0	0	91,892	0	0	0	0	0	32	0	0	0	89,428	580	0	1	0
2006-G2*	1,019	0	2,249	0	104,416	0	0	0	0	560	0	0	1,017	991	101,242	1,179	0	537	0
2007-P*	0	0	0	0	3,020	0	0	0	0	0	0	0	0	0	1,918	4	0	1,829	0
2009-P	0	0	0	0	2,067	0	0	0	0	0	0	0	0	0	614	0	0	0	0
2012-K	0	0	0	0	229	0	0	0	0	0	132	0	0	0	230	0	0	770	0
2014-K2	0	0	0	0	1,054	0	0	0	0	0	0	0	0	0	104	0	0	0	0
2014-K1	0	103	0	0	0	3	0	0	0	0	0	0	296	0	0	0	0	0	3
2014-K*	0	0	0	0	50,292	325	0	2,805	311	0	0	0	0	0	34,240	0	0	0	2,361
2016-B	1	0	1	0	573	1,482	0	2,262	0	0	0	0	1	0	597	2	0	16	1,823
2016-B2	16	0	0	0	199	166	0	18	0	0	182	0	1	1	46	4	0	0	67

The number of reads per million (rpm) showing a positive blast with the different genotypes identified is reported for each sample, for the two main ORF allowing the typing. The stars indicate samples in which a complete genome was identified. Green case: no rpm, yellow case <20,000rpm, red case: >20,000rpm.

be destroyed during sample preparation and the different purification steps. From a public health perspective, obtaining long contigs allowing viral genomes identification and investigation of transmission chains constitute a gold standard but remains a challenge when dealing with complex environmental samples such as wastewater (Aarestrup et al., 2021; Cobbin et al., 2021; Mazur et al., 2022). In a recent comment, Diamond et al. (2022) highlight the need to improve sewage sample analysis to be able to distinguish signal from noise to rapidly identify emerging Human pathogen (Diamond et al., 2022).

To achieve our goal to obtain long contigs allowing characterization of viral genomes, we used a previously described method (Bisseux et al., 2018; Strubbia et al., 2019). Combining a classic approach such as PEG precipitation, and additional purification steps including a special care given to protect viral RNA up to the guanidinium thiocyanate solution treatment contribute to successfully identify viral sequences.

Another issue when analyzing samples such as sewage is the representativeness of the sampling. Collecting 24-h composite sample increase the probability to detect a large diversity of strains compare to a grab sample, but the issue of low concentration of some viruses compared to other micro-organisms remains (Corpuz et al., 2020; Ahmed et al., 2022). One option can be to concentrate a large volume of wastewater, but this increases the risk to concentrate inhibitors of enzymes used during the library preparation (McCall et al., 2020; Adriaenssens et al., 2021). Instead, we chose to perform biological replicates, that is to repeat several extractions from the same sample and on different series (performed on different days) to limit any bias between samples. To evaluate the impact of the multiple purification steps in our method, and to verify that replicates provide comparable results, a control step in the form of a norovirus-specific qPCR was performed on the cDNA to verify that this viral target was not lost during the sample preparation up to RT. Overall, all replicates but one provided comparable results for norovirus GII to initial analysis, while for norovirus GI, more variations were observed presumably linked to the lower concentration of this genogroup as usually observed in sewage (Miura et al., 2018; Fumian et al., 2019). Importantly, this verification showed concentrations of norovirus (GI+GII) similar or somewhat higher than the initial one-step RT-qPCR screening, suggesting an efficient recovery of viral particles despite the multiple steps of the protocol, and a better elimination of inhibitors, compared to the initial screening approach. This norovirus quantification was used as a proxy to estimate the efficacy of the process to obtain viral genomes, assuming that other human enteric viruses behaved similarly. Adding some controls at different steps of sample preparation seems to be an important issue for further developments and subsequent risk assessment (Diamond et al., 2022; Fu et al., 2022).

In addition to the care given to sample preparation, we used a target enrichment step based on probe capture, whose added-value in yielding sequences of interest has been demonstrated in clinical, sewage or shellfish samples (Briese et al., 2015; Strubbia et al., 2019; Martinez-Puchol et al., 2020; Bonny et al., 2021). This enrichment step during the library preparation was shown efficient to identify novel viral sequence even in complex samples, however the relative abundance of the different sequences may be modified (Briese et al., 2015). Also, as we expected to recover a low proportion of viral reads, we chose a high sequencing depth by performing a Next-seq

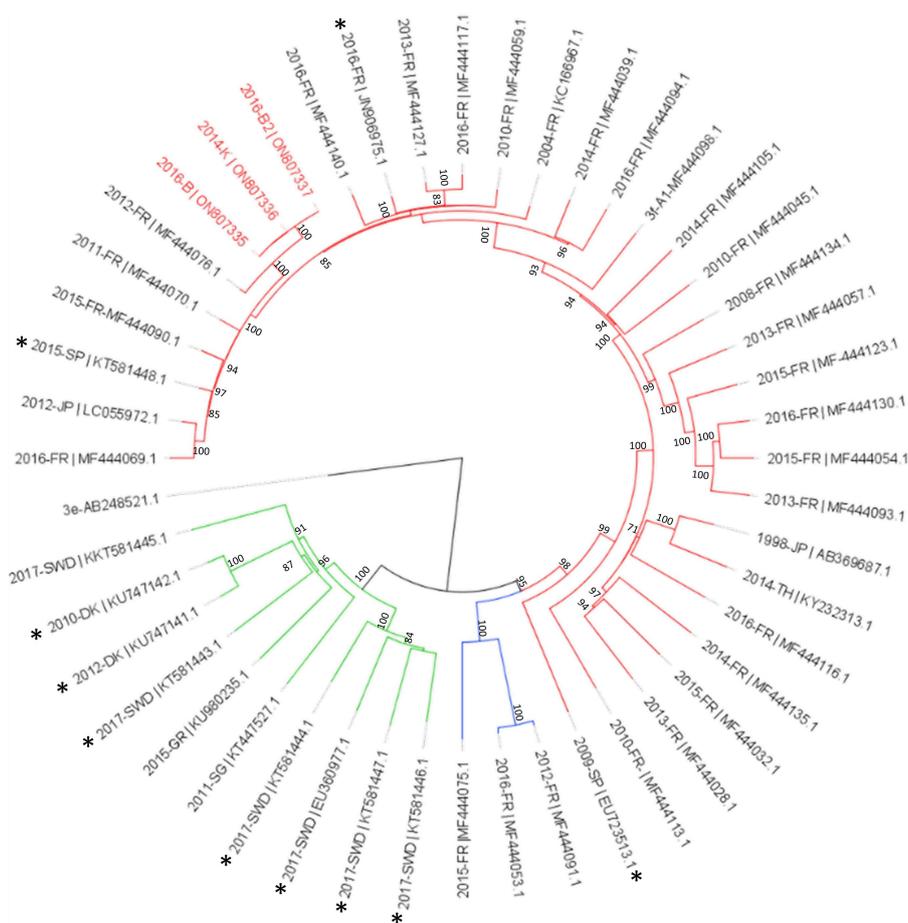


FIGURE 7 Phylogenetic analysis of HEV-3 subtype f subclusters using full genomes sequences identified in this study. The three complete genomes detected in this study (written in red) were aligned with reference sequences of subtype HEV-3f-A1 (red lines), HEV-3f-A2 (blue lines) and HEV-3f-B (green lines; Munoz-Chimeno et al., 2022). Sequences of animal origin are identified with an asterisk. The tree was constructed using IQtree with the GTR+F+I+G4 substitution model with ultrafast bootstrap (bootstrap=1,000) analysis, bootstrap values greater than 70% are shown. Rooted tree was drawn using FigTree.

TABLE 4 Heatmap profile showing the relative abundance for enterovirus sequences detected in the 12 samples.

Sample	Enterovirus A		Enterovirus B			Enterovirus C		Enterovirus J
	Untyped	A114	Untyped	E-9	CV-A9	Untyped	CV-A22	Untyped
2006-D	0	0	0	0	0	0	0	0
2006-G	0	0	0	0	0	1,085	0	526
2006-M	0	0	0	0	0	0	0	0
2006-G2	0	0	0	0	8	0	0	0
2007-P	0	0	0	0	0	0	0	0
2009-P	0	0	0	0	0	0	0	0
2012-K	0	0	11,124	0	0	0	0	0
2014-K2	0	0	211	0	0	0	0	0
2014-K1	0	0	0	0	0	0	0	0
2014-K	0	0	9,388	0	0	1,693	653	557
2016-B	943	1,436	174	0	43,685	53	0	0
2016-B2	2,912	1,291	4,528	528	0	102	0	0

The number of reads per million (rpm) showing a positive blast with the different strains identified is reported for each sample. Green case: no rpm, yellow case <20,000 rpm, red case: >20,000 rpm.

TABLE 5 Genotype constellation of human RVA sequences identified.

Sample	VP7 (1062)	VP4 (2362)	VP6 (1356)	VP1 (3302)	VP2 (2687)	VP3 (2592)	NSP1 (1581)	NSP2 (1059)	NSP3 (1074)	NSP4 (751)	NSP5 (666)
2006-D	G3	P[8]	I1	R1	C1	M1		N1	T1	E1	H1
2006-G	G1	P[8]	I1	R1	C1	M1	A1	N1	T1		H1
	G3										
2006-M	G1	P[8]	I1	R1	C1	M1			T1		
	G3										
2006-G2	G1	P[8]	I1	R1	C1	M1			T1		H1
	G9										
	G3										
2007-P				R1	C1	M1	A1		T1		
2009-P	G1	P[8]	I1	R1	C1	M1		N1	T1		
	G9								T2		
2012-K	G1	P[8]	I1	R1	C1	M1	A1	N1	T1		
	G9			R2					T2		
	G3										
2014-K2	G1	P[8]	I1	R1	C1	M1	A1	N1	T1	E1	
	G3										
2014-K1	G1	P[8]	I1	R1	C1	M1	A1	N1	T1	E1	H1
	G9										
2014-K	G1	P[8]	I1	R1	C1	M1	A1	N1	T1	E1	
	G3										
2016-B	G1	P[8]	I1	R1	C1	M1	A1	N1	T1	E1	H1
	G9			R2	C2	M2		N2	T2		
	G3						A3		T6		
2016-B2	G1	P[8]	I1	R1	C1	M1	A1	N1	T1		H1
	G9	P[4]	I2	R2	C2				T2		
	G3	P[1]			C3				T3		
% identity	92.8–98.9	94.5–98.9	97.7–98.9	97.2–98.2	88.6–98.6	91.2–99.1	97.6–98.6	98.1–98.5	96.2–98.7	98.46–98.80	

Sequences identified for the different genes are reported for each sample. The Wa-like genotype is shown in green, DS-1-like in red, AU-1-like in orange, and P [6] in purple, following genotype constellations (Matthijssens and Van Ranst, 2012), sequence identities are expressed in % for the lowest and highest homologies found. The number written after each segment is the expected length of the gene.

samples collected in this WWTP K two year later. This WWTP collects sewage from a small city close to the island (Figure 1), suggesting that some secondary cases occurred in the city, some probably being asymptomatic as observed during the epidemiological investigation of the initial outbreak (Guillois et al., 2016). In 2016, following the hepatitis A outbreak, some sewage samples collected from the small island were positive for hepatitis A virus by RT-PCR (unpublished data). Here, we were able to obtain a full genome sequence of a genotype IA, genotype circulating in France (Bisseux et al., 2018). These results confirmed the importance of metagenomic to identify viral sequences and contribution that can be made to epidemiological surveillance (Izopet et al., 2019; Diamond et al., 2022; Munoz-Chimeno et al., 2022). This also reinforce the need to associate environmental surveillance with epidemiological data.

As mentioned above despite the care given to purify viral genomes and to increase the number of viral reads the issue of contig identification remains critical for segmented genomes. Database such as NCBI Genbank or more dedicated tool like the norovirus typing

tool or the Rotac tool (for rotavirus) are really helpful. However, assembling the 11 rotavirus genes from a sewage sample, remain a delicate challenge, performed on previously described constellations obtained in Human or animal stool samples (Simsek et al., 2021; Portal et al., 2022). It is interesting to note that some genes were more frequently identified in our samples than other without any relationship to the length. This issue may need to be further investigated before wastewater samples can be used to evaluate the genetic diversity of the different segments and any changes in the patterns of the genotype distribution (Matthijssens and Van Ranst, 2012). We also faced this issue for the analysis of the bi-segmented double-stranded RNA genomes of the picobirnavirus. Many complete segments could not be clearly identified in GenBank, but the high number of contigs detected in this study suggest that picobirnaviruses may be under-detected in the human population, as hypothesized also by others (Ghosh and Malik, 2021).

The aim of this study was to evaluate the relevance of technical replicates and quality criteria such as a minimum of reads or contigs

length to identify human pathogenic virus sequence with confidence. We chose samples from diverse WWTP and different seasons to ensure that our protocols could be applied to diverse wastewater samples. As some samples were kept frozen for more than 10 years and could be degraded, we did not expect to raise conclusion in terms of sequence diversity or the presence of specific strains. Yet, we included samples linked to hepatitis outbreaks to assess the possibility to detect the causing agents retrospectively. Our approach was found efficient to identify some sequences and successfully described the viral diversity, yielding full genomes either for norovirus, enterovirus or even rotavirus, even if, for these segmented genomes, combining genes remains a difficult issue.

Metagenomics allowing the description of all genomes present in a sample applied to wastewater samples that received all different viruses replicating in the gut of the local population constitute a promising approach for epidemiological surveillance. Our results show that performing biological replicates and merging the obtained data allow to obtain longer contigs and a reliable strain identification, an important issue for public health surveillance. Indeed, confidence in result interpretation is strengthened by quality criteria in terms of read numbers and contig length (Diamond et al., 2022). Importantly, this wastewater-based approach yielded viral sequences, sometimes whole genomes, that could be phylogenetically related to those obtained during related outbreak investigations or public databases. With the improvements of methods and the enrichment of databases, wastewater sample analysis will become an important tool also to prevent further transmission by raising alerts in case of viral outbreaks or emergence.

Data availability statement

The sequencing data presented in this study are deposited in the NCBI repository. Accession numbers can be found in [Supplementary Table 1](#).

Author contributions

JS, MD, and FG contributed to the conception, design of the study, contributed to data interpretation, and manuscript preparation. JS performed experiments. JS and AB performed the bio-informatics

References

- Aarestrup, F. M., Bonten, M., and Koopmans, M. P. G. (2021). Pandemics-one health preparedness for the next. *Lancet Reg. Health Eur.* 9:100210. doi: 10.1016/j.lanepe.2021.100210
- Adriaenssens, E. M., Farkas, K., McDonald, J. E., Jones, D. L., Allison, H. E., and McCarthy, A. J. (2021). Tracing the fate of wastewater viruses reveals catchment-scale virome diversity and connectivity. *Water Res.* 203:117568. doi: 10.1016/j.watres.2021.117568
- Ahmed, W., Simpson, S. L., Bertsch, P. M., Bibby, K., Bivins, A., Blackall, L. L., et al. (2022). Minimizing errors in RT-PCR detection and quantification of SARS-CoV-2 RNA for wastewater surveillance. *Sci. Total Environ.* 805:149877. doi: 10.1016/j.scitotenv.2021.149877
- Barbé, L., Schaeffer, J., Besnard, A., Jousse, S., Wurtzer, S., Moulin, L., et al. (2022). SARS-CoV-2 whole genome sequencing using oxford nanopore technology for variant monitoring in wastewaters. *Front. Microbiol.* 13:889811. doi: 10.3389/fmicb.2022.889811
- Barril, P. A., Fumian, T. M., Prez, V. E., Gil, P. I., Martinez, L. C., Giordano, M. O., et al. (2015). Rotavirus seasonality in urban sewage from Argentina: effect of meteorological

analysis. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the COMPARE and VEO European projects (H2020 grant agreement N°643476 and H2020 SC1-2019-874735) and by support from the Direction Générale de l'Alimentation (DGAl) through the annual convention.

Acknowledgments

This work benefited from equipment and services from iGenSeq core facility, at ICM (Institut Cerveau Moelle, Paris, France). We thank B. Oude Munnink and D. Nieuwenhuijse for helpful scientific discussion. We thank Obepine consortium for helpful discussion. We are grateful to C. McLeods, Cawthron Institute (NZ) for careful review of the manuscript.

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmicb.2023.1161674/full#supplementary-material>

variables on the viral load and the genetic diversity. *Environ. Res.* 138, 409–415. doi: 10.1016/j.envres.2015.03.004

Bisseux, M., Colombet, J., Mirand, A., Roque-Alfonso, A.-M., Abranavel, F., Izopet, J., et al. (2018). Monitoring human enteric viruses in wastewater and relevance to infections encountered in the clinical setting: a one-year experiment in Central France, 2014 to 2015. *Euro Surveill.* 23, 17–00237. doi: 10.2807/1560-7917.ES.2018.23.7.17-00237

Bonny, P., Schaeffer, J., Besnard, A., Desdouts, E., Essia Ngang, J. J., and Le Guyader, F. S. (2021). Human and animal RNA virus diversity detected by metagenomics in Cameroonian clams. *Front. Microbiol.* 12:e770385. doi: 10.3389/fmicb.2021.770385

Briese, T., Kapoor, A., Mishra, N., Jain, K., Kumar, A., Jabado, O. J., et al. (2015). Virome capture sequencing enables sensitive viral diagnosis and comprehensive virome analysis. *MBio* 6, e01491–e01415. doi: 10.1128/mBio.01491-15

Cobbin, J. C. A., Charon, J., Harvey, E., Holmes, E. C., and Mahar, J. E. (2021). Current challenges to virus discovery by metatranscriptomics. *Cur. Op. Virol.* 51, 48–55. doi: 10.1016/j.coviro.2021.09.007

- Corpuz, M. V. A., Buonerba, A., Vigliotta, G., Zarra, T., Ballesteros, F., Campiglia, P., et al. (2020). Viruses in wastewater: occurrence, abundance and detection methods. *Sci. Total Environ.* 745:140910. doi: 10.1016/j.scitotenv.2020.140910
- da Silva, A., Le Saux, J.-C., Parnaudeau, S., Pommepuy, M., Elimelech, M., and Le Guyader, F. S. (2007). Evaluation of removal of noroviruses during wastewater treatment, using real-time reverse transcription-PCR: different behaviors of genogroups I and II. *Appl. Environ. Microbiol.* 73, 7891–7897. doi: 10.1128/aem.01428-07
- Deshpande, J. M., Sharma, D. K., Saxena, V. K., Shetty, S. A., Qureshi, T. H. I. H., and Nalavade, U. P. (2016). Genomic characterization of two new enterovirus types, EV-A114 and EV-A121. *J. Med. Microbiol.* 65, 1465–1471. doi: 10.1099/jmm.0.000380
- Diamond, M. B., Keshaviah, A., Bento, A. I., Conroy-Ben, O., Driver, E. M., Ensor, K. B., et al. (2022). Wastewater surveillance of pathogens can inform public health responses. *Nat. Med.* 28, 1992–1995. doi: 10.1038/s41591-022-01940-x
- Fu, S., Yang, Q., Sheng, Y., Wang, Q., Wu, J., Qiu, Z., et al. (2022). Metagenomics combined with comprehensive validation as a public health risk assessment tool for urban and agricultural run-off. *Water Res.* 209:117941. doi: 10.1016/j.watres.2021.117941
- Fumian, T. M., Fioretti, J. M., Lun, J. H., dos Santos, I. A. L., White, P. A., and Miagostovich, M. P. (2019). Detection of norovirus epidemic genotypes in raw sewage using next generation sequencing. *Environ. Int.* 123, 282–291. doi: 10.1016/j.envint.2018.11.054
- Ghosh, S., and Malik, Y. S. (2021). The true host/s of picobirnaviruses. *Front. Vet. Sci.* 7:615293. doi: 10.3389/fvets.2020.615293. Greninger, A. L.
- Greninger, A. L. (2018). A decade of RNA virus metagenomics is (not) enough. *Virus Res.* 244, 218–229. doi: 10.1016/j.virusres.2017.10.014
- Guajardo-Leiva, S., Chnaiderman, J., Gaggero, A., and Diez, B. (2020). Metagenomic insights into the sewage RNA virosphere of a large city. *Viruses* 12:1050. doi: 10.3390/v12091050
- Guerrero-Latorre, L., Romero, B., Bonifaz, E., Timoneda, N., Rusinol, M., Girones, R., et al. (2018). Quito's virome: metagenomic analysis of viral diversity in urban streams of Ecuador's capital city. *Sci. Total Environ.* 645, 1334–1343. doi: 10.1016/j.scitotenv.2018.07.213
- Guillois, Y., Abranavel, F., Miura, T., Pavio, N., Vaillant, V., Lhomme, S., et al. (2016). High proportion of asymptomatic infections in an outbreak of hepatitis E associated with a spit-roasted piglet, France, 2013. *Clin. Infect. Dis.* 62, 351–357. doi: 10.1093/cid/civ862
- Hata, A., Kitajima, M., Haramoto, E., Lee, S., Ihara, M., Gerba, C., et al. (2018). Next-generation amplicon sequencing identifies genetically diverse human astroviruses, including recombinant strains, in environmental waters. *Nat. Sci. Rep.* 8:11837. doi: 10.1038/s41598-018-30217-y
- Hendriksen, R. S., Lukjancenko, O., Munk, P., Hjelmso, M. H., Verani, J. R., Ng'eno, E., et al. (2019). Pathogen surveillance in the informal settlement, Kibera, Kenya, using a metagenomics approach. *PLoS One* 14:e0222531. doi: 10.1371/journal.pone.0222531
- Izopet, J., Tremieux, P., Marion, O., Miguersn, M., Capelli, N., Chapuy-Regaud, S., et al. (2019). Hepatitis E virus infections in Europe. *J. Clin. Virol.* 120, 20–26. doi: 10.1016/j.jcv.2019.09.004
- Kageyama, T., Kojima, S., Shinohara, M., Uchida, K., Fukushi, S., Hoshino, F. B., et al. (2003). Broadly reactive and highly sensitive assay for Norwalk-like viruses based on real-time quantitative reverse transcription-PCR. *J. Clin. Microbiol.* 41, 1548–1557. doi: 10.1128/JCM.41.4.1548-1557.2003
- Kim, Y., Van Bonn, W., Aw, T. G., and Rose, J. B. (2017). Aquarium viromes: viromes of human-managed aquatic systems. *Front. Microbiol.* 8:1231. doi: 10.3389/fmicb.2017.01231
- Kroneman, A., Vennema, H., Deforche, K., Avoort, H. V. D., Penaranda, S., Oberste, M. S., et al. (2011). An automated genotyping tool for enteroviruses and noroviruses. *J. Clin. Virol.* 51, 121–125. doi: 10.1016/j.jcv.2011.03.006
- Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2, 231–239. doi: 10.1016/0888-7543(88)90007-9
- Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/NMETH.1923
- Loisy, F., Atmar, R. L., Guillon, P., Le Cann, P., Pommepuy, M., and Le Guyader, F. S. (2005). Real-time RT-PCR for norovirus screening in shellfish. *J. Virol. Methods* 123, 1–7. doi: 10.1016/j.jviromet.2004.08.023
- Mabasa, V. V., van Zyl, W. B., Ismail, A., Allam, M., Taylor, M. B., and Mans, J. (2022). Multiple novel human norovirus recombinants identified in wastewater in Pretoria, South Africa by next generation sequencing. *Viruses* 14:2732. doi: 10.3390/v14122732
- Maes, P., Matthijnsens, J., Rahman, M., and Van Ranst, M. (2009). RotaC: a web-based tool for the complete genome classification of group A rotaviruses. *BMC Microbiol.* 9:238. doi: 10.1186/1471-2180-9-238
- Martinez-Puchol, S., Rusinol, M., Fernandez-Cassi, X., Timoneda, N., Itarte, M., Andres, C., et al. (2020). Characterisation of the sewage virome: comparison of NGS tools and occurrence of significant pathogens. *Sci. Total Environ.* 713:136604. doi: 10.1016/j.scitotenv.2020.136604
- Matthijnsens, J., and Van Ranst, L. M. (2012). Genotype constellation and evolution of group A rotaviruses infecting humans. *Cur. Op. Virol.* 2, 426–433. doi: 10.1016/j.coviro.2012.04.007
- Mazur, F. G., Morinisi, L. M., Martins, J. O., Pontes Bueno Guerra, P., and Freire, C. C. M. (2022). Exploring virome diversity in public data in South America as an approach for detecting viral sources from potentially emerging viruses. *Front. Genet.* 12:722857. doi: 10.3389/fgene.2021.722857
- McCall, C., Wu, H., Miyani, B., and Xagorarakis, I. (2020). Identification of multiple potential viral diseases in a large urban center using wastewater surveillance. *Water Res.* 184:116160. doi: 10.1016/j.watres.2020.116160
- Melnick, J. (1947). Poliomyelitis virus in urban sewage in epidemic and non-epidemic times. *Am. J. Hyg.* 45, 240–253.
- Metcalfe, T., Melnick, J. L., and Estes, M. K. (1995). Environmental microbiology: from detection of virus in sewage and water by isolation to identification by molecular biology—a trip of over 50 years. *An. Rev. Microbiol.* 49, 461–487. doi: 10.1146/annurev.mi.49.100195.002333
- Miura, T., Lhomme, S., Le Saux, J.-C., Le Mehaute, P., Guillois, Y., Couturier, E., et al. (2016). Detection of hepatitis E virus in sewage after an outbreak on a French island. *Food Environ. Virol.* 8, 194–199. doi: 10.1007/s12560-016-9241-9
- Miura, T., Schaeffer, J., Le Saux, J.-C., Le Mehaute, P., and Le Guyader, F. S. (2018). Virus type-specific removal in a full-scale membrane bioreactor treatment process. *Food Environ. Virol.* 10, 176–186. doi: 10.1007/s12560-017-9330-4
- Munoz-Chimeno, M., Barturen, S., Garcia-Lugo, M. A., Morago, L., Rodriguez, A., Galan, J. C., et al. (2022). Hepatitis E virus genotype 3 microbiological surveillance by the Spanish reference laboratory: geographic distribution and phylogenetic analysis of subtypes from 2009 to 2019. *Euro Surveill.* 27:2100542. doi: 10.2807/1560-7917.ES.2022.27.23.2100542
- Nieuwenhuijse, D., and Koopmans, M. P. G. (2017). Metagenomic sequencing for surveillance of food- and waterborne viral diseases. *Front. Microbiol.* 8:230. doi: 10.3389/fmicb.2017.00230
- Nieuwenhuijse, D. F., Munnink, B. O., Phan, M. V. T., the Global Sewage Surveillance project consortium Munk, P., Venkatakrishnan, S., et al. (2020). Setting a baseline for global urban virome surveillance in sewage. *Sci. Rep.* 10:13748. doi: 10.1038/s41598-020-69869-0
- Nurk, S., Meleshko, D., Korobeynikov, A., and Tevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi: 10.1038/s41598-020-69869-0
- Portal, T. M., Vanmechelen, B., Espen, L. V., Jansen, D., Teixeira, D. M., Alho de Sousa, E. S., et al. (2022). Molecular characterization of the gastrointestinal eukaryotic virome in elderly people in Belem, Para, Brazil. *Inf. Gen. Evol.* 99:105241. doi: 10.1016/j.meegid.2022.105241
- Rodriguez-Diaz, J., Querales, L., Caraballo, L., Vizzi, E., Liprandi, F., Takiff, H., et al. (2009). Detection and characterization of waterborne gastroenteritis viruses in urban sewage and sewage polluted river waters in Caracas, Venezuela. *Appl. Environ. Microbiol.* 75, 387–394. doi: 10.1128/AEM.02045-08
- Savini, F., Giacometti, F., Tomasello, F., Pollesel, M., Piva, S., Serraino, A., et al. (2021). Assessment of the impact on human health of the presence of norovirus in bivalve mollusc: what data do we miss? *Foods* 10:2444. doi: 10.3390/foods10102444
- Simá, L. C., Schaeffer, J., Le Saux, J.-C., Parnaudeau, S., Elimelech, M., and Le Guyader, F. S. (2011). Calicivirus removal in a membrane bioreactor wastewater treatment plant. *Appl. Environ. Microbiol.* 77, 5170–5177. doi: 10.1128/AEM.00583-11
- Simsek, C., Corman, V. M., Everling, H. U., Lukashev, A., Rasche, A., Maganga, G. D., et al. (2021). At least seven distinct rotavirus genotype constellations in bats with evidence of reassortment and zoonotic transmissions. *MBio* 12:e02755. doi: 10.1128/mBio.02755-20
- Strubbia, S., Schaeffer, J., Besnard, A., Wacrenier, C., Le Menec, C., Garry, P., et al. (2020). Metagenomic to evaluate norovirus genomic diversity in oysters: impact on hexamer selection and targeted capture-based enrichment. *Int. J. Food Microbiol.* 323:108588. doi: 10.1016/j.ijfoodmicro.2020.108588
- Strubbia, S., Schaeffer, J., Oude Munnink, B. B., Besnard, A., Phan, M. V. T., Nieuwenhuijse, D. F., et al. (2019). Metavirome sequencing to evaluate norovirus diversity in sewage and related bioaccumulated oysters. *Front. Microbiol.* 10:3389. doi: 10.3389/fmicb.2019.02394
- Suffredini, E., Iaconelli, M., Equestre, M., Valdazo-Gonzalez, B., Ciccaglione, A. R., Marcantonio, C., et al. (2018). Genetic diversity among genogroup II noroviruses and progressive emergence of GII.17 in wastewaters in Italy (2011–2016) revealed by next-generation and sanger sequencing. *Food Environ. Virol.* 10, 141–150. doi: 10.1007/s12560-017-9328-y
- Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Coordinators, T. O., Eveillard, D., et al. (2020). Tara oceans: towards global ocean ecosystems biology. *Nat. Rev.* 18, 428–445. doi: 10.1038/s41579-020-0364-5
- Svraka, S., Duizer, E., Vennema, H., de Bruin, E., van der Veer, B., Dorresteyn, B., et al. (2007). Etiological role of viruses in outbreaks of acute gastroenteritis in the Netherlands from 1994 through 2005. *J. Clin. Microbiol.* 45, 1389–1394. doi: 10.1128/JCM.02305-06
- Torok, V. A., Hodgson, K. R., Jolley, J., Turnbull, A., and McLeod, C. (2019). Estimating risk associated with human norovirus and hepatitis A virus in fresh Australian leafy greens and berries at retail. *Int. J. Food Microbiol.* 309:108327. doi: 10.1016/j.ijfoodmicro.2019.108327

Vilsker, M., Moosa, Y., Nooij, S., Fonseca, V., Ghysens, Y., Dumon, K., et al. (2019). Genome detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics* 35, 871–873. doi: 10.1093/bioinformatics/bty695

Wurtzer, S., Marechal, V., Mouchel, J. M., Maday, Y., Teyssou, R., Richard, E., et al. (2020). Evaluation of lockdown effect on SARS-CoV-2 dynamics through viral genome quantification in wastewater, greater Paris, France, 5 March to 23 April 2020. *Euro Surveill.* 25:2000776. doi: 10.2807/1560-7917

Yang, Q., Rivailier, P., Zhu, S., Yan, D., Xie, N., Tang, H., et al. (2021). Detection of multiple viruses potentially infecting humans in sewage water from Xinjiang Uygur autonomous region, China. *Sci. Total Environ.* 754:142322. doi: 10.1016/j.scitotenv.2020.142322

Zheng, H., Lu, J., Zhang, Y., Yoshida, H., Guo, X., Liu, L., et al. (2013). Prevalence of nonpolio enteroviruses in the sewage of Guangzhou city, China, from 2009 to 2012. *Appl. Environ. Microbiol.* 79, 7679–7683. doi: 10.1128/AEM.02058-13