Bias of library preparation for virome characterization in untreated and treated wastewaters

Alba Pérez-Cataluña[1,2*], Enric Cuevas-Ferrando[1], Walter Randazzo[1,2,3], Gloria Sánchez[1,2*]

[1]Department of Preservation and Food Safety Technologies, IATA-CSIC, Avda. Agustin Escardino 7, 46980 Paterna, Valencia, Spain

[2]European Virus Bioinformatics Center, Leutragraben 1, 07743 Jena, Germany

[3]Department of Microbiology and Ecology, University of Valencia, Av. Dr. Moliner, 50, 46100 Burjassot, Valencia, Spain

*Corresponding authors:

Gloria Sánchez. Department of Preservation and Food Safety Technologies (IATA-CSIC). Av. Agustín Escardino 7. 46980 Paterna. Valencia. Spain.

Tel.: + 34 96 3900022; Fax: + 34 96 3939301; E-mail: gloriasanchez@iata.csic.es

Twitter: @FoodViruses

Alba Pérez-Cataluña. Department of Preservation and Food Safety Technologies (IATA-CSIC). Av. Agustín Escardino 7. 46980 Paterna. Valencia. Spain.

Tel.: + 34 96 3900022; Fax: + 34 96 3939301; E-mail: alba.perez@iata.csic.es Twitter:

Twitter: @AlbaPerezCat

Abstract

The use of metagenomics for virome characterization and its implementation for wastewater analyses, including wastewater-based epidemiology, has increased in the last years. However, the lack of standardized methods can led to highly different results. The aim of this work was to analyze virome profiles in upstream and downstream wastewater samples collected from four wastewater treatment plants (WWTPs) using two different library preparation kits. Viral particles were enriched from wastewater concentrates using a filtration and nuclease digestion procedure prior to total nucleic acid (NA) extraction. Sequencing was performed using the ScriptSeq v2 RNA-Seq (LS) and the NEBNext® Ultra™ II RNA (NB) library preparation kits. Cleaned reads and contigs were annotated using a curated *in-house* database composed by reads assigned to viruses at NCBI. Significant differences in viral families and in the ratio of detection were shown between the two library kits used. The use of LS library showed *Virgaviridae*, *Microviridae* and *Siphoviridae* as the most abundant families; while *Ackermannviridae* and *Helleviridae* were highly represented within the NB library. Additionally, the two sequencing libraries produced outcomes that differed in the detection of viral indicators. These results highlighted the importance of library selection for studying viruses in untreated and treated wastewater. Our results underline the need for further studies to elucidate the influence of sequencing procedures in virome profiles in wastewater matrices in order to improve the knowledge of the virome in the water environment.

Keywords: Wastewater, Metagenomics, Enteric viruses, viability RT-qPCR.

1. Introduction

The reuse of water, including for irrigation, cooling, and other non-potable applications is an emerging topic due to climate change and water scarcity. Treatment and regeneration of household sewage water in urban regions are usually performed by wastewater treatment plants (WWTPs); however, they are not always able to completely eliminate the microbiological risks present in treated wastewaters (Chalmers et al., 2010; Randazzo et al., 2019; Sano et al., 2016). Fecal bacteria have traditionally been used as indicators for the presence of pathogenic microorganisms even though they fail to detect the presence of human pathogenic enteric viruses (Eslamian, 2016; Gerba et al., 2013; Kitajima et al., 2014). Thus, several viruses (i.e. crAssphage, Pepper mild mottle virus, adenovirus, polyomavirus, …) have been proposed as indicators because of their similarity to pathogenic viruses in terms of environmental stability and resistance to wastewater sanitation treatments (Farkas et al., 2020). The presence of human enteric viruses in treated wastewaters has been well documented (Gerba et al., 2018; Randazzo et al., 2019; Sano et al., 2016), posing public health risk-related concerns also because of their stability into the environment. Thus far, nearly one hundred different types of human enteric viruses are known, which cause a variety of illnesses and diseases in humans (Fong and Lipp, 2005), primarily gastroenteritis and hepatitis, and new pathogenic strains and species continue to be discovered. Among others, the viruses most commonly detected in untreated and treated wastewaters include human norovirus, adenovirus (AdV), enterovirus (EV), sapovirus (SaV), astrovirus (HAstV), rotavirus A (RV), and hepatitis A and E viruses (HAV and HEV) (Haramoto et al., 2018). Surveillance of human enteric viruses in untreated and treated wastewaters is performed by molecular procedures (e.g., real time PCR (qPCR) or digital PCR (dPCR)) (Haramoto et al., 2018). Currently, a wastewater-based epidemiology surveillance has been globally implemented to monitor COVID-19 disease, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with notable implications for public health response in local settings (Bivins et al., 2020; Polo et al., 2020). These approaches require reference sequences for primer and probe design which limit the number and variety of viruses to be analyzed. Alternatively, recent shotgun or untargeted metagenomic approaches enable the simultaneous identification of viral sequences from a sample, referred to as 'virome', which is a diverse community of mainly eukaryotic RNA and DNA eukaryotic viruses and bacteriophages. Virome characterization in wastewater provides a potential solution to the challenges associated with the traditional surveillance of viruses in sewage (Nieuwenhuijse and Koopmans, 2017).

In this pilot study, we have used metagenomics analyses using two different library preparation kits for metagenomic sequencing to characterize the virome composition in influent and effluent samples from four different WWTPs. Thus, the objectives of this study were to: 1) evaluate different sequencing libraries for virome characterization; and 2) investigate virome distribution and diversity in influent and effluent samples.

2. Materials and Methods

2.1. Sample processing

Five-hundred mL of influent (IW) and effluent (EW) grab samples from four different WWTPs were collected in November 2018 in Valencia (Spain). Treatment plants differed in the number of equivalent population, the volume of treated wastewater and the disinfection treatments (Table S1). *Escherichia coli* counts, expressed as Most Probably Number (MPN), were performed using the Colilert® kit (IDDEX Laboratories, Spain) following the ISO 9308-2:2012 standard on the same sampling day. Samples were kept for further analyses at -20°C and thawed for 12 h at approximately 20 °C before processing. After thawing, 200 mL of each sample were inoculated with 7 log PCRU/L of mengovirus (MgV) vMC$_0$ (CECT 100000), used as a process control. Samples were processed using the aluminum-based precipitation protocol described elsewhere (AAVV, 2018; Randazzo et al., 2019). Briefly, 200 mL of sample was adjusted to pH 6.0. The $Al(OH)_3$ precipitate was performed mixing 1 part of $AlCl_3$ 0.9N per 100 parts of sample and the solution was mixed at 150 rpm for 15 min. Then, samples were centrifuged at 1,700 x *g* for 20 min and the pellet was resuspended in 10 mL of 3% beef extract (pH 7.4) and shaken at room temperature (RT) for 10 min at 150 rpm. Finally, samples were centrifuged for 30 min at 1,900 x *g* and the resulting pellet was resuspended in 1 mL phosphate saline buffer (PBS, pH 7.4) and stored at -80°C.

2.2. Sample processing for metagenomics

Viral particles were enriched from sample concentrates (n=8) following the NetoVIR protocol, which includes both filtration and nuclease digestion steps (Conceição-Neto et al., 2015). In brief, 500 µL of concentrates were homogenized using the MP FastPrep24 5G (MP Biomedicals, Spain) for 40 seconds at a speed of 6.0. The homogenate was centrifuged at 16,000 × *g* for 3 min and 200 µL of the supernatant was filtered through a 0.8 µm PES filter (Sartorius, UK) to remove large particles. The filtrate was incubated with benzonase (Millipore, Spain) and micrococcal nuclease (New England Biolabs, USA) enzymes at 37°C for 2 h to degrade free nucleic acids. Capsid protected viral nucleic acids were extracted with the NucleoSpin®RNA virus kit (Macherey-Nagel GmbH & Co., Germany), according to the manufacturer's instructions, without adding carrier RNA. Thus, both DNA and RNA viral nucleic acids were concomitantly extracted. Nucleic acids were eluted in 50 µL RNase-free water. Libraries were generated from 1 to 50 ng of a DNA-RNA sample using two different library preparation kits. The first library preparation kit was the ScriptSeq v2 RNA-Seq Library

Preparation Kit (Illumina, USA), referenced as LS, with slight modifications. An initial denaturation step (95 °C for 5 min) was added to the protocol, and PCR cycles were increased to 20 to obtain enough library concentration to sequence. Additionally, the RT enzyme from the original library preparation kit was substituted by the AMV Reverse Transcriptase (Promega, Spain). The second library preparation kit was the NEBNext® Ultra™ II RNA Library Prep Kit (New England BioLabs Inc., Ipswich, UK) (referenced as NB) following manufacturer's instructions. The two libraries compared in this study differ in terms of fragmentation times, enzymes, cDNA synthesis conditions, primers used in the PCR, as well as the conditions for aforesaid amplification (Table S2). Libraries were normalized, pooled, and sequenced using the NextSeq™ 500 platform (Illumina), following the manufacturer's protocol, with a configuration of 150 cycles paired-end reads. Sequencing was performed by Lifesequencing S.L. (Valencia, Spain).

2.3. Data analyses

Obtained reads were cleaned for adaptor removal using cutadapt software (Martin, 2011) with a minimum overlap of 5 nucleotides between read and adaptor and a maximum error rate of 0.1. Reads were cleaned with the *reformat.sh* script from BBMap software (sourceforge.net/projects/bbmap/) in order to remove nucleotides from both ends with Phred scores lower than 20 and reads shorter than 50 bp. Cleaned reads were merged in to single reads with FLASH v1.2.11 (Magoč and Salzberg, 2011) allowing outies. Additionally, cleaned reads were assembled with Ray 2.3.1. (Boisvert et al., 2012) using 31-mers.

Merged reads and contigs were taxonomically annotated using BLASTn algorithm (Boratyn et al., 2013) with a manually curated *in-house* database constructed with all the viral sequences (NCBI:txid10239; release May, 5 2020) available at GenBank (https://www.ncbi.nlm.nih.gov/nuccore/?term=viruses%5Borganism%5D). For the

BLASTn analysis of viral reads against this curated *in-house* database, a cut off of 70% of query sequence coverage and 80% of identity was used, respectively. Rarefaction curves and diversity indexes Shannon and Simpson were calculated with R package *vegan* v2.5-6.

2.4. Virus quantification

For virus quantification an optimized viability RT-qPCR was applied as previously described (Randazzo et al., 2019). In brief, 150 µL sample concentrates were added to 50 µM PMAxx (Biotium, USA) and 0.5% Triton 100-X (Thermo Fisher Scientific, Spain) and incubated in the dark at RT for 10 min at 150 rpm. Then, samples were exposed to photo-activation using a photo-activation system (Led-Active Blue, GenIUL, Spain) for 15 min. RNA was extracted using the NucleoSpin® RNA virus kit (Macherey-Nagel GmbH & Co.) according to the manufacturer's instructions including the Plant RNA Isolation Aid (Ambion, Spain) pretreatment. Primers, probes and RT-qPCR conditions for norovirus GI, norovirus GII, RV, HAV, HEV, mengovirus and HAstV quantification have been previously reported (Randazzo et al., 2019, Cuevas-Ferrando et al., 2020).

For crAssphage quantification by qPCR, the primer set CPQ_064 described by Stachler et al. (2017) was used. PCR conditions were an initial denaturation step of 30 seconds at 95˚C followed by 45 cycles of 5s at 95˚C and 30s at 60˚C. The Premix Ex Taq master mix for probe-based real-time PCR kit (Takara, France) was used for the reaction. For the crAssphage quantification, the standard curve was performed with a customized gBlock® fragment (Integrated DNA Technologies, Spain) of 228 bp that contained the crAssphage sequences used for amplification.

Limit of quantification, qPCR efficiency and standard curve R^2 values for all the tested genes are displayed in Table S3. For all RT-qPCR assays, undiluted and ten-fold diluted RNA was tested to check for RT-qPCR inhibitors.

2.5. Correlation and similarity analyses

Correlation analyses were carried out between data sets obtained by both libraries at family level, and between metagenomics and RT-qPCR results using the R package *Hmisc* v4.2-0 (https://CRAN.R-project.org/package=Hmisc) and applying the Spearman method (ρ). Significance was set at 0.05. Representation of correlation matrix values was performed with the R library *corrplot* v0.84 (https://CRAN.R-project.org/package=corrplot).

For each individual sample, the Jaccard index was used to analyze the similarity among results obtained with both libraries. Calculations were performed using R package *betapart* v1.5.2 (Baselga, 2010) taking into account the beta.JAC values representing the overall beta diversity for each sample pair.

3. Results

3.1. Overview of bias due to library preparation

Each concentrated sample was sequenced using two sequencing libraries: the ScriptSeq v2 RNA-Seq Library Preparation Kit (LS) and the NEBNext® Ultra™ II RNA Library Prep Kit (NB). The average number of reads was 3.2 and 11.5 million for LS and NB libraries, respectively. Rarefaction analyses showed that 5 out of 8 samples sequenced by the LS library reached the plateau, while 2 out of 8 samples sequenced by NB library reached it. Despite that, remaining samples were close to stabilization with both libraries (Fig. S1). Merged viral reads were annotated through a BLASTn comparison with the curated *in-house* database that comprised all the viral sequences (CDS and complete genomes) available at GenBank. For the LS library, the percentage of viral reads ranged from 0.6% to 2.4% in influent and from 0.4% to 4.4% in effluent samples. For NB library, the BLASTn analysis showed a high number of sequences ascribed to the same taxon, suggesting an overrepresentation due to sequencing bias, representing between 33 and 60% of the total viral reads. For that reason, the relative calculations of subsequent analyses were made also taking into account this overrepresentation. These corrected calculations will be called NB-corrected. For the NB library, viral reads ranged from 38% to 58% in influent and from 14% to 24% in effluent samples. Taking into account the results of NB-corrected, these percentages ranged from 9% to 12% and from 7% to 12% in influent and effluent samples, respectively. Shannon and Simpson diversity indexes were calculated for each type of sample (influent or effluent) and for each library (LS or NB) (Fig. 1). Shannon indexes were higher in influent samples sequenced with LS library (mean values of 3.85±0.33 for LS and 1.30±0.12 for NB); however, for effluent samples both indexes showed similar means (1.81±1.21 for LS and 1.90±0.2 for NB), being effluent samples sequenced with LS library more variable (0.36-3.07). Similar results were obtained for Simpson index, even though the mean values for influent samples were highly different (0.93±0.02 for LS and 0.62±0.05 for NB).

Raw data was deposited at SRA under the Bioproject PRJNA67378 with the following accession numbers: SAMN16633937-SAMN16633944 for ScriptSeq v2 RNA-Seq Library Preparation Kit samples and SAMN16634071-SAMN16634078 for NEBNext® Ultra™ II RNA Library Prep Kit samples.

3.2. Mengovirus recovery

Mengovirus (MgV) was used as a process control to analyze the performance of each library to recover reads and the entire genome of MgV. Its recovery, represented as the percentage of viral reads and the percentage of MgV isolate M genome (L22089.1) obtained for each sample with each library, was different depending on the library used. For LS library, the percentage of viral reads of MgV ranged from 0.05% to 0.79% in influent and from 0.35% to 3.68% in effluent samples. For NB libraries, these values ranged from 0.01% to 0.16% in influent samples, and from 0.63% to 5.77% in effluent samples. However, the percentage of MgV reads with the NB-corrected values were higher in effluent samples, ranging from 1.38% to 11.38%. For the analysis of the recovery of MgV genome, assembled contigs belonging to this species were compared with the genome of Mengovirus isolate M (L22089.1). LS library genome recovery ranges from 6.0% to 95.1%. The highest recovery was obtained in the sample IW3. On the other hand, the coverage of this genome by the NB library ranged from 98.4% to 100% (EW3).

3.3. Virome comparison

Regarding the virome composition for each library at family levels, results showed high differences between the two approaches (Fig. 2). While the most represented families with the LS library were *Virgaviridae*, *Microviridae* and *Siphoviridae*; the most abundant families with the NB library were *Ackermannviridae* and *Helleviridae*. These differences allowed the detection of some viral families depending on the library used. For example, families as *Rhabdoviridae*, *Pospiviroidae* or *Mitoviridae* were only detected when the NB library was used for sequencing. Also the taxa uncultured human fecal virus (NCBI:txid239364) and uncultured marine virus (NCBI:txid186617) were only detected with the NB approach. Regarding the families detected with both libraries, only *Genomoviridae* showed total correlation (ρ=1) between values obtained with both libraries in influent and effluent samples. For influent wastewater samples, families *Nairoviridae* and *Virgaviridae* showed total correlations; however, this correlation was only observed for *Parvoviridae* in effluent samples. Other families showed high correlations (ρ=0.8) in influent wastewaters, as *Peribunyaviridae* and *Picornaviridae*; while *Podoviridae*, *Poxviridae*, *Reoviridae* and *Virgaviridae* families showed high correlations (ρ=0.8) in effluent samples. Jaccard indexes showed similarities between the same sample sequenced with each library that ranged from 0.76 (IW1) to 0.91 (IW2), with mean values of 0.83±0.09 for IWs and 0.81±0.04 for EWs.

3.4. Analyses of viral fecal indicators by NGS and correlation with enteric viruses detected by RT-qPCR

Each of the libraries used in this study showed different power for detecting fecal indicators. Similarly, influent and effluent samples showed different detections rates (Fig. 3A). For example, LS library detected CrAssphage with read percentages higher than 1% but these percentages decreased to less than 0.01% with NB library. Most importantly, LS library was unable to detect the fecal indicator adenovirus in effluent wastewaters, while the NB library detected adenoviruses in percentages between 0.1-1%. The same scenario was observed for the Picobirnavirus indicator. However, indicators families as *Inoviridae*, *Microviridae*, *Myoviridae* and *Podoviridae* showed a better detection with the LS library. *Siphoviridae* family detection did not show differences in its detection capacity between the two different tested libraries, with the exception of sample EW3 (Fig. 3A).

Correlation between the number of reads of proposed viral indicators obtained with both libraries and the quantifications obtained by RT-qPCR for enteric viruses along with the *E. coli* counts were calculated. Figure 3B shows the Spearman values of correlation (ρ) calculated with 95% confidence. Norovirus GI and GII showed high correlation values (ρ>0.8) with the indicators crAssphage, Picobirnavirus and *Inoviridae*. The highest correlation values (ρ=0.7) between RV and indicators reads were crAssphage, *Inoviridae* and *Microviridae*. For HAstrV, high correlations (ρ>0.8) only occurred with crAssphage and *Myoviridae*. HAV and HEV did not correlate with any of the indicators. Interestingly, the proposed indicator AdV showed negative correlations with all the enteric viruses analyzed (ρ>-0.7), with the exception of HAV and HEV. Results obtained by NGS for the pepper mild mottle virus (PMMoV), also proposed as viral indicator, showed no correlation with any of the enteric viruses.

3.5. **Virome comparison between influent and effluent wastewaters**

Clean reads obtained from each library and each sample were assembled and contigs longer than 200 bp were used for taxonomical classification by using BLASTn algorithm and the *in-house* database. Due to the different results observed at reads level for each library, contigs classification obtained with both libraries for each sample were merged for results representation and virome analysis. The relative abundances of different taxa are shown in Fig. 4. As observed in the heatmap graphic, the most abundant viruses were bacteriophages, as Dickeya phage or Listeria phage WIL-3, even with higher percentages in effluent samples. The higher detection of these phages in treated samples, as occurred with other species (i.e. cucumber green mottle mosaic virus, EBPR podovirus 2, PMMoV, Stealth virus 1, or Tobacco and tomato mosaic viruses), can be due to the decrease of other viruses after wastewater treatment that allows the its detection. Similarly, this effect could be the responsible of the detection of some viruses in effluent samples that were not detected in influent samples, as the case of human adenovirus and human gammaherpesvirus. Some viruses were only found in high percentages in influent samples, as the indicator crAssphage, some Aeromonas phages, Escherichia phages or viruses belonging to the *Microviridae* family. Wastewater treatments could be the factor that produce this decrease; however, more studies along time from the same WWTPs must be performed in order to ensure the effect of performed treatments.

4. Discussion

The virome of wastewaters have been previously characterized from samples collected around the world (Adriaenssens et al., 2018; Aw et al., 2014; Cantalupo et al., 2011; Fernandez-Cassi et al., 2018; Furtak et al., 2016; Nieuwenhuijse et al., 2020; Rusiñol et al., 2020; Strubbia et al., 2019a, 2019b; Wang et al., 2018); however, much less is known about the virome of effluent samples as only one study has analyzed two effluent samples collected in the UK (Adriaenssens et al., 2018). As far as we know, this is the first study that concomitantly analyzes the RNA and DNA viruses present in influent and effluent samples besides providing a comparison of viruses profiles detected using different sequencing library kits.

Results obtained in our study showed high differences regarding not only viruses, but also the power of detection of viral fecal indicators. Both aspects are important for the use of random metagenomics as tool for specific detection. Our results evidenced the influence of the library used for virome studies together with their variability. Additionally, by using MgV as process control for both metagenomic and RT-qPCR analyses, we further assessed the sensitivity of each library, being higher when using NB library. Recoveries of MgV complete genome were between 6.0 and 95.1% for LS library and between 98.4 and 100% for NB library. In contrast, in a recent study, MgV reads were not recovered from spiked water and sediment samples (Adriaenssens et al., 2020). According to the authors, this was likely due to an inclusion of an inactivation step of the DNase at 75°C, which potentially exacerbated the effect of the RNase step (Adriaenssens et al., 2018). The use of models of a virus of interest when comparing sequencing libraries can be an excellent tool for the library selection.

For the analysis of the virome of influent and effluent wastewaters, results obtained by both libraries were merged. Phages as crAssphage, *Aeromonas* phages, *Escherichia* phages or viruses belonging to the *Microviridae* family were found in high percentages in influent wastewaters. The absence of these viruses in effluent samples can be due to the sanitation treatments applied in WWTPs, even though further analysis that includes a wider sampling design needs to be performed. These results are in line with previous studies showing a high abundance of bacteriophages families (Aw et al., 2014; Cantalupo et al., 2011; Fernandez-Cassi et al., 2018; Rusiñol et al., 2020; Wang et al., 2018) in influent sewage samples. Nevertheless, other studies showed *Virgaviridae* as the most represented viruses (Furtak et al., 2016). Differences in virome profiling with other studies might be due to the influence of library sequencing and the intrinsic characteristics of the virome related to the sample itself and the area of study. On the other hand, the higher presence of some viruses or even its detection only in effluent samples could be produced by the decrease of other viruses that allowed its detection.

The presence of pathogenic viruses is an important aspect for defining the final use of treated waters as it may be the case of irrigation. Due to their high environmental resistance, the presence of human enteric viruses has been reported in treated wastewaters (Adriaenssens et al., 2018). However, some of these pathogenic viruses are not always detected by metagenomics analyses. For instance, in the study by Fernández-Cassi et al. (2018), human adenoviruses (HAdV) reads were not detected in samples concentrated from 10 liters of wastewater. *Adenoviridae* was also not detected in the study of Adrianssens et al. (2018), in which the sample was concentrated from 1 liter of wastewater. In our study, concentrating 200 mL of effluent samples, we were able to detect HAdV in percentages between 0.16% and 0.35%. In contrast, percentages of HAdV in influent wastewaters were lower than 0.01%. Overall, the majority of the annotated virome belonged to bacteriophages. This indicates that metagenomics is poor in sensitivity when used to detect a low abundance of viral pathogens against a large background of bacteriophages, as occurred for the enteric viruses detected by viability RT-qPCR. For example, in the present study, norovirus genomes could not be retrieved from the reads as reported elsewhere (Adriaenssens et al., 2018; Fernández-Cassi et al., 2018; Strubbia et al., 2019b). In the current study, the number of generated paired reads per sample was 3.2 and 11.5 million for LS and NB, respectively; while Adriaenssens et al., (2018) reported between 10 and 110 million, increasing significantly the probability to retrieve full or partial viral genomes. Alternatively, methods to detect and characterize specific viruses have been described and rely on the selection of target RNA prior to library preparation through a capture using VirCapSeq-VERT target enrichment, as reported for norovirus (Strubbia et al., 2019b).

5. Conclusion

The use of metagenomics for virome characterization and its implementation for wastewater analyses has increased in the last years (Nieuwenhuijse and Koopmans, 2017). However, the major problem of this approach is the lack of standardized procedures and the substantial differences among studies; thus, available data must be interpreted with caution. The present study showed a procedure that allows the detection and the characterization of viral populations in untreated and treated wastewater samples. Overall, this study sheds light on the diversity of the viral communities in untreated and treated wastewaters providing valuable information also in terms of viral fecal indicators. The study also evidences the bias on virome profiles obtained by tested sequencing libraries. Our results underline the need for further studies to elucidate the influence of sequencing procedures in virome profiles in wastewater matrices in order to improve the knowledge of the virome in the environment.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Names of specific vendors, manufacturers, or products are included for informational purposes only and does not imply endorsement by Authors or their affiliations.

Acknowledgements

This study was supported by the "VIRIDIANA" project AGL2017-82909 (AEI/FEDER, UE) funded by Spanish Ministry of Science, Innovation and Universities, CPS project 2018CPS10 from the Center for Produce Safety Grant Agreement, and the APOTI grant (APOTIP/2018/007) from the Generalitat Valenciana. WR was supported by a postdoctoral fellowship from Generalitat Valenciana (APOSTD/2018/150). EC-F is recipient of a predoctoral contract from the MICINN, Call 2018. The authors acknowledge to the "Entidad de Saneamiento de Aguas Residuales" (EPSAR) and to Global Omnium S.L. for authorizing the sampling. We gratefully acknowledge Agustín Garrido (IATA-CSIC) for providing technical assistance.

Contributions

GS designed the work. AP-C and EC-F processed the samples. AP-C performed the bioinformatic work and data analysis. AP-C, EC-F, WR and GS wrote the paper. All authors have read and agreed to the published version of the manuscript.

References

AAVV (2018). "Section 9510 D. Virus concentration by aluminum hydroxide adsorption-precipitation, chapter detection of enteric viruses," in Standard Methods for the Examination of Water and Wastewater, 23rd Edn, eds

E. W. Rice, R. B. Baird, and A. D. Eaton (Denver, CO: American Water Works Association)

Adriaenssens, E.M., Farkas, K., Harrison, C., Jones, D.L., Allison, H.E., McCarthy, A.J., 2018. Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. mSystems 3, 1–18. https://doi.org/10.1128/mSystems.00025-18

Aw, T.G., Howe, A., Rose, J.B., 2014. Metagenomic approaches for direct and cell culture evaluation of the virological quality of wastewater. J. Virol. Methods 210, 15–21. https://doi.org/10.1016/j.jviromet.2014.09.017

Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. Global Ecology and Biogeography 19:134-143. https://doi.org/10.1111/j.1466-8238.2009.00490.x

Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., Corbeil, J., 2012. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol. 13, R122. https://doi.org/10.1186/gb-2012-13-12-r122

Boratyn, G.M., Camacho, C., Cooper, P.S., Coulouris, G., Fong, A., Ma, N., Madden, T.L., Matten, W.T., McGinnis, S.D., Merezhuk, Y., Raytselis, Y., Sayers, E.W., Tao, T., Ye, J., Zaretskaya, I., 2013. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 41, W29-33. https://doi.org/10.1093/nar/gkt282

Bivins, A., North, D., Ahmad, A., Ahmed, W., Alm, E., Been, F., Bhattacharya, P., Bijlsma, L., Boehm, A.B., Brown, J., Buttiglieri, G., Calabro, V., Carducci, A., Castiglioni, S., Cetecioglu Gurol, Z., Chakraborty, S., Costa, F., Curcio, S., De Los Reyes, F.L., Delgado Vela, J., Farkas, K., Fernandez-Casi, X., Gerba, C., Gerrity, D., Girones, R., Gonzalez, R., Haramoto, E., Harris, A., Holden, P.A., Islam, M.T., Jones, D.L., Kasprzyk-Hordern, B., Kitajima, M., Kotlarz, N., Kumar, M., Kuroda, K., La Rosa, G., Malpei, F., Mautus, M., McLellan, S.L., Medema, G., Meschke, J.S., Mueller, J., Newton, R.J., Nilsson, D., Noble, R.T., Van Nuijs, A., Peccia, J., Perkins, T.A., Pickering, A.J., Rose, J., Sanchez, G., Smith, A., Stadler, L., Stauber, C., Thomas, K., Van Der Voorn, T., Wigginton, K., Zhu, K., Bibby, K., 2020. Wastewater-Based Epidemiology: Global Collaborative to Maximize Contributions in the Fight against COVID-19. Environ. Sci. Technol. https://doi.org/10.1021/acs.est.0c02388

Cantalupo, P.G., Calgua, B., Zhao, G., Hundesa, A., Wier, A.D., Katz, J.P., Grabe, M., 2011. Raw Sewage Harbors Diverse Viral Populations. MBio 2, e00180-11. https://doi.org/10.1128/mBio.00180-11.Editor

Chalmers, R.M., Robinson, G., Elwin, K., Hadfield, S.J., Thomas, E., Watkins, J., Casemore, D., Kay, D., 2010. Detection of Cryptosporidium species and sources of contamination with Cryptosporidium hominis during a waterborne outbreak in north west Wales. J. Water Health 8, 311–25. https://doi.org/10.2166/wh.2009.185

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C.K., Lavigne, R., Maes, P., Van Ranst, M., Heylen, E., Matthijnssens, J., 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci. Rep. 5, 16532. https://doi.org/10.1038/srep16532

Cuevas-Ferrando, E., Randazzo, W., Pérez-Cataluña, A., Sánchez, G., 2020. HEV Occurrence in Waste and Drinking Water Treatment Plants. Front. Microbiol. 10. https://doi.org/10.3389/fmicb.2019.02937

Eslamian, S. (Ed.), 2016. Urban Water Reuse Handbook. CRC Press. https://doi.org/10.1201/b19646

Farkas, K., Walker, D.I., Adriaenssens, E.M., McDonald, J.E., Hillary, L.S., Malham, S.K., Jones, D.L., 2020. Viral indicators for tracking domestic wastewater contamination in the aquatic environment. Water Res. https://doi.org/10.1016/j.watres.2020.115926

Fernandez-Cassi, X., Timoneda, N., Martínez-Puchol, S., Rusiñol, M., Rodriguez-Manzano, J., Figuerola, N., Bofill-Mas, S., Abril, J.F., Girones, R., 2018. Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. Sci. Total Environ. 618, 870–880. https://doi.org/10.1016/j.scitotenv.2017.08.249

Fong, T.-T., Lipp, E. K. 2005. Enteric viruses of humans and animals in aquatic environments: health risks, detection, and potential water quality assessment tools. MMBR, 69(2), 357–371. https://doi.org/10.1128/MMBR.69.2.357-371.2005

Furtak, V., Roivainen, M., Mirochnichenko, O., Zagorodnyaya, T., Laassri, M., Zaidi, S.Z., Rehman, L., Alam, M.M., Chizhikov, V., Chumakov, K., 2016. Environmental surveillance of viruses by tangential flow filtration and metagenomic reconstruction. Eurosurveillance 21, 30193. https://doi.org/10.2807/1560-7917.ES.2016.21.15.30193

Gerba, C.P., Betancourt, W.Q., Kitajima, M., Rock, C.M., 2018. Reducing uncertainty in estimating virus reduction by advanced water treatment processes. Water Res. 133, 282–288. https://doi.org/10.1016/j.watres.2018.01.044

Gerba, C.P., Kitajima, M., Iker, B., 2013. Viral presence in waste water and sewage and control methods. Viruses Food Water 293–315. https://doi.org/10.1533/9780857098870.3.293

Haramoto, E., Kitajima, M., Hata, A., Torrey, J.R., Masago, Y., Sano, D., Katayama, H., 2018. A review on recent progress in the detection methods and prevalence of human enteric viruses in water. Water Res. 135, 168–186. https://doi.org/10.1016/j.watres.2018.02.004

ISO 9308-2:2012. Water quality -- Enumeration of Escherichia coli and coliform bacteria -- Part 2: Most probable number method.

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. Sci. Total Environ. 488–489, 290–296. https://doi.org/10.1016/J.SCITOTENV.2014.04.087

Lamori, J.G., Xue, J., Rachmadi, A.T., Lopez, G.U., Kitajima, M., Gerba, C.P., Pepper, I.L., Brooks, J.P., Sherchan, S., 2019. Removal of fecal indicator bacteria and antibiotic resistant genes in constructed wetlands. Environ. Sci. Pollut. Res. 26, 10188–10197. https://doi.org/10.1007/s11356-019-04468-9

Magoč, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–63. https://doi.org/10.1093/bioinformatics/btr507

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10. https://doi.org/10.14806/ej.17.1.200

Nieuwenhuijse, D.F., Koopmans, M.P.G., 2017. Metagenomic sequencing for surveillance of food- and waterborne viral diseases. Front. Microbiol. 8, 1–11.

https://doi.org/10.3389/fmicb.2017.00230

Nieuwenhuijse, D.F., Oude Munnink, B.B., Phan, M.V.T., Hendriksen, R.S., Bego, A., Rees, C., Neilson, E.H., Coventry, K., Collignon, P., Allerberger, F., Rahube, T.O., Oliveira, G., Ivanov, I., Sopheak, T., Vuthy, Y., Yost, C.K., Tabo, D. adjim, Cuadros-Orellana, S., Ke, C., Zheng, H., Baisheng, L., Jiao, X., Donado-Godoy, P., Coulibaly, K.J., Hrenovic, J., Jergović, M., Karpíšková, R., Elsborg, B., Legesse, M., Eguale, T., Heikinheimo, A., Villacis, J.E., Sanneh, B., Malania, L., Nitsche, A., Brinkmann, A., Saba, C.K.S., Kocsis, B., Solymosi, N., Thorsteinsdottir, T.R., Hatha, A.M., Alebouyeh, M., Morris, D., O'Connor, L., Cormican, M., Moran-Gilad, J., Battisti, A., Alba, P., Shakenova, Z., Kiiyukia, C., Ng'eno, E., Raka, L., Bērziņš, A., Avsejenko, J., Bartkevics, V., Penny, C., Rajandas, H., Parimannan, S., Haber, M.V., Pal, P., Schmitt, H., van Passel, M., van de Schans, M.G.M., Zuidema, T., Jeunen, G.J., Gemmell, N., Fashae, K., Wester, A.L., Holmstad, R., Hasan, R., Shakoor, S., Rojas, M.L.Z., Wasyl, D., Bosevska, G., Kochubovski, M., Radu, C., Gassama†, A., Radosavljevic, V., Tay, M.Y.F., Zuniga-Montanez, R., Wuertz, S., Gavačová, D., Trkov, M., Keddy, K., Esterhuyse, K., Cerdà-Cuéllar, M., Pathirage, S., Larsson, D.G.J., Norrgren, L., Örn, S., Van der Heijden, T., Kumburu, H.H., de RodaHusman, A.M., Njanpop-Lafourcade, B.M., Bidjada, P., Nikiema-Pessinaba, S.C., Levent, B., Meschke, J.S., Beck, N.K., Van Dang, C., Tran, D.M.N., Do Phuc, N., Kwenda, G., Munk, P., Venkatakrishnan, S., Aarestrup, F.M., Cotten, M., Koopmans, M.P.G., 2020. Setting a baseline for global urban virome surveillance in sewage. Sci. Rep. 10, 1–13. https://doi.org/10.1038/s41598-020-69869-0

Polo, D.D., Quintela-Baluja, M., Corbishley, A., Jones, D.L., Singer, A.C., Graham, D.W., Romalde, P.J.L., 2020. Making waves: Wastewater-based epidemiology for

SARS-CoV-2 – Developing robust approaches for surveillance and prediction is harder than it looks. Water Res. 186, 116404. https://doi.org/10.1016/j.watres.2020.116404

Randazzo, W., Piqueras, J., Evtoski, Z., Sastre, G., Sancho, R., Gonzalez, C., Sánchez, G., 2019. Interlaboratory Comparative Study to Detect Potentially Infectious Human Enteric Viruses in Influent and Effluent Waters. Food Environ. Virol. https://doi.org/10.1007/s12560-019-09392-22

Rusiñol, M., Martínez-Puchol, S., Timoneda, N., Fernández-Cassi, X., Pérez-Cataluña, A., Fernández-Bravo, A., Moreno-Mesonero, L., Moreno, Y., Alonso, J.L., Figueras, M.J., Abril, J.F., Bofill-Mas, S., Girones, R., 2020. Metagenomic analysis of viruses, bacteria and protozoa in irrigation water. Int. J. Hyg. Environ. Health 224, 113440. https://doi.org/10.1016/j.ijheh.2019.113440

Sano, D., Amarasiri, M., Hata, A., Watanabe, T., Katayama, H., 2016. Risk management of viral infectious diseases in wastewater reclamation and reuse: Review. Environ. Int. 91, 220–229. https://doi.org/10.1016/J.ENVINT.2016.03.001

Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K., Shanks, O.C., 2017. Quantitative CrAssphage PCR Assays for Human Fecal Pollution Measurement. Environ. Sci. Technol. 51, 9146–9154. https://doi.org/10.1021/acs.est.7b02703

Strubbia, S., Phan, M.V.T., Schaeffer, J., Koopmans, M., Cotten, M., Le Guyader, F.S., 2019a. Characterization of Norovirus and Other Human Enteric Viruses in Sewage and Stool Samples Through Next-Generation Sequencing. Food Environ. Virol. https://doi.org/10.1007/s12560-019-09402-3

Strubbia, S., Schaeffer, J., Oude Munnink, B.B., Besnard, A., Phan, M.V.T., Nieuwenhuijse, D.F., de Graaf, M., Schapendonk, C.M.E., Wacrenier, C., Cotten, M., Koopmans, M.P.G., Le Guyader, F.S., 2019b. Metavirome Sequencing to Evaluate Norovirus Diversity in Sewage and Related Bioaccumulated Oysters. Front. Microbiol. 10, 2394. https://doi.org/10.3389/fmicb.2019.02394

Wang, Y., Jiang, X., Liu, L., Li, B., Zhang, T., 2018. High-Resolution Temporal and Spatial Patterns of Virome in Wastewater Treatment Systems. Environ. Sci. Technol. 52, 10337–10346. https://doi.org/10.1021/acs.est.8b03446

Figure legends

Figure 1: Shannon and Simpson diversity indexes for viral species in influent (IW) and effluent (EW) samples processed by using ScriptSeq v2 RNA-Seq Library Preparation kit (LS) and NEBNext® Ultra™ II RNA Library Prep Kit libraries (NB) for metagenomics characterization.

Figure 2. Relative abundance at family level of the viral population detected in influent and effluent samples from four different WWTPs by metagenomics with ScriptSeq v2 RNA-Seq Library Preparation kit (LS, green) and NEBNext® Ultra™ II RNA Library Prep Kit (NB, orange).

Figure 3. Viral indicators analysis in influent (IW) and effluent (EW) samples. Panel A, Detection of viral indicators with ScriptSeq v2 RNA-Seq Library Preparation kit (LS) and NEBNext® Ultra™ II RNA Library Prep Kit (NB and NB-corrected). Panel B, Correlation matrix between the reads of viral indicators obtained by NGS and the load of enteric viruses (RT-qPCR) and E. coli counts.

Figure 4. Heatmap showing the virome composition at species level obtained by merging the results of ScriptSeq v2 RNA-Seq Library Preparation kit (LS) and NEBNext® Ultra™ II RNA Library Prep Kit (NB). Only species with percentages higher than 1% are shown.