Tracing the fate of wastewater viruses reveals catchment-scale virome diversity and connectivity
Adriaenssens, Evelien M.; Farkas, Kata; McDonald, James E.; Jones, David L.; Allison, Heather E.; McCarthy, Alan J.

Water Research

DOI: 10.1016/j.watres.2021.117568
Published: 15/09/2021

Publisher's PDF, also known as Version of record

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):
Adriaenssens, E. M., Farkas, K., McDonald, J. E., Jones, D. L., Allison, H. E., & McCarthy, A. J. (2021). Tracing the fate of wastewater viruses reveals catchment-scale virome diversity and connectivity. Water Research, 203. https://doi.org/10.1016/j.watres.2021.117568

Hawliau Cyffredinol / General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Tracing the fate of wastewater viruses reveals catchment-scale virome diversity and connectivity

Evelien M. Adriaenssens a,b,*, Kata Farkas c,d, James E. McDonald c, David L. Jones c,e, Heather E. Allison a, Alan J. McCarthy a

a Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK
b Quadram Institute Bioscience, Norwich, NR4 7UQ, UK
c School of Natural Sciences, Bangor University, Bangor, LL57 2UW, UK
d School of Ocean Sciences, Bangor University, Bangor, LL59 5AB, UK
e UWA School of Agriculture and Environment, The University of Western Australia, Perth, WA 6009, Australia

ARTICLE INFO

Keywords: Viromics viral diversity wastewater viruses aquatic viruses shellfish viruses wastewater contamination virus ecology

ABSTRACT

The discharge of wastewater-derived viruses in aquatic environments impacts catchment-scale virome composition. To explore this, we used viromic analysis of RNA and DNA virus-like particles to holistically track virus communities entering and leaving wastewater treatment plants and the connecting river catchment system and estuary. We reconstructed >40,000 partial viral genomes into 10,149 species-level groups, dominated by dsDNA and (+)-ssRNA bacteriophages (Caudoviricetes and Leviviricetes) and a small number of genomes that could pose a risk to human health. We found substantial viral diversity and geographically distinct virus communities associated with different wastewater treatment plants. River and estuarine water bodies harboured more diverse viral communities in downstream locations, influenced by tidal movement and proximity to wastewater treatment plants. Shellfish and beach sand were enriched in viral communities when compared with the surrounding water, acting as entrapment matrices for virus particles. Extensive phylogenetic analyses of environmental-derived and reference sequences showed the presence of human-associated sapovirus GII in all sample types, multiple rotavirus A strains in wastewater and a diverse set of picorna-like viruses associated with shellfish. We conclude that wastewater-derived viral genetic material is commonly deposited in the environment and can be traced throughout the freshwater-marine continuum of the river catchment, where it is influenced by local geography, weather events and tidal effects. Our data illustrate the utility of viromic analyses for wastewater- and environment-based ecology and epidemiology, and we present a conceptual model for the circulation of all types of viruses in a freshwater catchment.

1. Introduction

Viruses are the most abundant biological entities in terrestrial and aquatic biomes, but their origin, distribution, and potential to spread disease via watercourses is poorly understood (Roux et al., 2020). Previous research has demonstrated that wastewater contains a plethora of viruses, including human-pathogenic and zoonotic viruses, and that wastewater treatment processes do not remove human viruses with sufficient efficacy (Da Silva et al., 2007; Farkas et al., 2018b; Fong et al., 2010; Girones et al., 2010; Gomes et al., 2019; Gulino et al., 2020; Hellmér et al., 2014; Kitajima et al., 2014; Prado et al., 2019; Qiu et al., 2015; Sidhu et al., 2017). Viral abundance, behaviour, infectivity and fate remain poorly understood because of knowledge gaps in the ecology and connectivity of viromes across human populations and the freshwater-marine continuum.

The current gold standard method for investigating enteric viruses in the environment is q(RT)-PCR, a technique that provides reliable quantitative information on the presence of the genomic material of target viruses, but requires prior knowledge on the identity of the virus and its genome sequence (Farkas et al., 2020, 2017b). As qPCR-based assays only detect a fragment of the genome, the question of virus integrity, and hence infectivity, remains open. Infectivity assays can offer a solution, but even where available, require specialised cultivation systems and are not likely to become generally applicable for
routine monitoring of public health risks (DiCaprio, 2017). As a more comprehensive and now potentially feasible alternative, we applied shotgun viromics, i.e. next-generation sequencing of the entire aquatic virome, to reconstruct full virus genomes from the environment and objectively scrutinise the ecological and health implications of viral diversity and geographical distribution, with minimal bias.

Virome analyses are transforming our understanding of viral diversity and function in the biosphere (Emerson et al., 2018; Gregory et al., 2020; Roux et al., 2016) and provide unprecedented opportunity to understand the connectivity and fate of human-derived viruses at the catchment scale. Here, we present the first integrated analysis of the full virome of a river catchment system and estuary including water, sediment, wastewater treatment plants, beaches and shellfish production areas (Fig. 1 and Supplementary Table 1). We assembled over 40,000 partial or near-complete genomes (UViGs Uncultivated Virus Genomes, Roux et al., 2019)) of ssRNA, dsRNA, ssDNA and dsDNA viruses, clustered into 10 149 species-level groupings (vOTUs, viral Operational Taxonomic Units). Our detailed bioinformatic analysis of the RNA and DNA virome, to reconstruct full virus genomes from the environment and enrichment, dilution and degradation, and the implications for human health.

2. Materials and Methods

2.1. Sample collection

This work builds on our pilot study of a single wastewater treatment plant, and a downstream water and sediment sampling site, in which we optimised methods and showed that we could reconstruct RNA virus genomes from environmental samples (Adriaenssens et al., 2018). We collected and processed four different types of samples for this study: wastewater (influent and effluent), surface water (river and estuarine), sediment and shellfish in June 2017 from the Conwy river catchment area located in North Wales (UK) (Fig. 1, Supplementary Table 1). Wastewater influent was collected from the four major treatment plants in the catchment and the corresponding effluent from three (the Galon plant effluent pipe exits directly into the open sea and therefore was not sampled separately); one litre per sample. Surface water was collected in four biological replicates of 50 L, resulting in two replicates per library type (RNA and DNA–based). At two locations, one in the river and one at a major recreational beach, sediment samples were taken in 4 biological replicates of at least 50 g, two replicates per library type. At low tide, samples were scooped from an accessible part of the river bank into sterile bags using a sterilised spade. Finally, mussels (Mytilus edulis) were collected from the two main commercial shellfishery locations in the estuary and divided into eight pseudoreplicates, as described below. Samples were collected at low tides, enabling the collection of shellfish samples from shore without a boat.

2.2. Sample processing

The wastewater (1 L) and surface water (50 L) samples were concentrated using a two-step protocol involving tangential flow ultrafiltration (TFUF) and beef extract elution as described in detail previously (Farkas et al., 2018c). Briefly, sample volumes were reduced to 50 mL by TFUF on a KrosFlo® Research IIi Tangential Flow Filtration System (Spectrum Labs, Phoenix, AZ, USA, Cat. no. SYR-U20-01N) using a 100 kDa cut-off mPES MiniKros® hollow fibre filter (Spectrum Labs). Virus particles were eluted using beef extract and NaNO₃ to a final concentration of 3% and 2 M (pH 5.5), respectively. After the solution’s pH was adjusted to 7.5, PEG 6000 was added to a final concentration of 15% with 2% NaCl, incubated overnight at 4°C and after centrifugation (30 min, 10 000 x g, 4°C) the pellet was resuspended in 10-15 ml PBS (pH 7.4). These suspensions were kept at ~80°C until nucleic acid extraction. The sediment samples were processed using beef extract elution and PEG 6000 precipitation as above and described previously (Farkas et al., 2017a).

For the mussel samples, approximately 200 mussels were collected from each location and stored on ice. Each mussel was dissected and the digestive tissue extracted and minced with a scalp. The tissue was...
pooled per location and then divided into four pseudoreplicates. Two replicates per location were mixed with SM buffer (0.1 M NaCl, 50 mM Tris/HCl pH 7.4, 10 mM MgSO4) and two with PBS at 25 g of digestive tissue to 20 mL of buffer. The samples were then shaken for 30 minutes (150 rpm) at room temperature to dissociate viral particles from the digestive tissue after which they were stored at -80°C. Due to a limit on the number of shellfish collected, we used both PBS and SM buffer as we could not perform a pre-optimisation of the protocol. No systematic differences between the two methods were detected.

2.3. RNA extraction

Wastewater, surface water and sediment concentrates were processed as follows. The concentrates were diluted in an equal volume 0.5 M NaCl to improve dissociation of viral particles before filtration. After centrifugation (5 min, 3200 x g) the supernatant was filtered through a 0.2 µm sterile syringe filter (Millipore). The filtrate was further concentrated using Vivaspin 20 spin filters (100 kDa) and centrifugation at 3200 x g. Once the volume was below 1 mL, 5 mL Tris buffer (5 mM Tris/HCl, 5 mM MgSO4, 75 mM NaCl, pH 7.5) was added and the volume reduced (two times) to reduce the NaCl content of the virus suspension. Centrifugation times ranged between 150 minutes and 20 hours to reduce the volume below 500 µL for the next step. A DNase treatment with 10 U Turbo DNase (Invitrogen) was performed to remove extracellular DNA (incubation at 37°C for 30 min, inactivation at 75°C for 10 min). Mussel samples were highly viscous and required separate processing, as we were unable to filter or concentrate with the Vivaspin filters. Instead, 2 × 1 mL aliquots per replicate were mixed with 0.1 mm glass beads (MoBio) and lysed in a PowerLyser (MoBio) shaker (2 × 30 seconds at 3400 rpm). Debris was removed by centrifugation (5 min at 3200 × g) and the supernatant was stored at -20°C for next-day processing.

For all sample types, the viral capsids were lysed using a combination of proteinase K (50 µg for clear samples, 100 µg for turbid samples), EDTA (0.5 M final concentration) and SDS (0.5% final concentration), and incubation for one hour at 56°C. Next, the RNA was extracted by TRIzol extraction derived from Kroger and colleagues (Kroger et al., 2012). In short, 500 µL of sample was mixed with 1 mL of TRIzol reagent (Invitrogen), shaken vigorously and centrifuged for 15 minutes at 13 000 x g. The aqueous phase was removed and transferred to a new tube. The phase separation was repeated for samples that remained turbid. The nucleic acid was recovered by isopropanol precipitation and resuspended in 50 µL of sterile, RNase-free water. Viral DNA was removed with an additional DNase step, adding 4 U Turbo DNase, 5 µL TD buffer, and incubating for 40 minutes at 37°C followed by inactivation of the DNase at 75°C for 10 minutes. The DNase was removed by a second isopropanol precipitation as above, the RNA resuspended in 50 µL of RNase-free water and stored at -80°C until sequencing. Alongside all samples, a positive extraction control comprising of Salmonella cells (Salmonella enterica subsp enterica serovar Typhimurium strain D23580, RefSeq acc NC_016854) and the process-control virus mengovirus (~10^9 particles/mL) was extracted, as was a negative Tris buffer control.

2.4. DNA extraction

Wastewater, surface water and sediment samples were processed similarly as for RNA extraction with a few amendments to the extraction process. The samples were diluted in 10 mL NaCl (0.5 M). For surface water and sediment samples one replicate (designated a) was treated with chloroform (1 mL) to lyse the cellular fraction (15 min incubation with gentle shaking) and the cellular debris removed by centrifugation (5 min, 3200 x g). The second replicate (b) was filtered through a 0.45 µm sterile syringe filter (Millipore). For the wastewater samples which consisted of only one replicate, the sample was split in two, half treated with chloroform and half filtered, and then merged. All samples were then concentrated and desalted as described above (using Vivaspin 20 spin filters (100 kDa) and centrifugation at 3200 x g, with centrifugation time between 100 min and 20h). All sample concentrates (approx. 500 µL each) were treated with 10 U of Turbo DNase (Invitrogen) and 10 µg of RNase A (Thermo Fisher Scientific) supplemented with Turbo DNase buffer for 30 min at 37°C and inactivation at 65°C for 10 min. The separation of filtering and chloroform treatment before proteinase and nuclease treatment theoretically allows for the recovery of both giant viruses (larger than the filter pore size) and viruses with lipid membranes, while simultaneously reducing the background cellular DNA. No differences between treatments were found for potentially pathogenic viruses, however, the systematic comparison of methods for all viruses is beyond the scope of this study and we encourage others to use our data for further exploration.

Mussel digestive tissue was processed exactly as during RNA extraction (mixed with 0.1 mm glass beads and lysed in PowerLyser) and no nuclease treatment was performed. From this point, all samples were extracted in the same manner. Capsids were lysed by adding proteinase K (50 µg/mL final concentration), EDTA (20 mM final concentration) and SDS (0.5% final concentration), followed by incubation at 56°C for one hour and extracted using phenol/chloroform/isooamylicol (25:24:1) in PhaseLock tubes (VWR). The aqueous phase was transferred to a new tube and the process was repeated at least once (twice for turbid samples), followed by one round of chloroform phase separation. Finally, samples were further cleaned and concentrated with ethanol precipitation (2.5 x volume 100% ethanol; 1/10 volume 3 M NaAc pH 5; incubation at -20°C for 30 minutes; precipitation 30 min at 15 000 x g, 4°C), washed with 70% ethanol and air-dried in a laminar flow cabinet.

In tandem with the whole process, control samples were extracted, starting with the dilution in NaCl. We used a negative control consisting of Tris buffer and a positive control consisting of 500 µl stationary culture Escherichia coli MG1655 cells (RefSeq acc NC_000913), 2.2 × 10^8 pfu of Escherichia phage T5 (RefSeq acc NC_005859) and 1.3 × 10^9 pfu of Escherichia phage vB_EcoP_phi24B (GenBank acc HM208303).

2.5. Sequencing

Sequence library preparation and sequencing was performed by the Centre for Genomics Research (CGR) NBaf facilities at the University of Liverpool, UK. RNA libraries were prepared as in the pilot study (Adriaenssens et al., 2018) using the NEBNext Ultra directional RNA library preparation kit of Illumina with dual indexes. During library preparation, the number of PCR cycles was increased to 30 to account for the low amounts of input RNA (< 1 ng). Dual-indexed DNA libraries were generated using the NEBNext Ultra II DNA Library Prep kit according to the manufacturer’s instructions. Libraries were pooled and sequenced on six lanes of the Illumina HiSeq 4000 generating paired-end 2 x 150 bp reads, three lanes for the RNA libraries in July 2017 and three lanes for the DNA libraries in March 2018.

The DNA sequencing run failed for libraries DNA_SW2a, DNA_TyCa/b, DNA_Sba/b and unfortunately, we were unable to reconstruct the libraries as the samples had been mistakenly stored at 4°C and the DNA had degraded. Furthermore, the read lengths obtained for the mussel DNA libraries were much lower as for all other libraries, as the DNA had been excessively sheared during the extraction procedure.

2.6. In silico processing

Reads went through an initial round of quality control at CGR to remove Illumina adapters (Cutadapt version 1.2.1, -O 3) and were trimmed with Sickle (version 1.2) removing all reads below an average quality of 20 and shorter than 20 bp (Joshi and Fass, 2011; Martin, 2011). The resulting fastq files were received as raw read files from the CGR and deposited into SRA under BioProject PRJNA509142, accession numbers SRR8299359 to SRR8299398.
The paired-end read files were further trimmed and filtered to increase quality using the prinseq-lite suite (Schmieder and Edwards, 2011) and the read pairs meeting the following criteria were retained: minimum length 35 bases, GC-content between 5 and 95%, maximum 1 N, trimmed until the average read quality was 30. For all exactly duplicated reads only one copy was retained. The reads for the control libraries were merged per library type (RNA & DNA) and used as a bowtie2 mapping reference (Langmead and Salzberg, 2012). Each of the sample libraries was then mapped against its control and only the unmapped reads were retained. These reads were then assembled per sample using SPAdes version 13.9 using the k-mers lengths 21,33,55,77, 95,107,121 (Nurk et al., 2013), with the exception of the mossel DNA libraries containing the shorter reads where the k-mers 21,33,55,77 were used. The control libraries were assembled using the same parameters and compared to the sample contigs using BLASTn (BLAST+, suite), and sample contigs that showed significant similarity (e value < 0.001) were removed from each of the sample contig datasets (Camacho et al., 2009).

From these contigs, an Anvi’o contig database was created according to the instructions of the metagenomics workflow (Eren et al., 2015). To be included in the database, contigs needed to meet the following criteria: RNA library assembles (i) contig length min 1000 nucleotides (nt); (ii) amino acid similarity with any known virus; (iii) recruit no reads from control libraries; DNA library assembles (i) contig length min 10,000 nt, (ii) identified by VirSorter as viral in categories 1 or 2 (Roux et al., 2015), (iii) recruit no reads from control libraries. VirSorter was run on all DNA contig sets using the microbiome decontamination mode on the iVirus Cyvserse infrastructure (Bolduc et al., 2017). The contig dataset comprising 40 000 UVIGs was merged and clustered at an approximation of the viral species level (95% average nucleotide identity over min 80% of contig length), according to the species definition for bacteriophages implemented by the International Committee on Taxonomy of Viruses (ICTV) and conventionally used in virome studies (Adriaenssens and Brister, 2017; Emerson et al., 2018; Gregory et al., 2016; Roux et al., 2019, 2016). We performed a final refinement by removing all contigs < 10 000 nt assembled from RNA libraries that showed amino acid similarity with dsDNA viruses, based on diamond BLASTx comparison (Buchfink et al., 2015) with the nr database downloaded from the NCBI in January 2018. The final database contained 10 149 UVIGs (Uncultivated Viral Genomes, (Roux et al., 2019)) that each represent a viral species-level population. Taxonomic information was added to the contigs database in Anvi’o using Kaiju with the built-in viral database (Menzel et al., 2016). All individual assemblies were also compared with the nr and viral ReSeq protein databases (version Jan 2018) separately in case the length thresholds for contig database creation excluded certain virus types.

To compare the incidence and abundance of UVIGs in the different samples, for each library the reads were mapped to the contigs database using kallisto (Bray et al., 2016). The abundances of contigs within and between samples were assessed by transforming the values into Transcript Per Million values (TPMs) where each contig (UVIG) was considered a transcript using the program tximport in R (Soneson et al., 2016). The resulting 10 149 by 58 matrix was visualised with Phantasus viewer. Sequences missing the conserved structural motifs present in RdRPs (Venkataraman et al., 2018) were removed, as were sequences missing more than 50% of the trimmed sites. Trees were computed using the IQ-Tree suite (Nguyen et al., 2015) including calculation of the best substitution model with ModelFinder (Kalyaanamoorthy et al., 2017), calculation of the approximate likelihood ratio test (1000 repetitions) (Anisimova et al., 2011) and ultrafast bootstrap approximation with UFBOOT2 (1000 repetitions) (Hoang et al., 2018). The resulting trees were analysed and annotated in iTOL (Letunic and Bork, 2007). For the picorna-calici tree, the alignments generated by Shi and colleagues were additionally used as references (Shi et al., 2018, 2016).

3. Results and Discussion

3.1. Viral species richness is environment-specific and geographically distinct

We generated a final species-level contig database containing 10 149 vOTUs from 44 897 assembled contigs, each represented by the longest viral genome (UVIG). We used the number of vOTUs per sample, normalised per volume (ml) or wet weight (g) of input material, as an approximation of species richness (Fig. 1). Normalised viral OTU (“species”) richness was highest in the shellfish digestive tissue and beach sediment samples, intermediate in wastewater influent and effluent and lowest in the surface water samples. The surface river water samples showed a trend of increasing viral richness moving downstream, as further inputs of wastewater from treatment plants and other anthropogenic sources occurred, reaching a plateau around the location of SW3 after which richness remained high (Fig. 1).

We further investigated the differential patterns of abundance of each UVIG in each library by mapping reads of all samples against the vOTU database and visualised the data with Anvi’o (Fig. 2, Supplementary Fig. 1), to identify 13 categories of viral species abundance and composition patterns (Table 1). The wastewater samples contained the most diverse set of UVIGs in absolute numbers, however, each wastewater treatment plant yielded a geographically distinct viral community. The river water samples contained a lower absolute richness of viruses than the wastewater, except for sample SW5 (and to a lesser extent, SW3) which showed a high degree of overlap in UVIGs from category 1 with the wastewater influent sample from the Tai-y-Bont treatment plant (RNA_TT) (Fig. 2). Many of the same UVIGs in this category (1) were also detected in the mussel (shellfish) samples and in the sediment samples. Comparing this pattern with the geographical origin of the samples (Fig. 1), revealed that the river water upstream and distant from wastewater effluent locations contained fewer detectable
Categories of UViGs observed in the dataset, binned using a combination of sequence composition and read mapping pattern.

| Groups | Number of UViGs | Total length (Mb) | N50 (nt) | Sample presence | Main virus types per category |
|--------|-----------------|-------------------|----------|-----------------|-----------------------------|
| cat_1  | 3257            | 35                | 18100    | WW, SW, SF      | dsDNA phages, NCLDVs        |
| cat_2  | 1514            | 27                | 29090    | WW, SW          | dsDNA phages, NCLDVs        |
| cat_3  | 636             | 18                | 38076    | WW, SW          | dsDNA phages, NCLDVs        |
| cat_4  | 890             | 1.9               | 2446     | WW, SW, SF, Sed | dsDNA phages, NCLDVs, (+) ssRNA viruses |
| cat_5  | 103             | 1.3               | 17154    | WW, SW          | dsDNA phages, NCLDVs        |
| cat_6  | 1077            | 3.9               | 5239     | WW, SW, SF      | dsDNA phages, NCLDVs        |
| cat_7  | 519             | 11                | 24406    | WW              | dsDNA phages, NCLDVs        |
| cat_8  | 671             | 2.1               | 3765     | WW              | (+) ssRNA viruses, dsDNA viruses, RNA plant viruses |
| cat_9  | 337             | 1.0               | 4133     | SF, Sed         | (+) ssRNA viruses, dsDNA viruses, RNA plant viruses |
| cat_10 | 200             | 0.36              | 1750     | WW              | (+) ssRNA viruses, dsDNA viruses, RNA plant viruses |
| cat_11 | 230             | 0.65              | 3703     | WW              | (+) ssRNA viruses, dsDNA viruses, RNA plant viruses |
| cat_12 | 309             | 0.88              | 3959     | SF, Sed         | (+) ssRNA viruses, dsDNA viruses, NCLDVs |
| cat_13 | 406             | 6.2               | 20764    | WW, SW, SF      | dsDNA phages, NCLDVs        |

* NCLDV: nucleo-cytoplasmic large DNA virus. Key WW – wastewater, SW – surface water, SF – shellfish, Sed – sediment.
In order to assign each UViG to a viral family and higher taxa, we used a combination of Diamond BLASTx against the viral RefSeq protein database (version 200, May 2020) and taxonomic binning using a lowest common ancestor approach with Megan6 (Buchfink et al., 2015; Huson and Weber, 2013). To reduce the number of different taxa displayed in Figure 3b, we assigned the UViGs at class or phylum level recently defined by the International Committee on Taxonomy of Viruses (ICTV) (Gorbunova et al., 2020; Koonin et al., 2020) and where this was not unambiguously possible at the Realm level, with the remainder designated as either “Viruses” (similarities to viruses belonging to multiple Realms) or “Unknown” (no similarity with any virus in the RefSeq database). Of all contigs in our set, 98% had at least one BLAST hit with the virus database (9935/10^149) and 88% (8904/10^149) were assigned to at least the viral Realm level.

The taxonomic composition of each library (Fig. 3b), normalised to 1 million reads per sample mapped to the UViGs, showed large differences in relative abundances of virus groups between both library types and samples. Scanning these data confirms the observation from Figure 2, that some of the RNA libraries were contaminated with DNA viruses. In these cases (all RNA river water libraries and RNA TI, RNA BI, RNA-TyCa/b, RNA_CP1/2/CS1), the relative abundance of dsDNA viruses, mainly tailed phases of the class Caudoviricetes, eclipsed the detected RNA virus signatures. The remainder of the RNA libraries recruited the most reads against several groups of RNA viruses, such as the phage class Leviviricetes (formerly family Leviviridae), phylum Lenarviricota and unknown RNA virus UViGs (Realm Riboviria) from a previously published study on the RNA virosphere of invertebrates (Shi et al., 2016). The majority of the DNA libraries were dominated by dsDNA bacteriophages associated with the class Caudovirales and its constituent families. Exceptions were libraries DNA TI and DNA LE, which were dominated by a small number of UViGs with ambiguous taxon assignments (i.e. classified as “Viruses” or “Unknown”). The read recruitment to the UViGs and their taxonomic binning clearly showed discrepancies between some of the replicates, most notably the RNA libraries of the shellfish digestive tissue samples. These differences are in line with a recent study by Pérez-Cataluna and colleagues who investigated library preparations for viromes of wastewater and showed that further standardisation of methods is necessary for quantitative viromics (Pérez-Cataluna et al., 2021).

In view of these discrepancies in read mapping patterns between replicates, we investigated the taxonomic bins per environment type as an indication for richness, not relative abundance (Supplementary Fig. 2). Overall, the most common RNA virus classification was the “UViG RNA virus” bin grouped within the realm Riboviria comprising a diverse set of metagenome-assembled RNA viruses [dsRNA, (+-)ssRNA, (-)ssRNA from invertebrates (Shi et al., 2016)], which contained the most UViGs from mussels, sediment and wastewater samples. The most abundant DNA virus bin was the class Caudoviricetes which groups all tailed phases of the order Caudovirales and its constituent families (Myoviridae, Siphoviridae, Podoviridae, Ackermannviridae, Autographiviridae, Drexleviridae, Herelleviridae) including crAss-like phages, and unidentified dsDNA viruses (probably tailed bacteriophages), which were particularly rich in wastewater, river water, and to a lesser extent in mussels. Wastewater was also host to a diverse group of (+-)ssRNA phages of the class Leviviricetes (~700 UViGs), with a smaller number of these viruses observed in mussels and sediment. About 6% of the total reads could not be assigned to a known group, not even at the Realm level, and were categorized as unknown viruses. While these unknown viruses represented only 6% of the total reads, they made up about a third (3 502/10 149) of the vOTUs.

3.2. Circulating human pathogens: Sapovirus, coxsackievirus and rotavirus

To investigate the potential environmental and public health impact of the UViGs, we focused on the near-complete genomes shared between wastewater and the other environments (Fig. 3a) and the taxonomic groups that contain known pathogens (human/animal). We identified 29 vOTUs of potential public health concern, further representing 73 UViGs from six families (Table 2). Interestingly, we were unable to unambiguously identify any potentially pathogenic dsDNA UViGs. The ability to reconstruct a complete papillomavirus genome in our pilot study from a subset of these sites sampled at an earlier date (Adriaenssens et al., 2018) suggests that there were in fact no predicted-pathogenic DNA viruses circulating (above the limit of detection) in the Conwy catchment at the time of sampling (June 2017). We speculate that the most likely reason for the absence of potentially pathogenic DNA viruses from the dataset is because their presence was below the limit of detection, with the discovery of the papillomavirus in the pilot study potentially due to a large shedding event. It is also possible that the diversity of dsDNA bacteriophages drives up the limit...
### Table 2

Potentially pathogenic virus groups in the UViG dataset.

| Family/group | Genus – closest relative | Potential host/ metagenome | # of contigs<sup>a</sup> | Cat<sup>b</sup> Present in samples (traces)<sup>c</sup> |
|--------------|--------------------------|---------------------------|-------------------------|-----------------------------------------------|
| *Astroviridae* | UvIG Bastro-like virus | Bat | 1 | 12 DM |
| (+n)RNA | Astrovirus - Astrovirus MLB1 | Human | 1 | 10 GI |
| *Caliciviridae* | Sapovirus - Sapovirus GIL5 | Human | 6 | 4 LI, LE, GI (SB, DM, SW5) |
| (+)nRNA | Sapovirus - Sapovirus GIL2 | Human | 2 | 4 LI, LE (GI, SB, DM, SW5) |
| *Picornaviridae* | Enterovirus - Human coxsackievirus A22 | Human | 1** | 10 GI (SB) |
| (−)nRNA | Enterovirus - Human coxsackievirus A19 | Human | 1** | 10 GI (SB) |
| *Reoviridae* | Rotavirus - Rotavirus A (NSP1) | Human | 2 | 4 LE, GI (LI, SB, DM) |
| dsRNA | Rotavirus - Rotavirus A (VP1) | Human | 2 | 4 LE, GI (LI, SB, DM) |
| | Rotavirus - Rotavirus A (VP2) | Human | 2 | 4 LE, GI (SB, DM) |
| | Rotavirus - Rotavirus A (VP3) | Human | 2 | 4 LE, GI (SB, DM) |
| | Rotavirus - Rotavirus A (NSP1) | Human | 4 | 4 BE, LI, LE, GI |
| | Rotavirus - Rotavirus A (NSP3) | Human | 4 | 4 BE, LI, LE, GI (TE, SB, DM, SW5) |
| | Rotavirus - Rotavirus A (VP1) | Human | 3 | 4 BE, LI, LE, GI (TE, SB, DM, SW5) |
| | Rotavirus - Rotavirus A (VP2) | Human | 4 | 4 BE, LI, LE, GI (TE, SB, DM, SW5) |
| | Rotavirus - Rotavirus A (VP3) | Human | 4 | 4 BE, LI, LE, GI (SB, DM, SW5) |
| | Rotavirus - Rotavirus A (VP4) | Human | 4 | 4 BE, LI, LE, GI (SB, DM, SW5) |
| | Rotavirus - Rotavirus A (VP7) | Human | 4 | 4 BE, LI, LE, GI (SB, DM, SW5) |
| | Rotavirus - Rotavirus A (NSP1) | Human | 1 | 6 LE |
| | Rotavirus - Rotavirus A (NSP3) | Human | 1 | 6 LE (LI) |

### Table 2 (continued)

| Family/group | Genus – closest relative | Potential host/ metagenome | # of contigs<sup>a</sup> | Cat<sup>b</sup> Present in samples (traces)<sup>c</sup> |
|--------------|--------------------------|---------------------------|-------------------------|-----------------------------------------------|
| *Rotavirus* | Rotavirus A (VP1) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (NSP3) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (VP1) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (VP4) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (VP7) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (NSP1) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (NSP3) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (VP1) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (VP4) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (VP7) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (NSP1) | Human | 1 | 10 GI (LI, LE, SB) |
| | Rotavirus - Rotavirus A (NSP3) | Human | 1 | 10 GI (LI, LE, SB) |

<sup>a</sup>This assignment was based on low similarity scores.

<sup>b</sup>These UViGs were partial genomes, not near-complete genomes.

<sup>c</sup>The number of UViGs clustered at 95% ANI (cd-hit-est) represented by one UViG in the dataset.

<sup>d</sup>Category as defined in Fig. 2 and Table 1.

of detection for other DNA viruses by skewing the data towards the most abundant genomes. Given the current pandemic, we also did a search for similarity of the UViG dataset with members of the *Coronaviridae* family, but no coronavirus signatures were identified in our dataset.

With respect to the family *Astroviridae*, we recovered one UViG related to bat-infecting astroviruses in muscel (*Mytilus edulis*) tissue from the Deganyw shellfishery, and one UViG in wastewater sample GI highly similar to Astrovirus MLB1 (*FJ222451*) which was sequenced from the stool of a child with acute diarrhoea (*Finkbeiner et al.*, 2008) (Supplementary Fig. 3a). For the family *Caliciviridae*, we were not able to identify any UViGs representing noroviruses, the leading cause of viral gastro-intestinal illness in the UK and indeed worldwide (*Ahmed et al.*, 2014; *FSA*, 2017; *Kirk et al.*, 2015) in contrast to our pilot study performed in autumn, where we assembled a norovirus GI.2 genome (*Adriaenssens et al.*, 2018). We did, however, find two near-complete sapovirus UViGs and six shorter contigs grouped with the near-complete genomes (*Fig. 3*, Supplementary Fig. 3b), most closely related to sapoviruses of genotype GII.2 and GII.5 that were collected from children with acute gastroenteritis in Nashville (US) (*Diez-Valcarce et al.*, 2018). This finding suggests that at the time of sampling for the dataset reported here (June 2017), sapoviruses replaced noroviruses (commonly associated with winter illness) as the main cause of gastro-intestinal disease. This theory is supported by our previous RT-qPCR detection study showing that sapovirus concentrations spiked between March and June in wastewater collected at the four WWTPs in the Conwy area (*Farkas et al.*, 2018a). However, this is difficult to formally prove as many norovirus-sapovirus cases are undiagnosed clinically, and the seasonality of norovirus and sapovirus is not consistent in all clinical settings in the UK (*Brown et al.*, 2016; *Inns et al.*, 2019).
We identified two potentially pathogenic UViGs in the Picornaviridae family (Table 2) among a host of distantly related picorna-like viruses (Fig. 4). The two potentially pathogenic picornavirus UViGs, which were represented by only partial genome sequences (Supplementary Fig. 3c), could be identified as coxsackieviruses of the species Enterovirus C, most closely related to human coxsackieviruses A19 and A22 reportedly involved in meningitis, gastroenteritis and herpangina (Tapparel et al., 2013; Zell et al., 2017). Detailed phylogenetic analysis of all calici- and picorna-like RNA-dependent RNA polymerase (RdRP) sequences (Fig. 4) showed that the majority of UViGs found in this study fell within a very diverse, ill-resolved clade (low branch support) comprised of environmental sequences nested within the order Picornavirales (bottom half of circle, Fig. 4). Based on the RdRP sequences, only three UViGs in the picorna-calici group were designated potential human pathogens (Fig. 4, black arrows), the two sapovirus UViGs and one of the coxsackievirus UViGs. Only the sapovirus UViG LI_NODE_9 was detected in all sample types, posing a potential risk for human health as it was detected in the mussel beds of the commercial shellfishery, sediment on the tourist beach and estuarine water (Supplementary Fig. 4). PCR-based detection of sapoviruses in older studies show that among cases of gastro-intestinal disease, sapoviruses accounted for only 4% of cases (vs 36% for noroviruses) (Amar et al., 2007), however, the primers used in that study (SR80 (Noel et al., 1997), JV33 (Vinjec et al., 2000)) did not match the two sapovirus genomes reconstructed in this study (data not shown). The detection of this complete genome sequence from two different wastewater treatment plants is another indication that sapoviruses are more common in the UK than previously reported, similar to its incidence reported in other countries (Mann and Liebert, 2019; Pang et al., 2019; Varela et al., 2018).

While the phylogenetic analysis does not provide enough evidence for the presence of plant-pathogenic picorna-like viruses in the Conway river catchment, there is a set of UViGs present that is mollusc-specific (coloured orange in Figure 4). It is therefore likely that we have sequenced and reconstructed a set of mussel/shellfish-associated or infecting viruses.

Within the non-redundant, species-level clustered dataset, 18 UViGs grouped into three categories according to read recruitment pattern, and were assigned to the species Rotavirus A in the family Reoviridae, representing a further 41 contigs. Analysis of roviruses is confounded by their segmented nature, i.e. members of the genus Rotavirus contain 11 segments of dsRNA, and the size of the smaller segments is below our 1000 nt contig length threshold. We therefore analysed all contigs larger than 500 nt for the presence of rotavirus signatures and assigned genotypes to each segment recovered (Fig. 5a). The most common rotavirus A (RVA) genome constellation recovered was G2-P[4]-I2-R2-C2-M2-A2-N2-T2-E2, with additional genotypes R1 for the RNA-dependent RNA polymerase (RdRP) segment, C1 for the segment encoding VP2, P[1] and P[14] for the outer capsid-encoding segment, A3 and A11 for NSP1 and T6 for NSP2. In many of the wastewater samples, we assembled multiple contigs of the same segment indicating the presence of several co-circulating population lineages of rotavirus A in the population. Phylogenetic analysis of the outer capsid proteins (VP4) confirmed the genotype clustering, and comparison with isolated rotavirus VP4 sequences points towards a human origin for the P[4] and P[14] genotypes (found in samples BE, LI, LE and GI) and a potential bovine zoonotic origin for the P[1] genotype segment (Fig. 5b). The RVA genome segments recovered here are markedly different to those recovered from wastewater influent from Llanrwst (LE-LL) in our pilot study 10 months prior (Adriaenssens et al., 2018), for which the dominant genotypes of RVA were G8/G10-P[1]-P[14]/P[41], and a diverse set of rotavirus C segments were also present. We can conclude that rotavirus shedding into wastewater within the population varied both spatially and temporally, but more data are required to investigate any possible seasonal patterns. From the distribution of the rotavirus fragments in shellfish, beach sediment and estuarine water (Table 2), we speculate that rotaviruses could pose a potential risk for human health in relation to shellfish consumption or recreational activities and bathing within the immediate coastal zone. However, rotaviruses mainly affect...
infants and children under the age of five (Hamborsky et al., 2015), who are less likely to engage with such activities which may be the reason for the lack of reported illnesses.

We identified a number of small contigs related to ssDNA circular circoviruses and parvoviruses, that were originally recovered from environmental or host-associated metagenomes (Dayaram et al., 2015; Phan et al., 2015; Zawar-Reza et al., 2014). Of these, four circovirus-associated vOTUs, representing 18 contigs, showed significant sequence similarity to previously described UViGs from animal or wastewater metagenomes. One parvovirus contig, assigned to the genus Ambidensovirus, was related to a bat metagenome sequence. However, for these types of ssDNA virus UViGs, any causative links with disease syndromes would be very tenuous.

3.3. A conceptual model for virus circulation in a freshwater catchment area

The data presented in this study support the following conceptual model of virus circulation in the river system (Fig. 6). Upstream, the more pristine regions of the river with low human and livestock inputs, viral species richness is low and the water virome is dominated by dsDNA tailed bacteriophages (caudoviruses) and a few algal viruses of...
the family *Phycodnaviridae*. At certain points along the river, wastewater effluent from large treatment plants and smaller scale septic tank discharge enters the water. This effluent is much less rich in viruses than untreated wastewater (influent) but can still contain over 1000 different viral species per litre. The entire spectrum of viral diversity detected in this study is represented in effluent (treated wastewater), with DNA and RNA bacteriophages (predicted to infect members of the human gut microbiome) the most commonly detected groups (caudoviruses, levi-viruses). Nucleo-cytoplasmic large DNA virus (NCLDV; phyco-cnaviruses, mimiviruses, iridoviruses) and common plant-derived viruses present in food and excreted by the human digestive tract (mainly tobamoviruses such as pepper mild mottle virus (Rosario et al., 2009; Zhang et al., 2006)) and groups of enteric viruses such as sapovirus, rotavirus and astrovirus within a wider collection of unclassified RNA viruses are also well represented. These communities are both spatially and temporally distinct. Upon entering the river, the pathogenic virus groups fall below the limit of detection by virome sequencing, which can be attributed primarily to dilution by the river water. However, close to an effluent site and at the estuary that is under tidal control, the number of viral species detected in water samples is much higher. Beach sediment and filter-feeding shellfish (in this case mussels, *Mytilus edulis*) then act as entrapment matrices enriching the viral content from the surrounding water (Maalouf et al., 2010; Whitman et al., 2014). In the majority of cases, the UVIgs that were assembled from wastewater recruited fewer reads from beach sediment, mussel tissue or estuary water libraries, and the read mapping over the genome length was often patchy, leading us to hypothesize that these genomes, and by extension the virions, are likely to be substantially degraded. At the same time, we observed sediment- and mussel-specific viral communities represented by full genomes, mainly picorna-like RNA viruses and unclassified UVIgs from invertebrates (Shi et al., 2016), thus excluding technical bias as the explanation for our failure to detect intact pathogenic virus genomes in sediment and shellfish. In the scenario that we propose, shellfish and sediment become enriched in viruses that are recruited from the environment by filter feeding and adsorption, respectively. Those viruses that do not undergo active replication in the newly occupied niche (human, animal and plant pathogens in particular) are degraded over time or diluted below the limit of detection, while viruses that infect the shellfish, the shellfish microbiome, diatoms or sediment-associated bacteria are maintained, enabling detection of their full genome sequences. In this scenario, the risk of illness due to consumption of shellfish, contact with sediment (beach sand) or swimming, would depend on the time interval between uptake/adsorption of pathogenic viruses in the matrix and ingestion by a human subject. To critically evaluate this, further experimental data on the infectivity/survival kinetics for each viral species are required, as this is likely to vary markedly between viral groups. However, this would be a Herculean endeavour, given the diversity of viruses detected here, the difficulty in propagation and the absence of routine infectivity assays. The conceptual model is supported by the results of our previous year-long q cycle.

## 4. Conclusion

Viruses and their genetic material are commonly discharged in the environment, but their risk to human health is driven by community outbreaks leading to viral shedding into the wastewater, leading to temporal and spatial variations in the specific genotypes detected. In the environment, these viruses are then subject to cycles of dilution, enrichment and virion degradation influenced by local geography, weather events and tidal effects. Our analyses show that viromics is a useful tool to assess viral diversity in the aquatic environment in order to explore new and emerging human and animal health threats.

### Availability of data and material

The datasets generated and analysed in this study are available from the Sequence Read Archive (SRA) under BioProject PRJNA509142, accession numbers SRR8299359 to SRR8299398.

### 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.

### Acknowledgements

We thank Dwr Cymru/Welsh Water for access to the wastewater treatment plants. We thank the following people for assistance with sample collection: Dr Emma Green and Harry Riley, Bangor University. This work was supported by the Natural Environment Research Council (NERC) and the Food Standards Agency (FSA) under the Environmental Microbiology and Human Health (EMHH) Programme (VIRAQUA; NE/M010996/1). E.M.A is currently funded by the Biotechnology and Biological Sciences Research Council (BBSRC) Institute Strategic Programme Gut Microbes and Health (BB/R012490/1) and its constituent projects BBS/E/F/000PR10353 and BBS/E/F/000PR10356.

### Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2021.117568.

### References

Adriaenssens, E., Brister, J.R., 2017. How to name and classify your phage: An informal guide. Viruses 9, 70. https://doi.org/10.3390/v9040070.

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

Ahmed, S.M., Hall, A.J., Robinson, A.E., Verhoef, L., Premkumar, P., Parashar, U.D., Koopmans, M., Lopman, B.A., 2014. Global prevalence of norovirus in cases of gastroenteritis: A systematic review and meta-analysis. Lancet Infect. Dis. 14, 725–730. https://doi.org/10.1016/s1473-3099(14)70767-4.

Amar, C.F.L., East, C.L., Gray, J., Iriurzita-Gomara, M., Maclure, E.A., McLauchlin, J., 2007. Detection by PCR of eight groups of enteric pathogens in 4,627 faecal samples: Re-examination of the English case-control Infectious Intestinal Disease Study (1993-1996). Eur. J. Clin. Microbiol. Infect. Dis. 26, 311–323. https://doi.org/10.1007/s10098-006-0290-8.

Anisimova, M., Kutil, M., Gotoh, J., Desimone, C., Gascuel, O., 2011. Selection of a model selection method for phylogenetics. J. Mol. Evol. 73, 307–320. https://doi.org/10.1007/s00239-011-9523-2.

Balduc, B., Youens-Clark, K., Roux, S., Hurwitz, B.L., Sullivan, M.B., 2017. iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. ISME J 11, 7–14. https://doi.org/10.1038/ ismej.2016.89.

Bray, N.L., Fumal, E., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527. https://doi.org/10.1038/ nbio.2016.52.

Brown, J.R., Shah, D., Breuer, J., 2016. Viral gastrointestinal infections and norovirus genotypes in a paediatric UK hospital, 2014-2015. J. Clin. Virol. 84, 1–6. https://doi.org/10.1016/j.jcv.2016.08.298.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176.
