



Comparison of Three Viral Nucleic Acid Pre-amplification Pipelines for Sewage Viral Metagenomics

Xavier Fernandez-Cassi^{1,2} · Tamar Kohn¹

Received: 19 October 2023 / Accepted: 1 March 2024
© The Author(s) 2024

Abstract

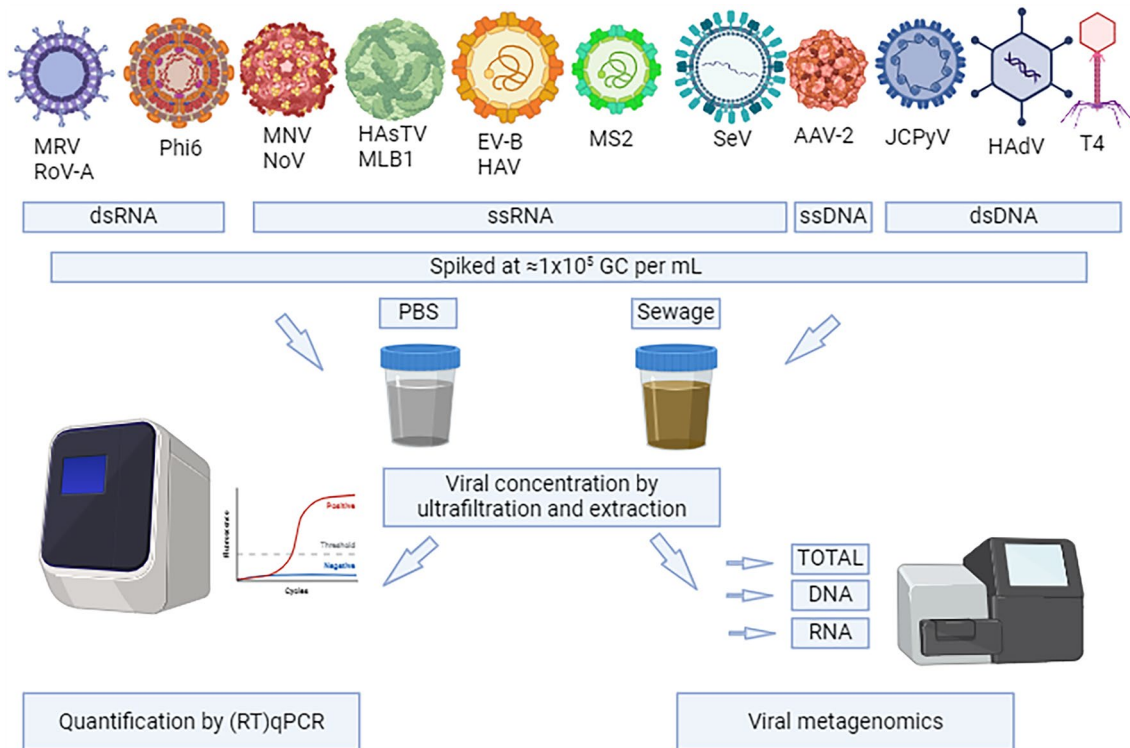
Viral metagenomics is a useful tool for detecting multiple human viruses in urban sewage. However, more refined protocols are required for its effective use in disease surveillance. In this study, we investigated the performance of three different pre-amplification pipelines (specific to RNA viruses, DNA viruses or both) for viral genome sequencing using spiked-in Phosphate Buffered Saline and sewage samples containing known concentrations of viruses. We found that compared to the pipeline targeting all genome types, the RNA pipeline performed better in detecting RNA viruses in both spiked and unspiked sewage samples, allowing the detection of various mammalian viruses including members from the *Reoviridae*, *Picornaviridae*, *Astroviridae* and *Caliciviridae*. However, the DNA-specific pipeline did not improve the detection of mammalian DNA viruses. We also measured viral recovery by quantitative reverse transcription polymerase chain reaction and assessed the impact of genetic background (non-viral genetic material) on viral coverage. Our results indicate that viral recoveries were generally lower in sewage (average of 11.0%) and higher in Phosphate Buffered Saline (average of 23.4%) for most viruses. Additionally, spiked-in viruses showed lower genome coverage in sewage, demonstrating the negative effect of genetic background on sequencing. Finally, correlation analysis revealed a relationship between virus concentration and genome normalized reads per million, indicating that viral metagenomic sequencing can be semiquantitative.

✉ Xavier Fernandez-Cassi
x.fernandez-cassi@ub.edu

¹ Laboratory of Environmental Chemistry, School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Vaud, Lausanne, Switzerland

² Department of Biology, Healthcare and Environment, Faculty of Pharmacy and Food Sciences, University of Barcelona (UB), Barcelona, Catalunya, Spain

Graphical Abstract



Keywords Viral metagenomics · Sewage monitoring · Environmental surveillance · Enteric viruses · Waterborne pathogens

Introduction

Sewage contains a broad variety of viruses (Bibby & Pecchia, 2013; Cantalupo et al., 2011; Ng et al., 2012) and has been suggested as a matrix for the surveillance of viruses circulating among a population (Fernandez-Cassi et al., 2018). Until recently, most of the commonly studied human viruses in sewage are enteric viruses with a fecal–oral transmission route such as noroviruses, enteroviruses or rotaviruses (Hellmér et al., 2014; Santiso-Bellón et al., 2020). Since the onset of the COVID-19 pandemic, however, sewage surveillance has also proven to be a useful tool to track the dynamics of respiratory viruses such as SARS-CoV-2 (Fernandez-Cassi et al., 2021; Randazzo et al., 2020), Respiratory Syncytial Virus (RSV) (Hughes et al., 2022), influenza (Heijnen & Medema, 2011) or monkeypox virus (de Jonge et al., 2022; Wolfe et al., 2023). Sewage has proven to be a useful tool to track the occurrence of variants (Jahn et al., 2021) and to derive epidemiological parameters of interest such as the effective reproductive number for SARS-CoV-2 as (Huisman et al., 2021). Accordingly, the interest in using sewage as a tool to monitor viral diseases has dramatically increased, and

surely the use of this matrix will be further expanded in the near future to monitor other viral diseases previously unexplored.

Traditionally, viruses in sewage have been studied using quantitative PCR (qPCR) for specific viral pathogens or viral indicators (Hellmér et al., 2014; Kitajima et al., 2014; Santiso-Bellón et al., 2020). In the last decade due to the democratization of next-generation sequencing, virus surveillance has been done using amplicon sequencing methods to study the diversity within a specific viral family (Bisseux et al., 2020; Fernandez-Cassi et al., 2018; Iaconelli et al., 2017), as well as using non-targeted metaviromic studies to explore the viral diversity without aiming at any particular viruses (Cantalupo et al., 2011; Fernandez-Cassi et al., 2018; McCall et al., 2020; Ng et al., 2012). Both methodologies provide qualitative information (presence/absence) on pathogens but can also infer important genomic information pertaining to viral species or variants. However, amplicon sequencing approaches have the limitation that they are primer-dependent and hence can only be applied to known and expected viral targets. In contrast, metaviromic approaches are primer-independent, and can thus be used to detect a broader range of viruses, including new or unexpected species.

Despite its potential as a catch-all approach, there are several challenges to the application of sewage viral metagenomics as a routine method for viral surveillance. First, mammalian viruses are present in sewage at concentrations that are too low for direct sequencing. The sewage sample must first be concentrated and extracted, and nucleic acids must be preamplified in order to meet the requirements for library preparation. Both procedures can be a source of bias (Nieuwenhuijse et al., 2020; Parras-Moltó et al., 2018). Second, mammalian viruses in sewage are vastly outnumbered by both phages and phytoviruses (Bačnik et al., 2020; Cantalupo et al., 2011; Duarte et al., 2023), which are enriched along with mammalian viruses during sample concentration. In addition, sewage also contains a high background of nucleic acids from non-viral organisms. Nucleic acids from phages, phytoviruses and non-viral organisms are co-amplified and sequenced with the viral targets, thereby reducing the sequencing depth and the coverage of the viral species of interest (Krishnamurthy & Wang, 2017; Tamaki et al., 2012). Third, mammalian viruses present in sewage can contain different genome conformations (single- and double-stranded, linear and circular genomes), and their genomes can be codified using DNA or RNA molecules. This implies the need of molecular biology methodologies that work efficiently with both DNA and RNA molecules. And finally, methods used to concentrate viral particles from sewage tend to co-concentrate inhibitory substances that interfere with subsequent molecular biology reactions (Fernandez-Cassi et al., 2018; Schrader et al., 2012).

For viral metagenomics to serve as a tool to study the presence of multiple human viruses in a single analysis, refined protocols are thus needed to increase both the method sensitivity and its ability to specifically capture viruses. Several studies have been developing strategies to increase this sensitivity by applying specific viral probes to capture mammalian viruses after library preparation (Martínez-Puchol et al., 2020; Tisza et al., 2023). These approximations aim to enhance the library preparation for vertebrate viruses by excluding the sequencing of bacteriophages and other sources of DNA, thereby increasing the number of reads and the coverage of the targeted viruses of interest.

In recent years, significant progress has been made in developing systematic protocols and techniques for implementing viral metagenomics in a clinical setting (Conceição-Neto et al., 2018; Hall et al., 2014; Kohl et al., 2015; Li et al., 2015; Tulloch et al., 2023). Although there has been a growing interest in using sewage as a sample to monitor circulating viruses within a population, especially since the emergence of COVID-19, there is still a lack of specific protocols to assess viral diversity through viral metagenomics in this specific context.

In this study, we assess how viral metagenomic sequencing is affected by virus recovery, background genetic material, and preamplification protocols. We conducted metagenomic analyses in two scenarios: a low genetic background scenario using phosphate buffered saline (PBS) and a high genetic background scenario using a sewage sample. Both matrices were spiked with a suite of diverse viruses at known concentrations to evaluate viral particle recovery. Finally, three preamplification protocols were applied. One protocol targeting both DNA and RNA viruses, one targeting only DNA viruses, and one targeting only RNA viruses. The idea behind this comparison is that individual pipelines for DNA and RNA viruses should improve the sensitivity compared to protocols which simultaneously captures both genome types. Finally, the results were validated in unspiked sewage samples.

Materials and Methods

Metaviromic analyses were conducted in spiked PBS and sewage samples, using different preamplification pipelines. The different the steps in workflow, from sample processing to bioinformatic analysis, are summarized in Fig. S1.

Preparation of Virus Spiking Solutions

For spike-in experiments, individual virus spiking solutions were prepared for 14 viruses with diverse characteristics. A list of all viruses used, their taxonomic classification and their genome and capsid characteristics, as well as their hosts cells is given in Table S1. As species representatives for double-stranded DNA viruses (dsDNA), human mastadenovirus 2 (HAdV-C2, kindly donated by Rosina Girones, University of Barcelona and from now on referred to as HAdV), JC Polyomavirus (JCPyV MAD4 strain, ATCC VR-1583) and bacteriophage T4 (DSMZ 4505) were selected. As a representatives of single-stranded DNA viruses (ssDNA), adeno-associated virus 2 (AAV-2; kindly donated by Cornel Fraefel, University of Zurich) was used. Single-stranded RNA viruses (ssRNA) included MS2 bacteriophage (ATCC 15597-B1), enterovirus B (Echovirus 11 (E11) Gregory strain, ATCC® VR37TM, and from now on referred to as EV-B throughout the manuscript), Sendai virus (SeV; kindly provided by Dominique Garcin, University of Geneva) and human astrovirus MLB1 (HAsV MLB1; kindly donated by Susana Guix and Albert Bosch, University of Barcelona), human norovirus GII (NoVGII), hepatitis A virus (HAV) and murine norovirus type I (MNV) (all kindly donated by Anna Charlotte Schultz, Technical University of Denmark). For double-stranded RNA (dsRNA), mammalian orthoreovirus 1 (MRV-1 Lang Strain, ATCC® VR-230), rotavirus A

(RoV-A; ATCC® VR-2018™) and *Pseudomonas* bacteriophage phi 6 (ϕ 6, DSMZ 21518) were added.

Viruses were propagated on their respective bacterial host or mammalian cell line (Table S1). Bacteriophages were harvested after an overnight (o/n) incubation with the bacterial host, were centrifuged at $3000\times g$ for 10 min to remove bacterial debris and were filtered through low-binding protein 0.45 μ m and 0.22 μ m Sterivex filters (Millipore, Massachusetts). Filtered solutions were preserved at 4 °C until further use.

Most mammalian viruses were propagated in-house. Propagation was performed in 125 cm² cell culture flasks containing sub-confluent host cell monolayers using a Multiplicity of Infection (MOI) of 0.1 to 1 depending on the virus. After 1 h incubation at 37 °C with regular agitation, cells were washed with PBS (140 mM NaCl, 2.68 mM KCl, 10 mM Na₂HPO₄) (Cat. No. 18912014, thermofisher Scientific) to remove non-adhered viral particles. Exposed cells were incubated at 37 °C in a 5% CO₂ atmosphere and were supplemented with DMEM at 2% of bovine serum for 5 to 10 days. Then viral particles were released from infected cells by three freeze–thaw cycles. Cell debris was removed by centrifugation at $3000\times g$ for 15 min and the supernatant was filtered using low-binding protein 0.45 μ m Sterivex filters (Millipore, Massachusetts). To prevent potential damage to viral capsids, the filtered solutions were preserved at 4 °C for a duration of one month before their subsequent use.

Four viruses, norovirus GII, SeV, AAV-2 and HAstV MLB1 were provided as stock solutions and were not further propagated. Specifically, NoV GII was received as a 2% fecal suspension from a clinical sample. To minimize companion microbiota present in the sample and reduce its genetic background, the fecal suspension was centrifuged at $3000\times g$ for 15 min. The supernatant was filtered and stored as described above. SeV was received as a stock solution produced by propagation in embryonated eggs as described elsewhere (Strahle et al., 2006). AAV-2 was received as a stock solution produced in HEK293T using recombinant plasmids as described by (Samulski et al., 1987).

The virus concentration in each individual stock solution was determined by (RT)qPCR as described below. Stock solutions were then diluted with PBS to a virus particle concentration corresponding to approximately 1×10^5 GC/mL. These solutions served as the final spiking solutions.

Nuclease Treatment, Nucleic Acid Extraction and (RT)qPCR Analysis

Prior to nucleic acid (NA) extraction, viral stock solutions were nuclease treated to remove any free RNA or DNA. For each viral stock solution, 150 μ l aliquots were treated with 30 U of TurboDNase (Ambion Cat. No. AM2238), 12U of Benzoase nuclease (Sigma Aldrich—E1014-5KU) and 5U

of RNase A and 200U RNase T1 (Thermo Fisher Cat. No. AM2286) for 2 h at 37 °C. After nuclease treatment, NAs were extracted using the Qiagen Viral RNA Mini Kit (Cat. No. 22906, Qiagen, Valencia, CA, USA) without RNA carrier. NA were eluted using 60 μ l of elution buffer.

(RT)qPCR analyses were conducted on tenfold dilutions of NA extracts in order to minimize the impact of inhibitors present in sewage samples. The different assays used are detailed in Table S2. For dsRNA and ssRNA viruses, quantifications were performed using the RNA Ultrasense One-Step Quantitative RT-PCR System kit (Invitrogen Cat. No.: 11732–927) by adding 5 μ l of NA template. For DNA TaqMan assays, TaqMan Environmental PCR Master Mix kit (Applied Biosystems Cat. No. 4396838) was used by adding 10 μ l of NA as template. Finally, a DNA SYBR green assay with TB Green™ Advantage® qPCR Premix kit (Takara, Cat. No. 639676) was used for bacteriophage T4 by adding 2.5 μ l of NA as template. All (RT)qPCR reactions had a final volume of 25 μ l and were run on a Mic qPCR Cycler device (Bio Molecular Systems). All (RT)qPCR runs included water as non-template controls (NTC). A negative extraction control (NEC) to discard cross-contamination during sample processing was also included. Calibration curves were generated using gblock standards (IDT, Coralville, IA, USA). These standards consisted of a minimum of 5 known concentrations that spanned the expected concentration range in the sample, were incorporated for each assay. For a comprehensive list of the standard concentrations, please refer to the supplementary section.

The PCR efficiencies, calibration curve slope and intercept of each assay is reported in Table S2, along with a checklist of experimental details as requested by the minimum information for publication of quantitative real-time PCR experiments (MIQE) guidelines in Table S3 (Bustin et al., 2009).

Spiking of PBS and Sewage Sample and Concentration of Viral Particles

Spiking experiments were performed in a PBS and in raw sewage collected from the Vidy wastewater treatment plant (Lausanne, Switzerland). Sampling was approved by Le service de l'eau de la Ville de Lausanne, which manages the Vidy wastewater treatment plant. Aliquots of 105 mL were spiked with 1 mL of each of the virus spiking solutions to yield a final volume of 120 mL. We aimed for a spiked concentration of viral particles corresponding to 1×10^5 GC per mL for each virus. This concentration was chosen as an acceptable starting point to detect spiked viruses by (RT)qPCR, assuming a LOD close to 50 GC per PCR reaction. The theoretical concentrations of spiked viruses are shown in Table 1. Spiked solutions were stirred for 1 h at room temperature. Then, 100 mL of the samples were mixed with

Table 1 Total amount of spiked viruses as genomic copies (GC) in low background sample (PBS) and high background sample (sewage)

	HAV	NoVGII	HAdV	EV-B	MNV	AAV-2	JCPyV	HAdV MLB1	MS2	RoV-A	SeV	Phi6	T4
Genome conformation	ssRNA +	ssRNA +	dsDNA	ssRNA +	ssRNA +	ssDNA	dsDNA	ssRNA +	ssRNA +	dsRNA	ssRNA-	dsRNA	dsDNA
Theoretical amount spiked (100 mL)	2.35×10^8	7.83×10^6	7.16×10^6	9.68×10^7	1.67×10^7	2.57×10^7	7.57×10^7	2.43×10^7	4.55×10^7	3.20×10^7	1.93×10^7	3.29×10^7	5.10×10^7
Low background sample (PBS) (100 mL)	1.64×10^7	3.58×10^6	3.96×10^6	2.45×10^6	1.47×10^5	2.04×10^6	4.14×10^7	7.03×10^4	1.10×10^7	3.18×10^7	1.05×10^7	2.40×10^6	3.75×10^7
Recovery (%)	7.0	45.6	55.2	2.5	0.9	7.9	54.6	0.3	24.3	99.5	54.3	7.3	73.6
High background sample (sewage) (100 mL)	8.17×10^5	2.01×10^6	1.32×10^6	2.01×10^5	4.48×10^4	1.13×10^6	1.68×10^8	4.29×10^5	4.97×10^6	5.09×10^7	9.33×10^6	1.15×10^6	1.96×10^7
Recovery (%)	0.4	25.6	18.4	0.2	0.3	4.4	221.2	1.8	11.0	159	48.3	3.5	38.5

The table shows also the obtained recovery for each viral type. Quantifications by (RT)-qPCR refer to 100 mL sample

200 mL of 0.25 N glycine (Sigma Aldrich) at pH 9.5 for 30 min with constant rocking at 180 rpm on ice to facilitate the elution of viruses from organic particles. The sample was filtered through a 0.45 µm SteriCup filters (Merck, Cat. No. S2HVU02RE) and subsequently by a 0.22 µm SteriCup filters (Merck, Cat. No. SCGVU02RE). 50 mL of the filtrate was transferred to two different centrifugal filter units with a cut-off size of 10 kDa (Centricon Plus-70; Millipore Cat. No. UFC701008) and were centrifuged 30 min at 3000 *xg*. This step was repeated three times until the entire sample (300 mL) was processed. Viral concentrate was recovered by inverting the unit and centrifuge for 3 min at 1000 *xg* the centrifugal unit. Obtained viral concentrates (approximately 900 to 1000 µl) were collected and stored at - 20 °C until further use.

Quantification of Virus Recovery

The concentrate was nuclease-treated and NA were extracted and enumerated by (RT)qPCR as detailed previously in Nuclease treatment, nucleic acid extraction and (RT)qPCR analysis section. Recovery was calculated as the ratio of spiked virus recovered after sample processing expressed in total genomic copies (GC) and the number of genomic copies (GC) spiked into the 120 mL PBS or sewage as expressed in Eq. 1. The GC per sample and GC added per mL of stock suspensions were calculated according to the formulas presented in Eq. 1 (A and B, respectively).

$$GC \text{ in samples}(A) = \left(C_{PCR} \times \frac{V_{extract}}{V_{PCR}} \right) \times 10 \times \frac{E_{concentrate}}{E_{sample}}$$

$$GC \text{ in stocks}(B) = \left(C_{PCR} \times \frac{V_{Extract}}{V_{PCR}} \right) \times 10 \times \frac{1000}{S} \times F_{stock} \times \frac{V_{processed}}{V_{sample}}$$

$$Recovery(\%) = \frac{\text{measured GC in sample (A)}}{\text{spiked virus GC added (B)}} \times 100$$

C_{PCR} is the template concentration (GC/reaction) determined by (RT)qPCR or qPCR, $V_{extract}$ is the total volume of RNA extract (60 µl), V_{PCR} is the volume of extract analyzed by (RT)qPCR (5 µl for RNA, 10 µl for DNA and 2,5 µl for DNA SYBR Green assay), the factor 10 accounts for the tenfold dilutions of the original NA extracts, S accounts for the volume of the stock (140 µl for stocks), E_{sample} is the volume of the viral concentrate extracted (150 µl or 300 µl); $E_{concentrate}$ is the total volume of the viral concentrate, F_{stock} is the fraction of the extracted sample that is composed by viral particles after nuclease treatment (0.78 for viral stocks),

V_{sample} is the volume of PBS or wastewater used to concentrate viral particles (100 mL), $V_{processed}$ is the total volume of PBS or sewage spiked (120 mL).

Preamplification Protocols

The remainder of the viral concentrates were used for metaviromic analysis. Each concentrate was subjected to three different preamplification protocols, which are summarized in Fig. 1A. In order to make the three protocols comparable, they were designed to capture a similar raw sewage equivalent (6.7 to 7.0 mL per preamplification reaction, see Fig. 1). This entailed those differing starting volumes of viral concentrate had to be used and reagent volumes were adjusted accordingly.

In a first step, all viral concentrates were nuclease treated and NA was extracted as described in detailed in Nuclease treatment, nucleic acid extraction and (RT)qPCR analysis section. Subsequently, each viral extract was preamplified by three different protocols. All the three protocols were started with two independent reactions using 6 μ L of NA. A diagram detailing the three procedures is presented in Fig. 2.

TOTAL (DNA + RNA) Pipeline

The TOTAL protocol (Fig. 2) is based on Sequence-Independent, Single-Primer Amplification (SISPA) and can capture both DNA and RNA viruses (Hjelmsø et al., 2017; Reyes & Kim, 1991; Wang et al., 2003). In brief, NAs were retrotranscribed using SuperScript III (Cat. No. 18080093, Life Technologies) and tagged using primer A (5'-GTTTCC

CAGTCACGATANNNNNNNNN'-3). Hybrid DNA-RNA chains were treated with RNase H (Cat. No. 18,021-071, Thermofisher) to remove RNA and leave only the tagged cDNA. Immediately thereafter, a second cDNA strain was constructed using Exo Klenow polymerase (Cat. No. EP0421, Thermofisher). In order to obtain enough dsDNA as starting material for library preparation, a PCR amplification step with AmpliTaqGold (Cat. No. 4311806, Life Technologies, Austin, Texas, USA) was performed, using the constant region present in primer A targeted by primer B (5'-GTTTCCCAGTCACGATA'-3). After 10 min at 95 °C to activate DNA polymerase, the following PCR program was used: 25 cycles of 30 s at 94 °C, 30 s at 40 °C, and 30 s at 50 °C with a final step of 60 s at 72 °C. Finally, dsDNA amplified products were purified and eluted in 15 μ L of water using the Zymo kit clean and concentrator (D4013, Zymo research, USA).

DNA Pipeline

The DNA pipeline selectively targets DNA viruses and is based on the approach by Kramná & Cinek, (2018). The DNA protocol follows the same SISPA procedure as the TOTAL protocol, but omits the retrotranscription and RNase H steps (Fig. 2). Briefly, DNA strands were separated by heating at 95 °C for 5 min and subsequently chilled on ice to keep dsDNA strands separated. Immediately after, the separated strands were tagged using primer A and Exo Klenow polymerase. A second cycle of Klenow polymerase reaction was performed to ensure the capture of ssDNA viruses. After obtaining tagged dsDNA genomes, a PCR

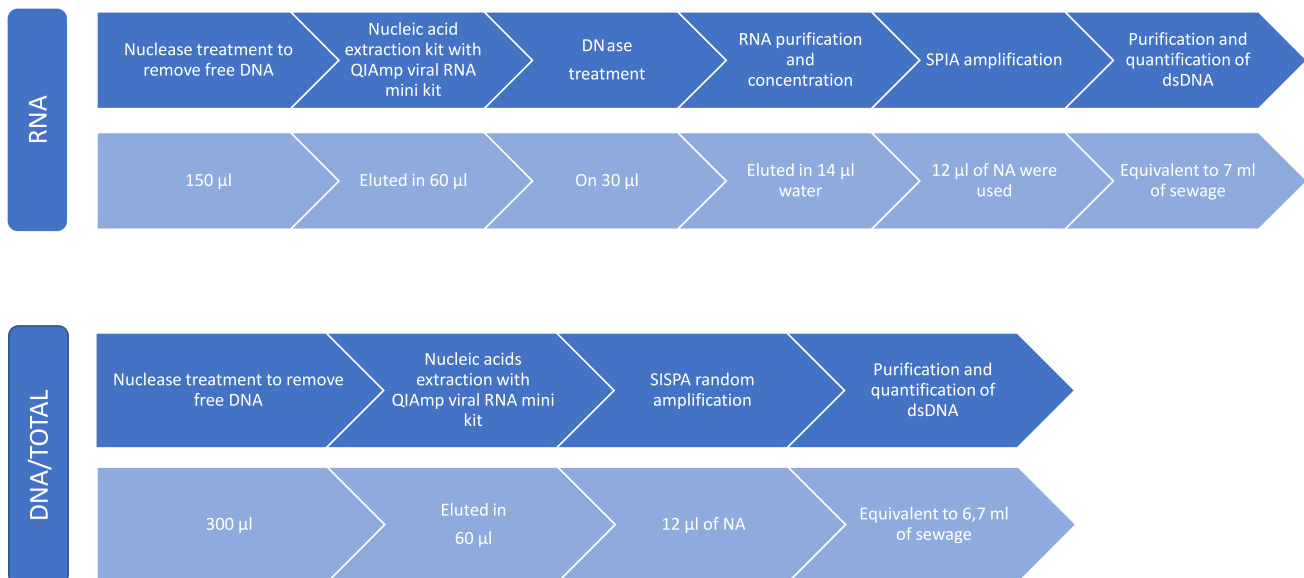


Fig. 1 Summary of the different preamplification protocols tested. The different steps of the protocols are indicated in dark blue. The light blue field indicate the volumes used in each step and the corresponding raw sewage equivalent (colour figure online)

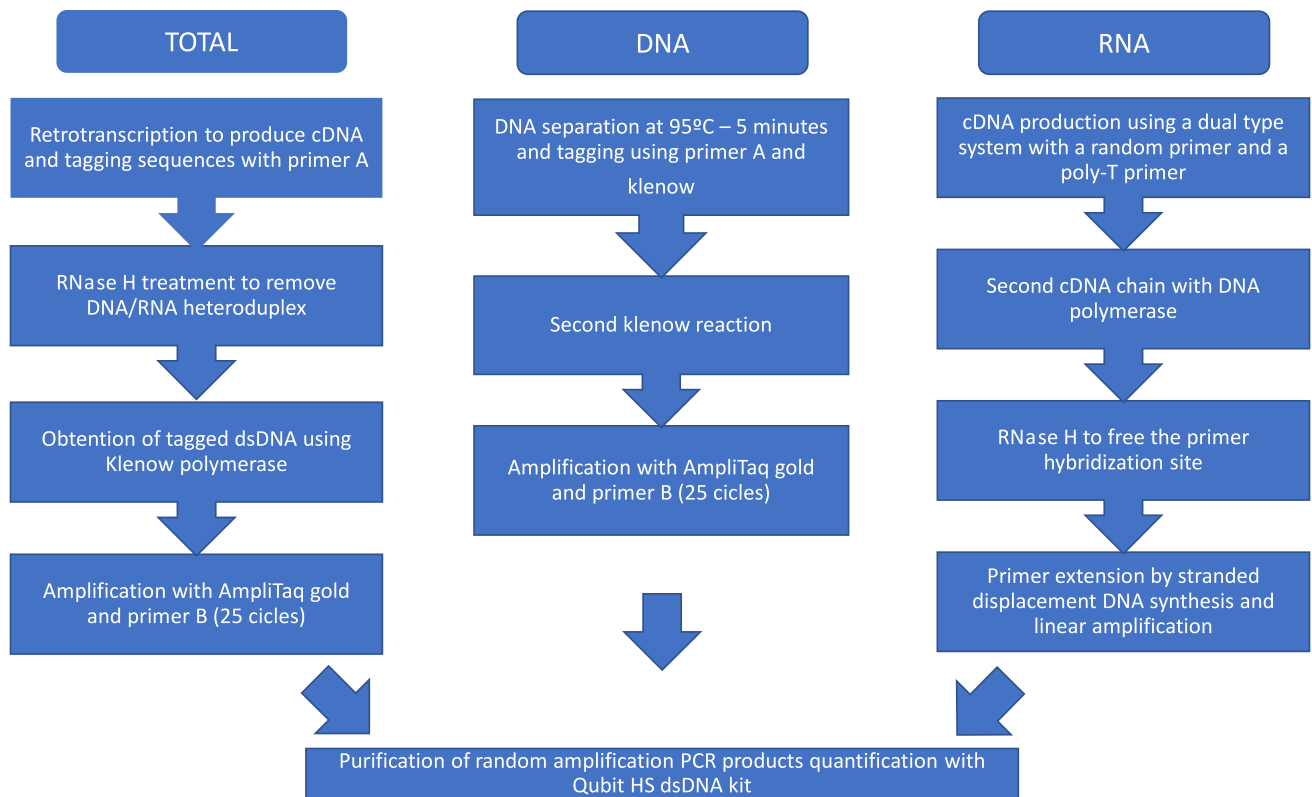


Fig. 2 Detailed scheme with important steps involved in the preamplification for TOTAL, DNA and RNA pipelines (colour figure online)

amplification step using primer B and AmpliTaqGold was performed as described in TOTAL (DNA + RNA) pipeline. Amplified dsDNA products were purified using the Zymo kit clean and concentration and eluted in 15 μ l of water.

RNA Pipeline

RNA was selectively analyzed using a Single Primer Isothermal Amplification (SPIA) approach adapted from (Myrmel et al., 2017). For this protocol, the extracted NA were subjected to a second nuclease step, to remove any viral DNA obtained during NA extraction. Specifically, 30 μ l of NA were DNase treated using 1 μ l of TurboDNase (Ambion Cat. No. AM1907) for 20 min at 37 $^{\circ}$ C. The nuclease reaction was stopped by using DNase Inactivation Reagent included in the TurboDNase kit (Ambion Cat. No. AM1907) according to the manufacturer's instructions. The nuclease-treated NA then underwent an additional purification and concentration step using the RNeasy MinElute Cleanup Kit (QIAGEN, Cat. No. 74204) and purified RNA was eluted in a final volume of 14 μ l. cDNA was synthesized in two independent reactions using 6 μ l of RNA as input with the Ovation RNA-Seq System V2 (NuGen, CA, USA; Part No. 7102) as detailed in the manufacturer's manual. Briefly, cDNA was produced using a combination of random RNA and poly-T

primers. A complementary chain for the cDNA was synthesized with a DNA polymerase. After second strand synthesis, RNA fragments of the SPIA primers were removed by using RNase H and DNA polymerase extended and isothermally amplified the tagged genomes. Amplification products from each of the SPIA reactions were pooled and purified as detailed in TOTAL (DNA + RNA) pipeline section.

Quality Control, Library Preparation Sequencing

All purified dsDNA viral preparations were quantified using Qubit 2.0 (Cat. No. Q32854, Life Technologies, Oregon, USA) and the HS dsDNA kit (Cat. No. Q32851). Libraries were constructed using Nextera XT DNA sample preparation kit (Illumina Inc.) following manufacturer's instructions. Samples were sequenced on two separate runs using Illumina MiSeq 2 \times 300 with paired end reads.

Bioinformatic Analysis

For the bioinformatic analysis, we utilized Genome detective, a web-based bioinformatics pipeline specifically designed to accurately and efficiently identify, assemble, and classify all known viruses present in viral metagenomics data (Vilsker et al., 2019). This platform, accessible at <https://www.genomedetective.com>.

genomedetective.com/, offers a comprehensive solution for conducting robust viral analysis within viral metagenomics datasets. This bioinformatic pipeline involves several steps that are summarized below. Firstly, low-quality reads and adapters were trimmed using Trimmomatic (Bolger et al., 2014). This process ensured that only high-quality reads were retained for subsequent analysis, but duplicate reads were not removed. Identified candidate viral reads were selected using the protein-based alignment method DIAMOND (Buchfink et al., 2014) and non-viral sequences were discarded. Subsequently, viral paired-end reads were assembled with metaSPAdes (Nurk et al., 2017) and taxonomically classified with NCBI-BLASTX and NCBI-BLASTN against NCBI RefSeq viral database (Vilsker et al., 2019; Wheeler et al., 2007). The generated contigs are globally aligned against the closer reference genome included in the viral RefSeq to calculate the genome coverage (Vilsker et al., 2019).

The performance of the three preamplification protocols was evaluated by assessing the number of reads taxonomically assigned to the spiked viruses and their coverage in both the low (PBS) and high (sewage) genomic content.

In an attempt to type retrieve viral species of Enterovirus and Norovirus detected, assembled contigs were additionally analyzed using the RIVM Enterovirus (<https://www.rivm.nl/mpf/typingtool/enterovirus/>) and Norovirus (<https://www.rivm.nl/mpf/typingtool/norovirus/>) typing tools (Kroneman et al., 2011).

Application of the Preamplification Protocols to Determine the Presence of Mammalian Viruses in Sewage

The performance of the preamplification protocols was assessed in unspiked sewage samples. Specifically, three 24-hour composite samples of raw sewage were collected from the Vidy wastewater treatment plant and were concentrated. The first sample (100 mL) was processed as described above. The other two samples had a starting volume of 300 mL, and they were concentrated in their entirety as described above, but without the amendment of the glycine solution. The glycine treatment was omitted to increase the sewage equivalents in the final sequencing reaction.

Results

(RT)qPCR Quantification and Recoveries of Spiked Viruses in PBS and Sewage

Viral quantifications and recoveries in both low genetic background sample (PBS) and high genetic background sample (Sewage) are presented in Table 1. Data to calculate the recovery for spiked viruses is provided in table S4. Data

to generate the calibration standard curves is provided in table S5.

All NTCs and NECs were negative, whereas all spiked viral targets in both matrices fell within the range of the (RT)qPCR calibration curves. In PBS, the recoveries for spiked viruses ranged from a minimum of 0.3% for HAsV MLB1 to a maximum of 99.5% for RoV. In sewage, recoveries ranged from a minimum of 0.2% for EV to a maximum of 221.2% for JCPyV, respectively. Recoveries over 100%, as obtained for JCPyV and RoV, are indicative of a high background concentration of these viruses in sewage. For mammalian reovirus, recoveries over 100% were obtained in both PBS and sewage. Given that the negative NTCs and NECs indicated an absence of contamination, we attribute the high recovery to either a faulty quantification of the spiked virus concentration or a problem with the real-time assay employed. Due to the abnormal values retrieved, mammalian reovirus (RT)qPCR data was excluded from Table 1.

Figure 3 compares the recovery for each virus in PBS and sewage (data given in Table S6). The recoveries of the viruses in PBS were positively correlated with the respective recoveries in sewage ($r=0.67$). This correlation became even stronger when viruses with higher background concentrations in sewage (JCPyV and RoV) were excluded ($r=0.91$).

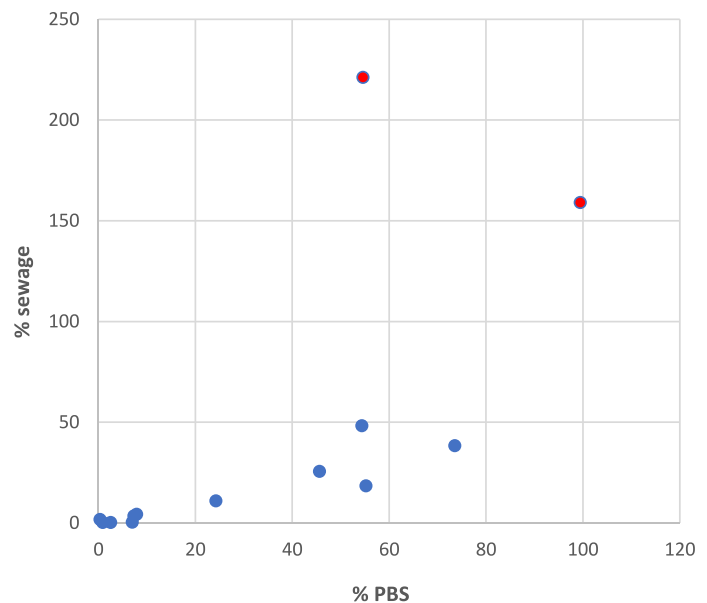
Spiked Viral Reads and Coverage Retrieved According to Preamplification Pipeline

A summary of the mammalian viral species detected in spiked PBS and sewage samples are provided in Tables 2 and 3, respectively. A list of detected mammalian viruses in unspiked sewage samples is provided in Table 4. A summary of the MiSeq output and statistics regarding the number of viral and non-viral sequences, the number of mapped reads and other run related and bioinformatic parameters are given in the Table S7. Additionally, an excel file containing a complete list of retrieved viral reads from spiked samples as well as unspiked sewage samples obtained with genome detective is presented in Tables S8 and S9.

Despite attempts to remove bacteria and free DNA by nuclease treatment, a high percentage of reads were taxonomically classified as non-viral (Table S7). In the PBS sample, which contained only the spiked viruses, both TOTAL and DNA pipelines retrieved 76 and 74% non-viral reads, respectively. For RNA, the number of non-viral reads retrieved was only 14% in spiked PBS. In spiked sewage, the number of non-viral reads was 69 and 70% for TOTAL and DNA pipelines and 59% for the RNA pipeline.

The number of total reads associated with the spiked viruses are plotted in Fig. 4. Data to plot Fig. 4 is presented in Table S10. The RNA pipeline yielded the highest number of reads in both matrices studied, whereas the DNA pipeline yielded the lowest number. Given that

Fig. 3 Comparison of the calculated recovery for all spiked viruses quantified by (RT)qPCR depending on the matrix tested. Presented data omitted mammalian orthoreoviruses as the calculated recovery presented unrealistic recoveries for both PBS and sewage. Plots the recovery (%) of spiked viruses in PBS (X-axis) vs sewage (Y-axis) including RoV and JCPyV (red dots). Higher recoveries for JCPyV and Rotaviruses, can be explained by their high natural presence in sewage. $R=0.67$ (including JCPyV and RoVA); $R=0.91$ (excluding JCPyV and RoVA)



most of the spiked viral species had an RNA genome, this discrepancy is not surprising. The TOTAL pipeline captured a lower number of reads compared to the RNA pipeline, indicating that it was less efficient at capturing RNA viruses. The observed suboptimal performance of the TOTAL pipeline may be attributed to potential inhibition of the reverse transcriptase enzyme during the cDNA production process, though this remains to be confirmed in future work.

The specific number of reads for each spiked virus species, along with the genome coverage obtained, is listed in Tables 2 and 3. For the DNA pipeline, the maximum genome coverage obtained was 100% (JCPyV) in both sewage and PBS. The lowest coverage in PBS was 37% for AAV-2, whereas in sewage, no reads taxonomically assigned to this virus were identified. As expected, the DNA pipeline did not detect any RNA viruses. In the TOTAL pipeline with PBS, genome coverage ranged from 100% (JCPyV) to as low as 7.0% (T4 bacteriophage). Surprisingly, the TOTAL approach performed very poorly for RNA with only Enterovirus B being detectable in PBS but not in sewage. In the RNA pipeline, genome coverage in PBS ranged from a maximum of 100% (EV) to a minimum of 36.2% (MNV). In sewage, the coverage ranged from a maximum of 100% for MS2 bacteriophage to a minimum of 1.5% (MNV). This low genome coverage for MNV is in line with its low recovery (0.23%) by (RT)qPCR. For the segmented dsRNA viruses (Pseudomonas bacteriophage Phi6, RoV-A and MRV), all segments were found in PBS while in sewage 5 out of 11 segments of MRV were undetectable.

We also correlated the number of reads retrieved for each viral species or genome segment in PBS against the corresponding reads in sewage (Fig. 5). A correlation

coefficient of $r=0.81$ was found when data from all the three preamplification protocols were used. When only the RNA preamplification was considered, r was still 0.81. However, sewage systematically yielded fewer reads compared to PBS, up to 3 orders of magnitude. Data used to calculate the correlation coefficients are provided in Table S11.

Relation Between the Number of Reads Detected by Next Generation Sequencing (NGS) and (RT)qPCR Quantifications

To determine if the number of reads associated with a given virus is representative of its concentration, we correlated the Reads Per Kilobase Million (RPKM) of each virus against the corresponding genome copies measured by (RT)qPCR. This analysis was performed separately for spiked PBS and sewage, and individually for each pipeline, whereby only DNA viruses were considered in the DNA and TOTAL pipelines, and only RNA viruses in the RNA pipeline. Furthermore, given the much larger spread of the NGS data compared to the (RT)qPCR data, this analysis was conducted on a log–log scale. A positive correlation could be found between the RPKM and the GC measured by each pipeline for each matrix considered (r ranged from 0.55 to 0.82 in PBS and from 0.62 to 0.82 in sewage, Table S12). High virus concentrations are thus associated with high viral reads (Fig. 6).

Table 2 Reads and contigs taxonomically assigned to viral species in spiked PBS

Virus—Accession number	TOTAL				DNA				RNA			
	n.° Reads	Contigs	Coverage (%)		n.° Reads	Contigs	Coverage (%)		n.° Reads	Contigs	Coverage (%)	
Non-segmented viruses												
JC polyomavirus (JCPyV)—NC_001699.1	83288	1	100		35586	1	100		273	9	66.6	
Human mastadenovirus C (HAdV)—AC_000007.1	10578	2	99.9		4900	6	98.3		117	8	9.64	
Bacteriophage T4—NC_000864.4	13099	3	100		7540	129	90.5		—	—	—	
Adeno-associated virus 2 (AAV-2)—NC_001401.2	247	4	90.8		38	2	37.7		—	—	—	
Hepatovirus A (HAV)—NC_001489.1	—	—	—		—	—	—		1115051	1	99.8	
Sendai virus (SeV)—NC_001552.1	—	—	—		—	—	—		243199	1	99.5	
MS2 bacteriophage—NC_001417.2	—	—	—		—	—	—		536961	1	98.5	
Enterovirus B (EV-B)—NC_038307.1	80	4	47.5		6	2	11		311263	1	100	
Astrovirus MLB1 (HAstV MLB1)—NC_011400.1	—	—	—		—	—	—		995	4	96.5	
Norwalk virus (NoVGII)—NC_008311.1	—	—	—		—	—	—		7622	1	95.3	
Murine norovirus (MNV)—NC_008311.1	—	—	—		—	—	—		147	7	36.2	
Segmented viruses												
Pseudomonas bacteriophage Phi6	—	—	—		—	—	—		3343	1	98.9	
NC_003714.1—NC_003715.1	—	—	—		—	—	—		898	1	95.9	
	—	—	—		—	—	—		677	1	98.1	
	—	—	—		—	—	—		4144	1	99.9	
Mamalian orthoreovirus (MRV)	—	—	—		—	—	—		2335	1	100	
NC_013225.1—NC_013234.1	—	—	—		—	—	—		1803	1	98.1	
	—	—	—		—	—	—		2320	1	98.0	
	—	—	—		—	—	—		565	1	99.9	
	—	—	—		—	—	—		1120	1	99.6	
	—	—	—		—	—	—		679	1	77.7	
	—	—	—		—	—	—		520	1	95.4	
	—	—	—		—	—	—		647	1	94.2	
	—	—	—		—	—	—		186	1	92.3	

Table 2 (continued)

Virus—Accession number	TOTAL						DNA			RNA		
	n.° Reads	Contigs	Coverage (%)	n.° Reads	Contigs	Coverage (%)	n.° Reads	Contigs	Coverage (%)	n.° Reads	Contigs	Coverage (%)
Rotavirus A	—	—	—	—	—	—	—	—	—	54406	1	99.9
(RoV-A)	—	—	—	—	—	—	—	—	—	15606	1	99.0
NC_011500.2—NC_011510.2	—	—	—	—	—	—	—	—	—	68663	1	96.8
Segment 1	—	—	—	—	—	—	—	—	—	25714	1	99.9
Segment 2	—	—	—	—	—	—	—	—	—	12877	1	87.9
Segment 3	—	—	—	—	—	—	—	—	—	4024	1	99.0
Segment 4	—	—	—	—	—	—	—	—	—	50445	1	86.2
Segment 5	—	—	—	—	—	—	—	—	—	15086	1	99.8
Segment 6	—	—	—	—	—	—	—	—	—	11772	1	99.7
Segment 7	—	—	—	—	—	—	—	—	—	2012	1	76.2
Segment 8	—	—	—	—	—	—	—	—	—	1048	1	98.5
Segment 9	—	—	—	—	—	—	—	—	—	—	—	—
Segment 10	—	—	—	—	—	—	—	—	—	—	—	—
Segment 11	—	—	—	—	—	—	—	—	—	—	—	—

The accession number used to establish the genome coverage for each virus and segment are provided

Viral Species with Putative Pathogenic Potential in Unspiked Sewage Samples

A complete inventory of retrieved viral reads from unspiked sewage obtained with genome detective is presented in Table S9. The three preamplification protocols were applied to three unspiked sewage samples to detect the presence of viruses with potential to cause human disease and/or infecting mammalian hosts. As for the spiked samples, the number of non-viral reads in the unspiked samples was high in all of the preamplification protocol used (Table S7). A detailed list of all the viral species identified according to the preamplification protocol used, as well as their genomic coverage is given in Table 4. Detected families which contain human pathogenic viruses include members of the families *Picornaviridae*, *Caliciviridae*, *Astroviridae*, *Reoviridae*, *Adenoviridae*, *Polyomaviridae* and *Parvoviridae* and are presented in Fig. 7.

The TOTAL protocol identified viral reads belonging to the families *Adenoviridae*, *Polyomaviridae* (BKPyV and JCPyV) and *Picornaviridae*. However, the number of reads and the coverage retrieved for *Adenoviridae* and *Picornaviridae* contigs did not allow a proper typing of these viral species. In addition, enteroviruses were only detected in one of the three samples. Obtained results suggests a similar problem with RNA in the TOTAL pipeline as observed for the spiked experiment. DNA preparation identified members of the *Parvoviridae* family but no other any viral pathogens.

The RNA protocol yielded members of the families *Reoviridae*, *Picornaviridae*, *Astroviridae* and *Caliciviridae*. For HAsV MLB1, the coverage was as high as 91.7%, retrieving almost it's full genome. In addition, several genera within the *Picornaviridae* family were identified, including *Enterovirus* (EV species C and Rhinovirus A and B), *Salivirus*, *Kobuvirus* and *Cardiovirus*. For the *Caliciviridae* family, important pathogens such as NoVGI and NoVGII as well as sapoviruses were detected. In the case of NoVGI and NoVGII the genome coverage for one of the samples was high enough to enable viral typing of NoVGI, identifying a NoVGI.P1 | GI.9. Other retrieved contigs of NoVGII did not fully cover the ORF1/ORF2 junction and could not be typed.

Within *Reoviridae*, up to 6 different rotavirus A segments could be identified in one of the studied samples. However, none of the segments belonged to the VP4 region and only a partial sequence of VP7 was retrieved (Rotavirus A strain RVA/Human-wt/RUS/NN1918-16/2016/G1P8; Accession number: MN337577, sharing 99.0% nucleotide identity over 100% coverage).

Among non-human virus sequences, the DNA protocol detected mostly head-tail bacteriophages from the previously classified families *Podoviridae*, *Myoviridae* and *Siphoviridae* and some unclassified ssDNA circular viruses. Recently, these families have been abolished and reclassified

Table 3 (continued)

Virus	TOTAL			DNA			RNA		
	n.° Reads	Contigs	Coverage (%)	n.° Reads	Contigs	Coverage (%)	n.° Reads	Contigs	Coverage (%)
Rotavirus A	-	-	-	-	-	-	21781	2	89.5
(RoV-A)	-	-	-	-	-	-	4214	1	90.0
NC_011500.2—NC_011510.2	-	-	-	-	-	-	55097	1	94.8
Segment 1	-	-	-	-	-	-	28312	1	64.8
Segment 2	-	-	-	-	-	-	148	1	35.2
Segment 3	-	-	-	-	-	-	1089	1	39.1
Segment 4	-	-	-	-	-	-	32734	1	93.7
Segment 5	-	-	-	-	-	-	2296	1	94.3
Segment 6	-	-	-	-	-	-	2417	1	96.2
Segment 7	-	-	-	-	-	-	155	1	53.3
Segment 8	-	-	-	-	-	-	237	1	24.7
Segment 9	-	-	-	-	-	-	-	-	-
Segment 10	-	-	-	-	-	-	-	-	-
Segment 11	-	-	-	-	-	-	-	-	-

into 22 different viral families as detailed in Turner et al., 2023. Finally, the RNA pipeline was dominated by phytoviruses from the families *Virgaviridae*, *Tombusviridae*, *Alphaflexiviridae* and *Tymoviridae*. Sequences related to unclassified picorna-like viruses and some insect viruses from the family *Dicistroviridae* were detected in all samples with some mammalian viruses detected sporadically. The predominant viral phage family was *Leviviridae*.

Discussion

High Genetic Background Reduces the Sensitivity of Viral Metagenomics of dsDNA and dsRNA Viruses

Three different protocols to preamplify nucleic acids before library preparation and NGS were tested in PBS and sewage spiked with several viruses at a known concentration. Across all virus species, the number of associated reads was higher in the PBS sample, which has a low genetic background, compared to the sewage sample with a high genetic background. The lower number of retrieved reads in sewage associated with the spiked viruses resulted in a reduced coverage in this matrix. The difference in the number of reads between the two matrices was more pronounced for double-stranded viruses (dsRNA and dsDNA) when compared to single-stranded viruses (ssRNA), if indigenous viruses in sewage (JCPyV and RoV-A) are excluded. The difficulty in capturing double-stranded RNA genomes may be due to an incomplete denaturation and tagging of their chains during cDNA production. Despite the high prevalence and viral loads of dsRNA viruses such as rotaviruses in sewage measured by (RT)qPCR they are rarely detected by NGS (Wang et al., 2020), with rare exceptions in viral metagenomics studies that use enrichment approaches (Martínez-Puchol et al., 2020; Strubbia et al., 2019) or a different matrix like sludge (Bibby & Peccia, 2013). Difficulties in sequencing dsRNA genomes have been highlighted and addressed by some authors by developing specific protocols to increase the reads associated to this particular conformation (Wilcox et al., 2019). Similarly, this issue might affect negatively double-stranded DNA conformations, which have been detected by qPCR in sewage but unable to be detected by using a viral metagenomics approach (Fernandez-Cassi et al., 2018).

An association Exists Between Virus Recovery in Sewage and PBS and Between Virus Concentration and RPKM

To assess virus recovery, 14 different viruses were spiked at approximately the same concentration both in Sewage and PBS samples. Recoveries of spiked viruses were assessed

Table 4 Detailed list of mammalian viral species detected in unspiked sewage samples

Pipeline used	Total			DNA			RNA		
	n.° reads (sample 1/2/3)	Contigs(sample 1/2/3)	Coverage (%) (sample 1/2/3)	n.° reads (sample 1/2/3)	Contigs (sample 1/2/3)	Coverage (%) (sample 1/2/3)	n.° reads (sample 1/2/3)	Contigs (sample 1/2/3)	Coverage (%) (sample 1/2/3)
human mastadenovirus F	4/-/-	2/-/-	2.4/-/-	-	-	-	-	-	-
Human polyomavirus 1	-/6/-	-/2/-	-/16.5/-	-	-	-	-	-	-
Human polyomavirus 2	-/8/-	-/2/-	-/23.5/-	-	-	-	-	-	-
Human bocavirus 3	-/2/-	-/1/-	-/7.3/-	-/2/-	-/1/-	-/7.1/-	-	-	-
Adeno-associated virus—2	-/2/-	-/1/-	-/8.5/-	-/8/-	-/2/-	-/19.1/-	-	-	-
Enterovirus A	-	-	-	-	-	-	-/6/-	-/1/-	-/7.9/-
Enterovirus B	-/2/-	-/2/-	-/5.7/-	-	-	-	24/84/3	1/2/2	3.6/6.3/9.2
Enterovirus C	-	-	-	-	-	-	654/100/16	6/9/4	43.3/50.2/26.2
Enterovirus D	-	-	-	-	-	-	-/4	-/1	-/5.6
Rhinovirus A	-	-	-	-	-	-	-/6	-/3	-/13.6
Rhinovirus B	-	-	-	-	-	-	-/4	-/2	-/11.1
Parechovirus A	-	-	-	-	-	-	-/13	-/4	-/23.5
Salivirus A	-	-	-	-	-	-	197/67/162	7/5/7	24/16.4/63.2
Aichi virus	-	-	-	-	-	-	1451/509/290	8/9/7	46.8/63/55.4
Cardiovirus B	-	-	-	-	-	-	31/34/-	4/2/-	14.4/9.5/-
Cosavirus E	-	-	-	-	-	-	-/6/-	-/1/-	-/5.8/-
Cosavirus A	-	-	-	-	-	-	-/6/-	-/1/-	-/4.6/-
Rabbit kobuvirus	-	-	-	-	-	-	-/6/-	-/1/-	-/4.7/-
Norovirus GI	-	-	-	-	-	-	-/11/156	-/3/5	-/10.6/75
Norovirus GII	-	-	-	-	-	-	-/65/35	-/6/7	-/34.4/41.7
Norwalk virus	-	-	-	-	-	-	6/-/-	1/-/-	3.6/-/-
Sapporo virus	-	-	-	-	-	-	27/24/25	2/2/4	5.2/8.2/30.5
Mamastrovirus 1	-	-	-	-	-	-	-/2038/-	-/2/-	-/91.7/-
Mamastrovirus	-	-	-	-	-	-	1210/-/-	10/-/-	61.9/-/-
Mamastrovirus 2	-	-	-	-	-	-	-/4	-/1	-/15.7
Mamastrovirus 9	-	-	-	-	-	-	-/20/-	-/3/-	-/19.1/-
Astrovirus MLB1	-	-	-	-	-	-	13/58/38	2/5/4	3.5/40.3/47.3
Astrovirus MLB2	-	-	-	-	-	-	33/-/4	3/-/1	8.9/-/6.7
HMO astrovirus A	-	-	-	-	-	-	-/132/218	-/6/4	-/55.8/80.7
Astrovirus VA3	-	-	-	-	-	-	-/6/22	-/1/3	-/6.2/31.5
Bastrovirus	-	-	-	-	-	-	-/28/-	-/3/-	-/11.9/-
Chicken astrovirus	-	-	-	-	-	-	-/10/-	-/1/-	-/5.1/-

Table 4 (continued)

Pipeline used	Total			DNA			RNA		
	n.° reads (sample 1/2/3)	Contigs(sample 1/2/3)	Coverage (%) (sample 1/2/3)	n.° reads (sample 1/2/3)	Contigs (sample 1/2/3)	Coverage (%) (sample 1/2/3)	n.° reads (sample 1/2/3)	Contigs (sample 1/2/3)	Coverage (%) (sample 1/2/3)
Viral Species									
Rotavirus A	-	-	-	-	-	-	200/-/-	2/-/-	23.3/-/-
	-	-	-	-	-	-	77/-/-	2/-/-	25.4/-/-
	-	-	-	-	-	-	148/-/-	1/-/-	14.1/-/-
	-	-	-	-	-	-	121/-/-	1/-/-	40.6/-/-
	-	-	-	-	-	-	179/-/-	1/-/-	32.3/-/-
	-	-	-	-	-	-	8/-/-	1/-/-	28.9/-/-

The character “/” is used to delimit results from three different sewage samples tested. For each sample, only the best match for a given species is indicated. More information about the detected species and each of the samples is given in Supplementary Table 9

by using specific (RT)qPCR. For most of the viruses tested, recoveries during sample concentration were somewhat higher in PBS compared to sewage, yet recoveries in the two matrices were positively correlated. Lower recoveries in sewage were expected due adsorption of viruses to particulate matter and subsequent losses during sample filtration steps (Gutierrez & Nguyen, 2012). Furthermore, higher quantifications of in PBS is expected by (RT)qPCR, due to the absence of inhibitory substances, which are abundant in sewage samples (Schrader et al., 2012). Only two spiked viruses, namely JCPyV and RoV-A, exhibited a higher concentration in sewage compared to PBS, resulting in recoveries over 100%. These viruses are commonly detected in high concentrations in raw sewage (Rusiñol et al., 2015; Wang et al., 2020). Both viruses were detected using (RT)qPCR on a different aliquot of the sample, presenting 2.32×10^7 GC/L for RoV-A as published in Li et al., (2023) hence recoveries over 100% are not unexpected.

Despite EV-B and HAV having a very low (<0.5%) recovery in sewage when quantified by (RT)qPCR, we retrieved a high number of reads and nearly complete coverage of the genome. In addition, in each matrix, a positive correlation between the log transformed (RT)qPCR and RPKM for each spiked virus and pipeline were found, indicating that sequencing can be used in a semiquantitative way. The preamplification procedure to achieve enough DNA for library preparation unavoidably introduces a bias in the relative abundance of the original sample which depends on the number of amplification cycles or the duration of the amplification cycle in isothermal methods (Regnault et al., 2021; Wang et al., 2023). This bias might be conditioned by other factors such as genome size of the virus or its GC% composition (Cremers et al., 2018).

Other studies have highlighted the difficulty to draw quantitative conclusion from metaviromic studies applied in the context of sewage sequencing (Hjelmsø et al., 2017; Wang et al., 2020). In addition, it has been reported the existing difficulties to detect by NGS means viruses that were highly abundant by (RT)qPCR (Bibby & Peccia, 2013; Fernandez-Cassi et al., 2018). Other factors besides the number of genomic copies such as genomic secondary structures might play a role in the ability to detect viral species by NGS (Price & Garhyan, 2017).

Virus Selective Sample Preparation do Not Suppress Non-viral Reads

We made a concerted effort to enrich the samples in viruses while removing bacteria. Specifically, we used a gentle bacterial removal procedure (disposable stericup filters) to avoid bacterial membrane disruption and applied nuclease treatment prior to viral genome extraction to remove free genetic material. Nevertheless, a high proportion of reads

Fig. 4 Number of total reads associated with the spiked viruses in each of the three preamplification protocols, in spiked PBS and spiked sewage

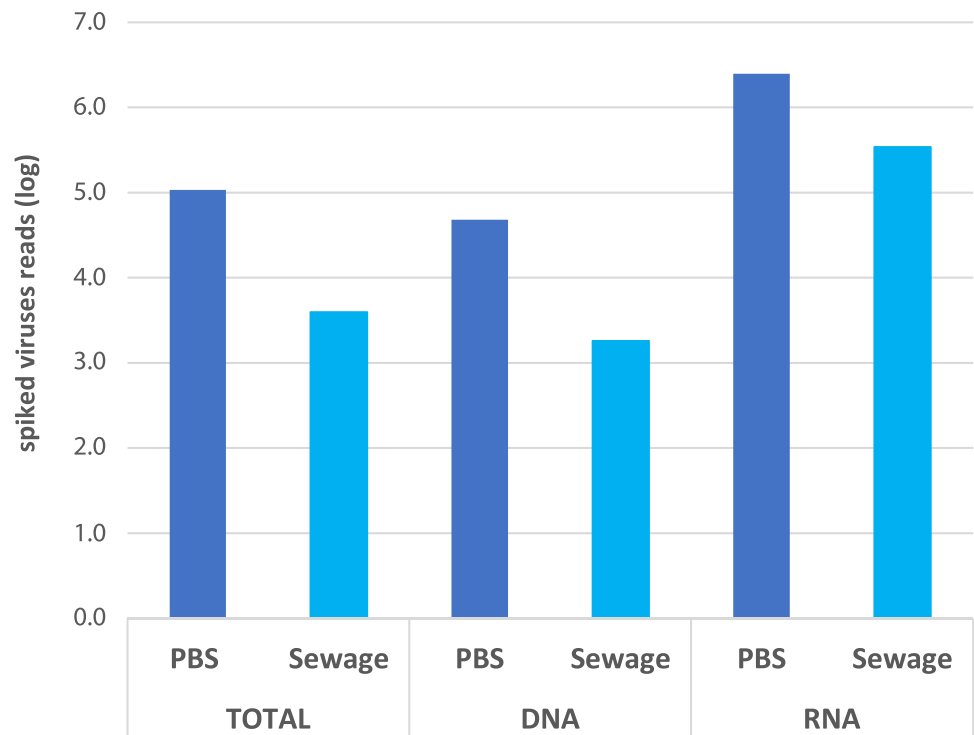
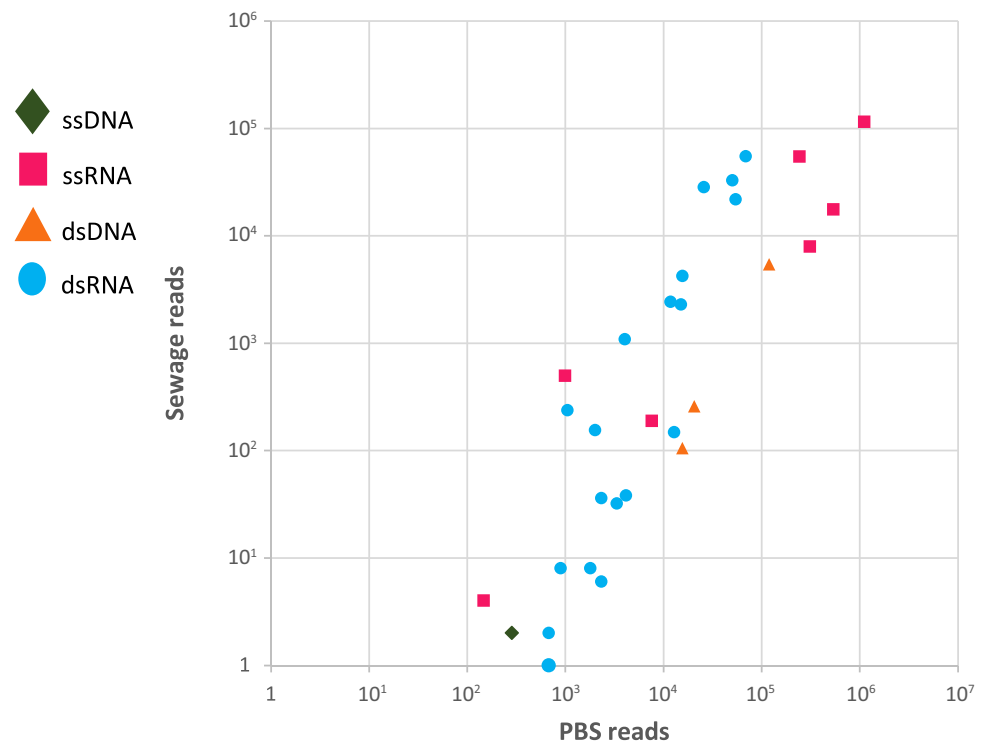


Fig. 5 Number of reads in PBS (a low background genomic content) vs sewage (a high background genomic content) for all viruses and segments detected for all pipelines (TOTAL, DNA and RNA). Each symbol represents a virus or viral segment. Each possible genomic conformation is presented in a different color and shape.



were taxonomically classified as non-viral (from 59 to 76% of the total reads), in line with values reported in previous studies (Cantalupo et al., 2011; Fernandez-Cassi et al., 2018; Hjelmsø et al., 2019). Poor performance of negative selection methods to reduce bacterial contamination have been

previously reported in the context of viral metagenomics (Hall et al., 2014). The presence of contaminants is a common issue on viromics studies, where these contaminants can be introduced at any point from stock production to subsequent concentration, amplification and sequencing steps

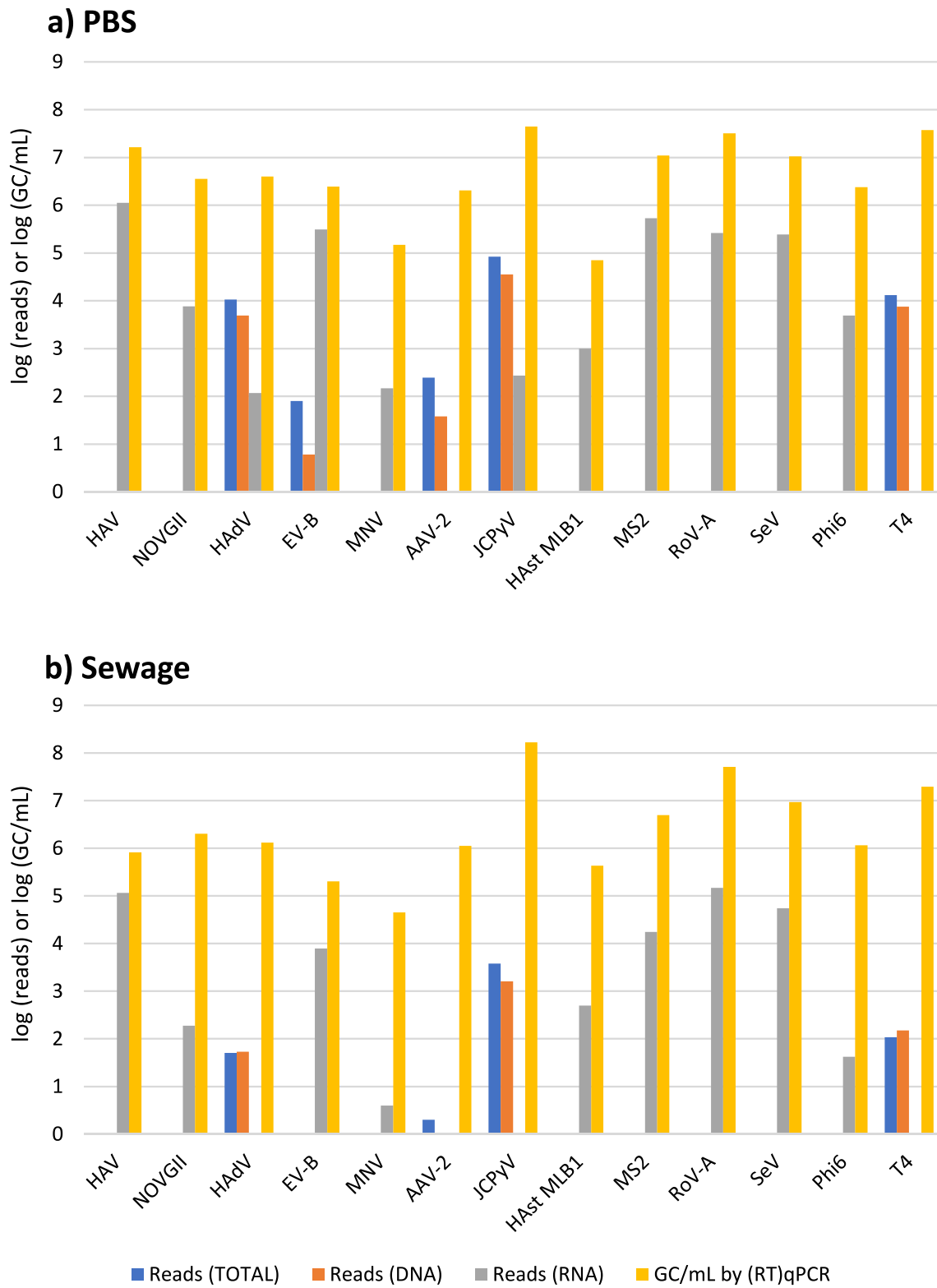


Fig. 6 Stacked barplots comparing the log (RT)qPCR expressed as GC/mL with the log number of reads retrieved for PBS (a) and Sewage (b)

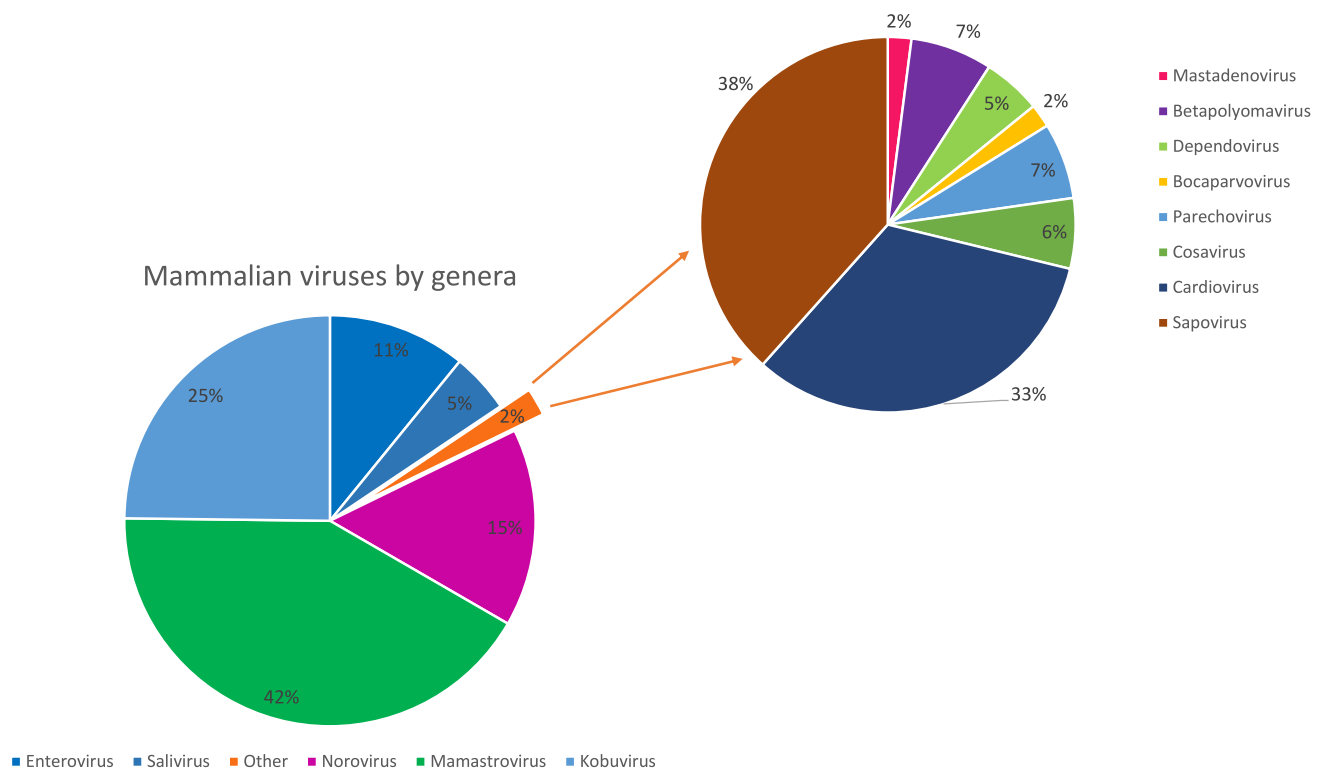


Fig. 7 Detected genera in all the unspiked sewage samples, including the three different preamplification methods

as reviewed by Juraszet et al., (2021). These contaminants are inherent and unique in any processing pipeline, associated to sample processing, nucleic acids extraction and library preparation (Asplund et al., 2019). Our hypothesis is that the DNA pipeline is more sensitive to capture these contaminants in comparison to the TOTAL pipeline. Despite this lower efficiency, an important part of these sequences could be incorrectly considered as bacterial due to the high homology between phages and their hosts or viral sequences yet to be characterized also known as viral dark matter (Santiago-Rodriguez & Hollister, 2022). Better bioinformatic tools not limited by alignment methodologies and the lower diversity within viral databases are needed, especially in the context of environmental studies (Krishnamurthy & Wang, 2017). Further research is needed to improve the selective amplification of viral targets in complex matrices.

Specific RNA Amplification in Unspiked Sewage Samples Improves the Detection of Mammalian RNA Viruses

The application of the three tested protocols to three different unspiked sewage samples showed an improved performance for the RNA pipeline compared to the TOTAL one to identify viruses within a genus that harbor viral species with pathogenic potential. The fact that this preamplification

method contains a post-extraction DNase step might increase the removal of important dsDNA bacteriophages that are predominant in sewage. In unspiked sewage samples, the RNA pipeline based on a SPIA amplification detected viral reads related to mammalian viruses. Important families with pathogenic potential detected in the analyzed samples include *Reoviridae*, *Picornaviridae*, *Astroviridae* and *Caliciviridae* members. Some of the retrieved viral species had enough genome coverage to allow viral typing and sporadically, it was possible to retrieve almost their full genome. However, the low relative abundance of mammalian viruses in proportion with phytoviruses produced non-overlapping contigs of mammalian viruses, limiting the capacity of an appropriate typing. Interestingly, several reads related to human rotavirus segments were detected in one of the unspiked samples. This could be indicating that this pipeline is capable of tagging dsRNA genomes.

Specific DNA Amplification in Unspiked Sewage Samples Does Not Improve the Detection of Mammalian DNA Viruses

An inferior performance of the DNA compared to the TOTAL pipeline was observed in the unspiked sewage samples. Here, the only genus containing viruses with a putative pathogenic potential identified by the DNA pipeline

was a species from the *Parvoviridae* family, whereas the TOTAL pipeline identified members of *Adenoviridae* and *Polyomaviridae*. Aiming to solely amplify DNA viruses did not improve the detection of mammalian viruses neither in the unspiked sewage samples nor the spiked ones when compared to the TOTAL pipeline. The underperformance of the DNA pipeline may be rationalized by a more efficient tagging of the dsDNA bacteriophages from the class Caudoviricetes which are predominant in viral metagenomes from environmental samples including sewage (Cantalupo et al., 2011; Fernandez-Cassi et al., 2018; Ng et al., 2012). In addition, this pipeline included a denaturation step at 95 °C which aimed to more efficiently tag mammalian dsDNA viruses, but also led to more efficient tagging of indigenous head–tail dsDNA bacteriophages.

Consequently, the abundance of mammalian dsDNA viruses may be further biased and reduced by the subsequent PCR amplification (Gutierrez & Nguyen, 2012; Karlsson et al., 2013) and the limited output capacity of the MiSeq platform. Other alternatives to preamplify viral genomes should be pursued to specifically amplify DNA viruses (e.g., Multiple displacement amplification (MDA) amplification).

Strengths and Limitations of the Current Study

An important limitation of the study is the number of samples processed. A higher number of processed samples, spiked PBS and sewage and unspiked sewage, using the three described pipelines would provide more robustness to the results presented in the current study. Studies with a similar approach but with a higher number of samples should be pursued to confirm or discard the relationship between (RT) qPCR and the retrieved results by NGS. Recently, research conducted by Crossette et al., (2021) and Langenfeld et al., (2022) have used spike-in approaches for quantifying gene concentrations, including Antibiotic Resistant Genes (ARGs), marine phages and human DNA viruses such as polyomavirus, papillomavirus, adenovirus, in environmental samples to prove the potential of quantitative metagenomics. Unfortunately, these studies were focused on DNA targets, avoiding the inclusion of RNA standards. Furthermore, it is imperative that these studies encompass viruses with larger genome sizes, such as Herpesvirus or Poxvirus, to obtain a more comprehensive understanding. In this work, viruses with larger genome sizes (e.g., HAdV, RoV, SeV, T4; see Table S12) tended to yield a greater number of reads in sewage (Fig. 6). In this sense, future research following a similar approach but including more varied genome sizes needs to be conducted.

The preamplification methods have proven to present inherent biases as reported by Roux et al., (2016), which showed a preference of MDA for ssDNA viruses. Despite

being widely used, SISPA has been associated with biases too and it's not recommended for the study of RNA viruses according to López-Labrador et al., (2021). These biases do not only affect the preamplification phase but also, they can be linked to the library preparation (Pérez-Cataluña et al., 2021), the impact of the bioinformatic pipeline selected (Sutton et al., 2019) or the sequencing technology employed (de Vries et al., 2021).

Given the various biases involved in preparing viromic samples, it is necessary to establish comprehensive protocols that assess multiple variables throughout both the wet laboratory procedures and the subsequent data and bioinformatic analyses. To address this issue, a methodology similar to the studies conducted by Conceição-Neto et al., (2015, 2018) or Regnault et al., (2021) should be employed, wherein a systematic examination of the viral losses occurring during the wet lab processing is conducted specifically within the context of sewage viral metagenomics. These protocols should be validated using high-quality reference materials that encompass a range of viral genome conformations, lengths, and genome content, as outlined by Santiago-Rodríguez and Hollister, (2020).

A problem with cDNA production during the retrotranscription step in the TOTAL pipeline cannot be ruled out, as very few RNA viruses were detected in both the spiked sample and the native unspiked sewage samples. To identify issues during the retrotranscription step, it would be beneficial to include an RNA process control such as mengovirus, as recommended by Van Borm et al. (2020). This measure would help in pinpointing performance issues, including potential inhibition during cDNA production by retrotranscriptase. In addition, previous studies conducted in the field have indicated that the inaccurate termination of RNases may lead to the loss of RNA process controls that were spiked into the samples (Adriaenssens et al., 2018). This loss could potentially arise from an incompatibility in the inactivation of various nucleases. We acknowledge this possibility, which could not be definitively ruled out in the current study.

Conclusions

1. We developed a RNA pipeline based on SPIA to study RNA viruses in sewage. The developed protocol provided good coverage for RNA viruses both in spiked PBS and sewage, allowing the detection of all RNA viruses spiked. Its application in unspiked sewage samples allowed the detection of putatively pathogenic RNA viruses. This method could also be used as a tool to perform ecological studies on the RNA virome from sewage.

2. The specific preamplification for DNA viruses didn't increase the detection of mammalian viruses neither in the spiked samples nor the unspiked sewage samples. Alternative metaviromic preamplification protocols to increase the sensitivity for DNA animal viruses in sewage are needed.
3. Recoveries of the spiked viruses were higher in PBS for all viruses except for those viruses naturally occurring in Sewage used for the experiment. A log–log correlation between the (RT)qPCR quantifications in both matrices and the Reads Per Kilobase Million (RPKM) was found.
4. To implements sewage viral metagenomics as a surveillance tool for epidemiological purposes, positive selection methods such as post-library enrichment panels targeting viral families with pathologic potential are recommended to increase the coverage and sensitivity. This seems particularly relevant for DNA viruses for which specific amplification of ssDNA and DNA seems to favor viral phages.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12560-024-09594-3>.

Acknowledgements This work was supported by EPFL discretionary funding. Xavier Fernandez Cassi was a fellow of the European Union's Horizon2020 research and innovation program under the Marie Skłodowska–Curie Grant Agreement No. 754462.

Author contributions X.F.C and T.K. wrote the main manuscript text. X.F.C. conducted all the experimental research. X.F.C. prepared all figures and tables. T.K. and X.F.C. conducted all the analyses. T.K. and X.F.C. designed the study. All authors reviewed the manuscript.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. Marie Skłodowska–Curie Grant Agreement, 754462, EPFL discretionary funding.

Data Availability The complete sequencing data are available at <https://www.ncbi.nlm.nih.gov/bioproject/948647> with the BioSample accessions numbers SAMN33911627, SAMN33911628, SAMN33911629, SAMN33911630 and SAMN33911631 (SRA SRR23986800—SRR23986814).

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will

need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- 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*. <https://doi.org/10.1128/msystems.00025-18>
- Asplund, M., Kjartansdóttir, K. R., Møllerup, S., Vinner, L., Fridholm, H., Herrera, J. A. R., et al. (2019). Contaminating viral sequences in high-throughput sequencing viromics: A linkage study of 700 sequencing libraries. *Clinical Microbiology and Infection*, 25(10), 1277–1285. <https://doi.org/10.1016/j.cmi.2019.04.028>
- Bačnik, K., Kutnjak, D., Pecman, A., Mehle, N., Tušek Žnidarič, M., Gutiérrez Aguirre, I., & Ravnikar, M. (2020). Viromics and infectivity analysis reveal the release of infective plant viruses from wastewater into the environment. *Water Research*. <https://doi.org/10.1016/j.watres.2020.115628>
- Bibby, K., & Peccia, J. (2013). Identification of viral pathogen diversity in sewage sludge by metagenome analysis. *Environmental Science & Technology*, 47(4), 1945–1951. <https://doi.org/10.1021/es305181x>
- Bisseux, M., Didier, D., Audrey, M., Christine, A., Hélène, P. L., Jean-Luc, B., & Cécile, H. (2020). Monitoring of enterovirus diversity in wastewater by ultra-deep sequencing: An effective complementary tool for clinical enterovirus surveillance. *Water Research*. <https://doi.org/10.1016/j.watres.2019.115246>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*. <https://doi.org/10.1038/nmeth.3176>
- Bustin, S. A., Benes, V., Garson, J. A., Hellems, J., Huggett, J., Kubista, M., et al. (2009). The MIQE guidelines: Minimum Information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, 55(4), 611–622. <https://doi.org/10.1373/clinchem.2008.112797>
- Cantalupo, P. G., Calgua, B., Zhao, G., Hundesa, A., Wier, A. D., Katz, J. P., et al. (2011). Raw sewage harbors diverse viral populations. *Mbio*, 2(5), e00180–e211. <https://doi.org/10.1128/mBio.00180-11>
- Conceição-Neto, N., Yinda, K. C., Van Ranst, M., & Matthijssens, J. (2018). NetoVIR: Modular Approach to customize sample preparation procedures for viral metagenomics. *Methods in Molecular Biology*, 1838, 85–95. https://doi.org/10.1007/978-1-4939-8682-8_7
- Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., et al. (2015). Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Scientific Reports*. <https://doi.org/10.1038/SREP16532>
- Cremers, G., Gambelli, L., van Alen, T., van Niftrik, L., & Op den Camp, H. J. M. (2018). Bioreactor virome metagenomics sequencing using DNA spike-ins. *PeerJ*. <https://doi.org/10.7717/PEERJ.4351>
- Crossette, E., Gumm, J., Langenfeld, K., Raskin, L., Duhaime, M., & Wigginton, K. (2021). Metagenomic quantification of genes with internal standards. *Mbio*, 12(1), 1–12.
- de Jonge, E. F., Peterse, C. M., Koelwijn, J. M., van der Drift, A. M. R., van der Beek, R. F. H. J., Nagelkerke, E., & Lodder, W. J. (2022). The detection of monkeypox virus DNA in wastewater samples in the Netherlands. *The Science of the Total Environment*. <https://doi.org/10.1016/J.SCITOTENV.2022.158265>
- de Vries, J. J. C., Brown, J. R., Couto, N., Beer, M., Le Mercier, P., Sidorov, I., et al. (2021). Recommendations for the introduction of

- metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting. *Journal of Clinical Virology: the Official Publication of the Pan American Society for Clinical Virology*. <https://doi.org/10.1016/J.JCV.2021.104812>
- Duarte, M. F., de Andrade, I. A., Silva, J. M. F., de Melo, F. L., Machado, A. M., Inoue-Nagata, A. K., & Nagata, T. (2023). Metagenomic analyses of plant virus sequences in sewage water for plant viruses monitoring. *Tropical Plant Pathology*. <https://doi.org/10.1007/s40858-023-00575-8>
- Fernandez-Cassi, X., Scheidegger, A., Bänziger, C., Cariti, F., Tuñas Corzon, A., Ganesanandamoorthy, P., et al. (2021). Wastewater monitoring outperforms case numbers as a tool to track COVID-19 incidence dynamics when test positivity rates are high. *Water Research*. <https://doi.org/10.1016/J.WATRES.2021.117252>
- Fernandez-Cassi, X., Timoneda, N., Martínez-Puchol, S., Rusiñol, M., Rodríguez-Manzano, J., Figuerola, N., et al. (2018). Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. *Science of the Total Environment*, 618, 870–880. <https://doi.org/10.1016/j.scitotenv.2017.08.249>
- Gutierrez, L., & Nguyen, T. H. (2012). Interactions between rotavirus and swansee river organic matter: Aggregation, deposition, and adhesion force measurement. *Environmental Science and Technology*, 46(16), 8705–8713. <https://doi.org/10.1021/ES301336U>
- Hall, R. J., Wang, J., Todd, A. K., Bissielo, A. B., Yen, S., Strydom, H., et al. (2014). Evaluation of rapid and simple techniques for the enrichment of viruses prior to metagenomic virus discovery. *Journal of Virological Methods*, 195, 194–204. <https://doi.org/10.1016/j.jviro.2013.08.035>
- Heijnen, L., & Medema, G. (2011). Surveillance of influenza A and the pandemic influenza A (H1N1) 2009 in sewage and surface water in the Netherlands. *Journal of Water and Health*. <https://doi.org/10.2166/wh.2011.019>
- Hellmér, M., Paxéus, N., Magnius, L., Enache, L., Arnholm, B., Johansson, A., et al. (2014). Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. *Applied and Environmental Microbiology*, 80(21), 6771–6781. <https://doi.org/10.1128/AEM.01981-14>
- Hjelmsø, M. H., Hellmér, M., Fernandez-Cassi, X., Timoneda, N., Lukjancenko, O., Seidel, M., et al. (2017). Evaluation of methods for the concentration and extraction of viruses from sewage in the context of metagenomic sequencing. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0170199>
- Hjelmsø, M. H., Møllerup, S., Jensen, R. H., Pietroni, C., Lukjancenko, O., Schultz, A. C., et al. (2019). Metagenomic analysis of viruses in toilet waste from long distance flights—A new procedure for global infectious disease surveillance. *PLoS ONE*. <https://doi.org/10.1371/JOURNAL.PONE.0210368>
- Hughes, B., Duong, D., White, B. J., Wigginton, K. R., Chan, E. M. G., Wolfe, M. K., & Boehm, A. B. (2022). Respiratory Syncytial Virus (RSV) RNA in wastewater settled solids reflects RSV clinical positivity rates. *Environmental Science and Technology Letters*, 9(2), 173–178.
- Huisman, J. S., Scire, J., Caduff, L., Fernandez-Cassi, X., Ganesanandamoorthy, P., Kull, A., et al. (2021). Wastewater-based estimation of the effective reproductive number of SARS-CoV-2. *medRxiv*. <https://doi.org/10.1101/2021.04.29.21255961>
- Iaconelli, M., Valdazo-González, B., Equestre, M., Ciccaglione, A. R., Marcantonio, C., Della Libera, S., & La Rosa, G. (2017). Molecular characterization of human adenoviruses in urban wastewaters using next generation and Sanger sequencing. *Water Research*, 121, 240–247. <https://doi.org/10.1016/j.watres.2017.05.039>
- Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., et al. (2021). Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of wastewater samples. *medRxiv*. <https://doi.org/10.1101/2021.01.08.21249379>
- Jurasz, H., Pawłowski, T., & Perlejewski, K. (2021). Contamination Issue in Viral Metagenomics: Problems, Solutions, and Clinical Perspectives. *Frontiers in Microbiology*. <https://doi.org/10.3389/FMICB.2021.745076>
- Karlsson, O. E., Belák, S., & Granberg, F. (2013). The effect of pre-processing by sequence-independent, single-primer amplification (SISPA) on metagenomic detection of viruses. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*. <https://doi.org/10.1089/bsp.2013.0008>
- Kitajima, M., Iker, B. C., Pepper, I. L., & Gerba, C. P. (2014). Relative abundance and treatment reduction of viruses during wastewater treatment processes — Identification of potential viral indicators. *Science of the Total Environment*, 488, 290–296. <https://doi.org/10.1016/j.scitotenv.2014.04.087>
- Kohl, C., Brinkmann, A., Dabrowski, P. W., Radonic, A., Nitsche, A., & Kurth, A. (2015). Protocol for metagenomic virus detection in clinical specimens. *Emerging Infectious Diseases*, 21(1), 48–57. <https://doi.org/10.3201/eid2101.140766>
- Kramná, L., & Cinek, O. (2018). Virome sequencing of stool samples. *Methods in Molecular Biology*, 1838, 59–83. https://doi.org/10.1007/978-1-4939-8682-8_6
- Krishnamurthy, S. R., & Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Research*, 239, 136–142. <https://doi.org/10.1016/J.VIRUSRES.2017.02.002>
- Kroneman, A., Vennema, H., Deforche, K., v d Avoort, H., Penaranda, S., Oberste, M. S., et al. (2011). An automated genotyping tool for enteroviruses and noroviruses. *Journal of Clinical Virology: THE Official Publication of the Pan American Society for Clinical Virology*, 51(2), 121–125. <https://doi.org/10.1016/j.jcv.2011.03.006>
- Langenfeld, K., Hegarty, B., Vidaurri, S., Crossette, E., Duhaime, M., & Wigginton, K. (2022). A quantitative metagenomic approach to determine population concentrations with examination of quantitative limitations. *bioRxiv*. <https://doi.org/10.1101/2022.07.08.499345>
- Li, C., Sylvestre, É., Fernandez-Cassi, X., Julian, T. R., & Kohn, T. (2023). Waterborne virus transport and the associated risks in a large lake. *Water Research*. <https://doi.org/10.1016/J.WATRES.2022.119437>
- Li, L., Deng, X., Mee, E. T., Collot-teixeira, S., Anderson, R., Schepelmann, S., et al. (2015). Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent. *Journal of Virological Methods*, 213, 139–146. <https://doi.org/10.1016/j.jviro.2014.12.002>
- López-Labrador, F. X., Brown, J. R., Fischer, N., Harvala, H., Van Boheemen, S., Cinek, O., et al. (2021). Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. *Journal of Clinical Virology*. <https://doi.org/10.1016/j.jcv.2020.104691>
- Martínez-Puchol, S., Rusiñol, M., Fernández-Cassi, X., Timoneda, N., Itarte, M., Andrés, C., et al. (2020). Characterisation of the sewage virome: Comparison of NGS tools and occurrence of significant pathogens. *Science of the Total Environment*. <https://doi.org/10.1016/j.scitotenv.2020.136604>
- McCall, C., Wu, H., Miyani, B., & Xagorarakis, I. (2020). Identification of multiple potential viral diseases in a large urban center using wastewater surveillance. *Water Research*. <https://doi.org/10.1016/j.watres.2020.116160>
- Myrmel, M., Oma, V., Khatri, M., Hansen, H. H., Stokstad, M., Berg, M., & Blomström, A.-L. (2017). Single primer isothermal amplification (SPIA) combined with next generation sequencing provides complete bovine coronavirus genome coverage and higher sequence depth compared to sequence-independent single primer amplification (SISPA). *PLoS ONE*, 12(11), e0187780. <https://doi.org/10.1371/JOURNAL.PONE.0187780>
- Ng, T. F. F., Marine, R., Wang, C., Simmonds, P., Kapusinszky, B., Bodhidatta, L., et al. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *Journal of Virology*, 86(22), 12161–12175. <https://doi.org/10.1128/JVI.00869-12>

- Nieuwenhuijse, D. F., Oude Munnink, B. B., Phan, M. V. T., Munk, P., Venkatakrishnan, S., Aarestrup, F. M., et al. (2020). Setting a baseline for global urban virome surveillance in sewage. *Scientific Reports*, *10*(1), 1–13. <https://doi.org/10.1038/s41598-020-69869-0>
- Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*. <https://doi.org/10.1101/gr.213959.116>
- Parras-Moltó, M., Rodríguez-Galet, A., Suárez-Rodríguez, P., & López-Bueno, A. (2018). Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome*, *6*(1), 119. <https://doi.org/10.1186/s40168-018-0507-3>
- Pérez-Cataluña, A., Cuevas-Ferrando, E., Randazzo, W., & Sánchez, G. (2021). Bias of library preparation for virome characterization in untreated and treated wastewaters. *Science of the Total Environment*. <https://doi.org/10.1016/j.scitotenv.2020.144589>
- Price, A., & Garhyan, J. (2017). The impact of RNA secondary structure on read start locations on the Illumina sequencing platform. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0173023>
- Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simón, P., Allende, A., & Sánchez, G. (2020). SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water Research*. <https://doi.org/10.1016/j.watres.2020.115942>
- Regnault, B., Bigot, T., Ma, L., Pérot, P., Temmam, S., & Eloit, M. (2021). Deep impact of random amplification and library construction methods on viral metagenomics results. *Viruses*. <https://doi.org/10.3390/v13020253r>
- Reyes, G. R., & Kim, J. P. (1991). Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Molecular and Cellular Probes*, *5*(6), 473–481.
- Roux, S., Solonenko, N. E., Dang, V. T., Poulos, B. T., Schwenck, S. M., Goldsmith, D. B., et al. (2016). Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ*, *4*(12), e2777–e2777. <https://doi.org/10.7717/PEERJ.2777>
- Rusiñol, M., Fernandez-Cassi, X., Timoneda, N., Carratalà, A., Abril, J. F., Silvera, C., et al. (2015). Evidence of viral dissemination and seasonality in a Mediterranean river catchment: Implications for water pollution management. *Journal of Environmental Management*, *159*, 58–67. <https://doi.org/10.1016/j.jenvman.2015.05.019>
- Samulski, R. J., Chang, L.-S., & Shenk, T. (1987). A recombinant plasmid from which an infectious adeno-associated virus genome can be excised in vitro and its use to study viral replication. *Journal of Virology*, *61*(10), 3096. <https://doi.org/10.1128/jvi.61.10.3096-3101.1987>
- Santiago-Rodríguez, T. M., & Hollister, E. B. (2020). Potential applications of human viral metagenomics and reference materials: Considerations for current and future viruses. *Applied and Environmental Microbiology*, *86*(22), 1–12. <https://doi.org/10.1128/AEM.01794-20>
- Santiago-Rodríguez, T. M., & Hollister, E. B. (2022). Unraveling the viral dark matter through viral metagenomics. *Frontiers in Immunology*, *13*, 1005107. <https://doi.org/10.3389/FIMMU.2022.1005107/BIBTEX>
- Santiso-Bellón, C., Randazzo, W., Pérez-Cataluña, A., Vila-Vicent, S., Gozalbo-Rovira, R., Muñoz, C., et al. (2020). Epidemiological surveillance of norovirus and rotavirus in sewage (2016–2017) in Valencia (Spain). *Microorganisms*. <https://doi.org/10.3390/MICROORGANISMS8030458>
- Schrader, C., Schielke, A., Ellerbroek, L., & Johne, R. (2012). PCR inhibitors - occurrence, properties and removal. *Journal of Applied Microbiology*. <https://doi.org/10.1111/j.1365-2672.2012.05384.x>
- Strahle, L., Garcin, D., & Kolakofsky, D. (2006). Sendai virus defective-interfering genomes and the activation of interferon-beta. *Virology*, *351*(1), 101–111. <https://doi.org/10.1016/j.virol.2006.03.022>
- Strubbia, S., Phan, M. V. T., Schaeffer, J., Koopmans, M., Cotten, M., & Le Guyader, F. S. (2019). Characterization of Norovirus and other human enteric viruses in sewage and stool samples through next-generation sequencing. *Food and Environmental Virology*, *11*(4), 400–409. <https://doi.org/10.1007/s12560-019-09402-3>
- Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P., & Hill, C. (2019). Choice of assembly software has a critical impact on virome characterisation. *Microbiome*. <https://doi.org/10.1186/s40168-019-0626-5>
- Tamaki, H., Zhang, R., Angly, F. E., Nakamura, S., Hong, P.-Y., Yasunaga, T., et al. (2012). Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environmental Microbiology*, *14*(2), 441–452. <https://doi.org/10.1111/j.1462-2920.2011.02630.x>
- Tisza, M., Javornik Cregeen, S., Avadhanula, V., Zhang, P., Ayvaz, T., Feliz, K., et al. (2023). Title: Comprehensive wastewater sequencing reveals community and variant dynamics of the collective human virome. *medRxiv*. <https://doi.org/10.1101/2023.05.03.23289441>
- Tulloch, R. L., Kim, K., Sikazwe, C., Michie, A., Burrell, R., Holmes, E. C., et al. (2023). RAPIDprep: A simple, fast protocol for rna metagenomic sequencing of clinical samples. *Viruses*. <https://doi.org/10.3390/v15041006>
- Turner, D., Shkoporov, A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfnas-Zerbini, P., et al. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Archives of Virology*, *168*(2), 74. <https://doi.org/10.1007/S00705-022-05694-2>
- Van Borm, S., Fu, Q., Winand, R., Vanneste, K., Hakhverdyan, M., Höper, D., & Vandenbussche, F. (2020). Evaluation of a commercial exogenous internal process control for diagnostic RNA virus metagenomics from different animal clinical samples. *Journal of Virological Methods*, *283*(April), 113916. <https://doi.org/10.1016/j.jviromet.2020.113916>
- 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*(5), 871. <https://doi.org/10.1093/BIOINFORMA/TICS/BTY695>
- Wang, D., Urisman, A., Liu, Y. T., Springer, M., Ksiazek, T. G., Erdman, D. D., et al. (2003). Viral discovery and sequence recovery using DNA microarrays. *PLoS Biology*. <https://doi.org/10.1371/journal.pbio.0000002>
- Wang, G., Li, S., Yan, Q., Guo, R., Zhang, Y., Chen, F., et al. (2023). Optimization and evaluation of viral metagenomic amplification and sequencing procedures toward a genome-level resolution of the human fecal DNA virome. *Journal of Advanced Research*. <https://doi.org/10.1016/J.JARE.2022.08.011>
- Wang, H., Neyvaldt, J., Enache, L., Sikora, P., Mattsson, A., Johansson, A., et al. (2020). Variations among viruses in influent water and effluent water at a wastewater plant over one year as assessed by quantitative PCR and metagenomics. *Applied and Environmental Microbiology*. <https://doi.org/10.1128/AEM.02073-20>
- Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., et al. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Research*. <https://doi.org/10.1093/NAR/GKL1031>
- Wilcox, A. H., Delwart, E., & Díaz-Muñoz, S. L. (2019). Next-generation sequencing of dsRNA is greatly improved by treatment with the inexpensive denaturing reagent DMSO. *Microbial Genomics*. <https://doi.org/10.1099/mgen.0.000315>
- Wolfe, M. K., Yu, A. T., Duong, D., Rane, M. S., Hughes, B., Chan-Herur, V., et al. (2023). Use of wastewater for mpox outbreak surveillance in California. *New England Journal of Medicine*, *388*(6), 570–572. <https://doi.org/10.1056/nejmc2213882>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.