Tracing the fate of wastewater viruses reveals catchment-scale virome diversity and connectivity

Evelien Adriaenssens (evelien.adriaenssens@quadram.ac.uk)
Quadram Institute Bioscience https://orcid.org/0000-0003-4826-5406
Kata Farkas
Bangor University
James McDonald
Bangor University https://orcid.org/0000-0002-6328-3752
David Jones
University of Western Australia https://orcid.org/0000-0002-1482-4209
Heather Allison
University of Liverpool https://orcid.org/0000-0003-0017-7992
Alan McCarthy
University of Liverpool

Article

Keywords: wastewater, virus, treatment, environmental epidemiology, freshwater catchment

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-62137/v1

Version of Record: A version of this preprint was published at Water Research on August 1st, 2021. See the published version at https://doi.org/10.1016/j.watres.2021.117568.

RUNNING TITLE: Illuminating virome diversity and ecology in a wastewater-impacted catchment

Tracing the fate of wastewater viruses reveals catchment-scale virome diversity and connectivity

Evelien M. Adriaenssens [1,2]✉, Kata Farkas[3], James E. McDonald[3], David L. Jones[3,5], Heather E. Allison[1] and Alan J. McCarthy[1]

[1] Institute of Integrative Biology, University of Liverpool, Liverpool, UK; [2] Quadram Institute Bioscience, Norwich, UK; [3] School of Natural Sciences, Bangor University, Bangor, UK; [4] School of Ocean Sciences, Bangor University, Bangor, UK; [5] AWA School of Agriculture and Environment, The University of Western Australia, Perth, Australia

✉ evelien.adriaenssens@quadram.ac.uk

Abstract

The discharge of wastewater-derived viruses in aquatic environments impacts catchment-scale virome composition and is a potential hazard to human health. Here, we used viromic analysis of RNA and DNA virus-like particle preparations to track virus communities entering and leaving wastewater treatment plants and the connecting river catchment system and estuary. 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. We reconstructed >40,000 partial viral genomes into 10,149 species-level groups, dominated by dsDNA and (+)ssRNA bacteriophages (*Caudovirales* and *Leviviridae*). We identified 73 (partial) genomes comprising six families that could pose a risk to human health; *Astroviridae*, *Caliciviridae* (sapovirus), *Picornaviridae* (cocksackievirus),

Reoviridae (rotavirus), *Parvoviridae* and *Circoviridae*. Based on the pattern of viral incidence, we observe that wastewater-derived viral genetic material is commonly deposited in the environment, but due to fragemented nature of these viral genomes, the risk to human health is low, and is more likely driven by community transmission, with wastewater-derived viruses subject to cycles of dilution, enrichment and virion degradation influenced by local geography, weather events and tidal effects. Our data illustrate the utility of viromic analyses for wastewater- and environment-based epidemiology, and we present a conceptual model for the circulation of viruses in a freshwater catchment.

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. 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 [1–10]. 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.

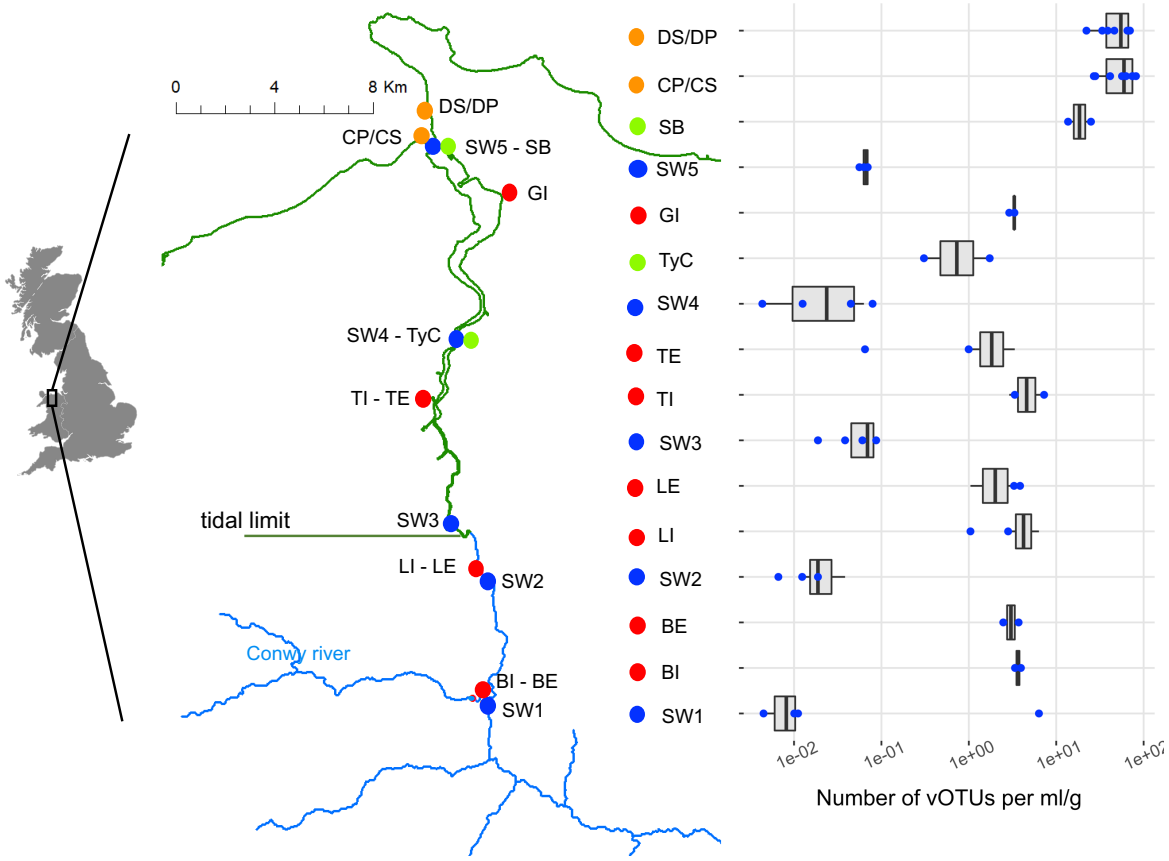
The current gold standard method for investigating enteric viruses in the environment is q(RT)-PCR, a technique that provides reliable quantitative information on the presence of the genomic material of target viruses, but requires prior knowledge on the identity of the virus and its genome sequence [11,12]. 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 [13]. 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 virus diversity and geographical distribution, with minimal bias.

Virome analyses are transforming our understanding of viral diversity and function in the biosphere 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 Extended Data Table 1). We assembled over 40,000 partial or near-complete genomes (UViGs Uncultivated Virus Genomes, [14]) 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 viromes provides an assessment of viral diversity in the wastewater-impacted Conwy river catchment, encompassing information on the dynamics of viral deposition along the river system leading to viral enrichment at the estuary, including shellfish destined for human consumption and a recreational bathing beach. Viral genome reconstruction revealed general patterns of viral enrichment, dilution and degradation, and the implications for human health.

Results & Discussion

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 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 (Fig. 1).



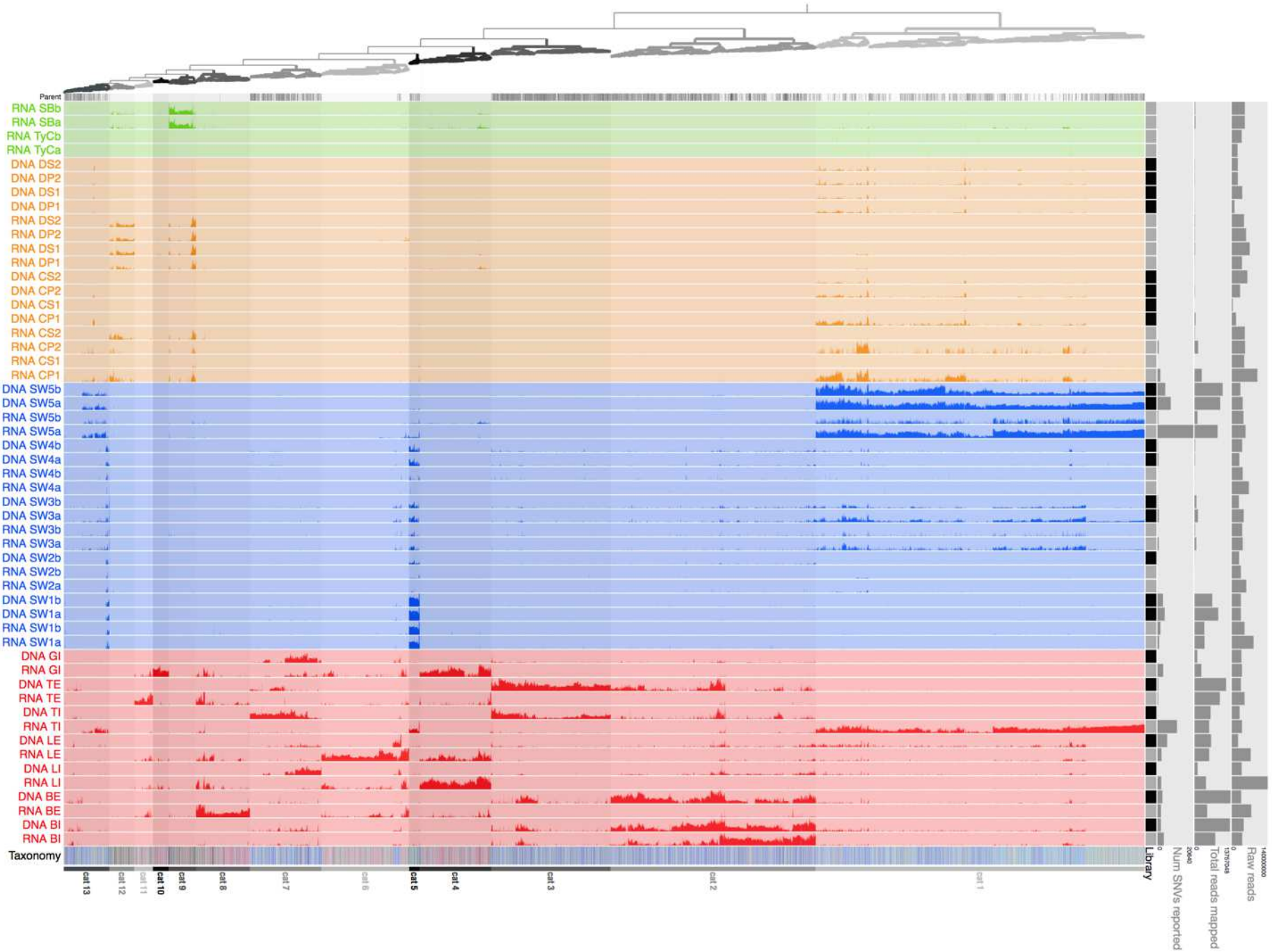
77

78 **Fig. 1|Viral abundance along the wastewater impacted Conwy river catchment and coastal zone. Left**: Schematic of the
79 Conwy river catchment with sampling sites designated by colour-coded dots (red – wastewater, blue – surface water, green
80 – sediment, orange – shellfish). The section of the river within the tidal limit is designated in green. Map of Great Britain by
81 Free Vector Maps. **Right**: Boxplot representation of the number of species (vOTUs) detected in each sample per ml or g of
82 sample extracted, composed of RNA and DNA libraries and biological replicates, species numbers for single libraries in blue
83 dots.

84 We further investigated the differential patterns of abundance of each UViG in each library by
85 mapping reads of all samples against the vOTU database and visualised the data with Anvi'o (Fig. 2),
86 to identify 13 categories of viral species abundance and composition patterns (Table 1). The
87 wastewater samples contained the most diverse set of UViGs in absolute numbers, however, each
88 wastewater treatment plant yielded a geographically distinct viral community. The river water
89 samples contained a lower absolute richness of viruses than the wastewater, except for sample SW5
90 (and to a lesser extent, SW3) which showed a high degree of overlap in UViGs from category 1 with
91 the wastewater influent sample from the Tal-y-Bont treatment plant (RNA_TI) (Fig. 2). Many of the
92 same UViGs in this category (1) were also detected in the mussel (shellfish) samples and in the
93 sediment samples. Comparing this pattern with the geographical origin of the samples (Fig. 1),

94 revealed that the river water upstream and distant from wastewater effluent locations contained
95 fewer detectable virus species, while the locations immediately downstream of an effluent pipe (SW3)
96 or in the tidal estuary (SW5, mussels, beach sediment) were enriched for UViGs. The high virus
97 richness in the tidal estuary (SW5) can be explained by the mixing of river and marine waters during
98 tidal flow [15]. The SW5 wastewater treatment Combined Sewage Outflow (CSO) periodically discharges
99 untreated sewage directly upstream of SW5, representing a sewage input that largely avoids the
100 dilution effect of estuary water and is consistent with the higher detection of faecal indicator bacteria
101 previously at this location [16]. The viral species count per sample (Fig. 1) also demonstrated that
102 wastewater effluent samples generally had a lower species tally than influent, indicating that
103 wastewater treatment reduced viral species richness, although it was not possible to test the
104 statistical significance of these data.

105


Parent
RNA SBb
RNA SBa
RNA TyCb
RNA TyCa
DNA DS2
DNA DP2
DNA DS1
DNA DP1
RNA DS2
RNA DP2
RNA DS1
RNA DP1
DNA CS2
DNA CP2
DNA CS1
DNA CP1
RNA CS2
RNA CP2
RNA CS1
RNA CP1
DNA SW5b
DNA SW5a
RNA SW5b
RNA SW5a
DNA SW4b
DNA SW4a
RNA SW4b
RNA SW4a
DNA SW3b
DNA SW3a
RNA SW3b
RNA SW3a
DNA SW2b
RNA SW2b
RNA SW2a
DNA SW1b
DNA SW1a
RNA SW1b
RNA SW1a
DNA GI
RNA GI
DNA TE
RNA TE
DNA TI
RNA TI
DNA LE
RNA LE
DNA LI
RNA LI
DNA BE
RNA BE
DNA BI
RNA BI
Taxonomy
cat 13
cat 12
cat 11
cat 10
cat 9
cat 8
cat 7
cat 6
cat 5
cat 4
cat 3
cat 2
cat 1
Library
Num SNVs reported
Total reads mapped
Raw reads

6

Fig. 2 | **Differential patterns of abundance of each viral genome (UViG) along the wastewater impacted Conwy river and coastal zone.** Anvi'o - mean coverage per contig (split). Each row is a sequencing library, coloured by its sample type (green = sediment; orange = mussels; blue = river/estuary water; red = wastewater). Each column (leaf in top dendrogram) is a contig or a split of a contig (in cases where contigs were larger than 11 kb). The height of the bar in each row is the log mean coverage across the contig or contig split length. The contigs are clustered (top dendrogram) according to their sequence composition and differential coverage using Euclidean distance and Ward linkage. Based on this clustering, we identified 13 categories of UViGs, indicated by shades of grey in the dendrogram and numbered at the bottom of the plot. The bottom row represents the taxonomy assigned by Kaiju (using its viral database) to the predicted genes in each contig. Contigs without assigned taxonomy are depicted in grey, dsDNA bacteriophages in shades of blue, other dsDNA viruses in shades of green, ssDNA viruses in shades of yellow, RNA (ds, (+)ss, (-)ss) in shades of purple/red. The right hand panels show the library type (RNA = grey; DNA = black), the number of single nucleotide variants (SNVs) found after read mapping (0-20,640), the total number of reads mapped to contigs (0-13,757,048) and the total number of raw sequencing reads (before QC and contamination screen; 0-140,000,000).

Examination of UViGs grouped per environment type for shared viral species [cut-off for detection 10 TPM (transcripts per million, see methods)] confirmed our observation that absolute richness was highest in wastewater samples (2692 unique vOTUs; Fig. 3a). River/estuarine water (82 unique UViGs), mussels (137) and sediment (100) all contained an order of magnitude fewer unique vOTUs. The majority of the vOTUs present in mussels and sediment were shared with wastewater; out of 4692 vOTUs detected in mussels, 3917 were also detected in wastewater (83%), and for sediment this was 1464 out of 1944 (75%) (Fig. 3a). Even though most of these vOTUs were likely bacteriophages, the high connectivity of these environments is a cause for concern, indicating potential sources of contamination that pose a risk to human health, and is investigated in more detail below. The categories of vOTUs in wastewater encompassed all virus types detected in this study (Table 1), whereas those specific to mussel shellfish and sediment comprised primarily RNA viruses. In the Materials and Methods Sequencing section, we describe technical difficulties during sequencing library construction that may have resulted in the underestimation, or even failure to detect, a group of mussel and sediment-specific DNA viruses.

Table 1| 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[a]
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	(+)ssRNA phages, dsRNA viruses, RNA plant viruses
cat_5	103	1.3	17154	WW, SW	dsDNA phages, NCLDVs
cat_6	1077	3.9	5239	WW, SW, SF	(+)ssRNA viruses, dsRNA viruses, dsDNA phages
cat_7	519	11	24406	WW	dsDNA phages, NCLDVs
cat_8	671	2.1	3765	WW	(+)ssRNA viruses, ssDNA viruses
cat_9	337	1.0	4133	SF, Sed	(+)ssRNA viruses, dsRNA viruses, ssDNA viruses
cat_10	200	0.36	1750	WW	(+)ssRNA viruses, dsRNA viruses, ssDNA viruses
cat_11	230	0.65	3703	WW	(+)ssRNA viruses, ssDNA viruses
cat_12	309	0.88	3959	SF, Sed	(+)ssRNA viruses, dsDNA phages
cat_13	406	6.2	20764	WW, SW, SF	dsDNA phages, NCLDVs

[a] 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 [17,18]. 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) [19,20] 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 phages 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 family *Leviviridae*, class *Lenarviricota* and unknown RNA virus UViGs (Realm *Riboviria*) from a previously published study on the RNA virosphere of invertebrates [21]. The majority of the DNA libraries were dominated by dsDNA bacteriophages associated with the class *Caudoviricetes* 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. 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.

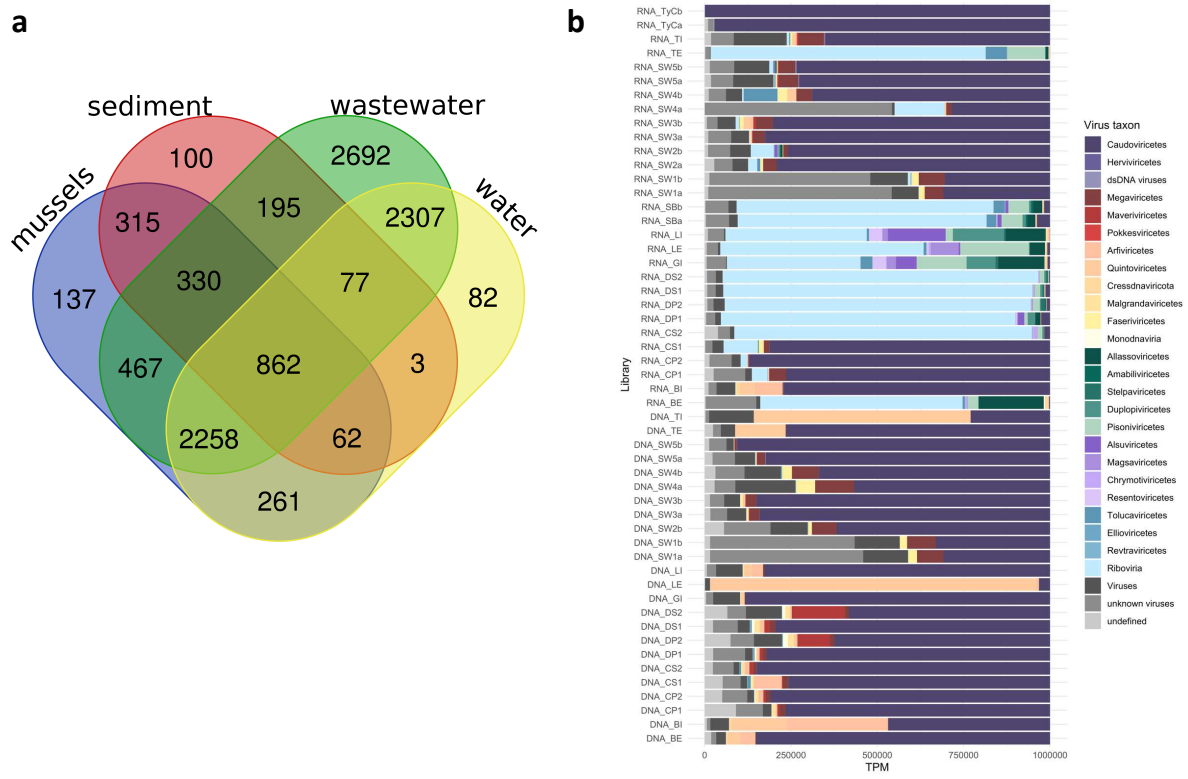


Fig. 3 | Commonality and taxonomic composition of viral genomes (UViG) in samples types from the wastewater impacted Conwy river and coastal zone. a, Venn diagram representation of the number of UViGs shared between different

environment types (min 10 TPM for detection). **b,** Relative abundances of the UViGs at the virus class level per sequencing library, normalized per library as transcipts (=contig) per million (TPM). dsDNA viruses in shades of dark purple and red; ssDNA viruses in shades of pink and yellow; RNA viruses in shades of green, purple and blue; unknown viruses in shades of grey.

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 (Extended Data 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 [21]], 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 phages of the order *Caudovirales* and its constituent families (*Myoviridae*, *Siphoviridae*, *Podoviridae*, *Ackermannviridae*, *Autographiviridae*, *Drexlerviridae*, *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 family *Leviviridae* (~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.

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

198 date [22] suggests that there were in fact no predicted-pathogenic dsDNA viruses circulating (above the
199 limit of detection) in the Conwy catchment at the time of sampling (June 2017). As an additional check,
200 we did a search for similarity of the UViG dataset with members of the *Coronaviridae* family, but no
201 coronavirus signatures were identified in our dataset.

202

203 **Table 2**| **Potentially pathogenic virus groups in the UViG dataset.**

Family/group	Genus – closest relative	Potential host/ metagenome	# of contigs[a]	Cat[b]	Present in samples (traces)[c]
Astroviridae	UviG Bastro-like virus*	Bat	1	12	DM
(+)ssRNA	*Astrovirus* - Astrovirus MLB1	Human	1	10	GI
Caliciviridae	*Sapovirus* - Sapovirus GII.5	Human	6	4	LI, LE, GI (SB, DM, SW5)
(+)ssRNA	*Sapovirus* - Sapovirus GII.2	Human	2	4	LI, LE (GI, SB, DM, SW5)
Picornaviridae	*Enterovirus* - Human coxsackievirus A22	Human	1**	10	GI (SB)
(+)ssRNA	*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 (BI, LI, SB, DM)
	Rotavirus - Rotavirus A (VP3)	Human	2	4	LI, 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, TE, GI (SB, DM, SW5)
	Rotavirus - Rotavirus A (VP4)	Human	4	4	BE, LI, LE, GI (TE, SB, DM, SW5)
	Rotavirus - Rotavirus A (VP7)	Human	4	4	BE, LI, LE, GI (TE, SB, DM, SW5)
	Rotavirus - Rotavirus A (NSP1)	Human	1	6	LE
	Rotavirus - Rotavirus A (NSP3)	Human	1	6	LE (LI)
	Rotavirus - Rotavirus A (VP1)	Human	1	6	LE (LI)
	Rotavirus - Rotavirus A (NSP3)	Human	1	10	GI (LI, LE, SB)
	Rotavirus - Rotavirus A (VP1)	Human	1	10	GI (LI, SB, DM)
	Rotavirus - Rotavirus A (VP4)	Human	1	10	GI (LE, SB)
	Rotavirus - Rotavirus A (VP7)	Human	1	10	GI
Circoviridae	UviG CRESS-like virus	Animals	8	1	BE (LI, LE, TI, SW4)
ssDNA	UviG CRESS-like virus	Animals	1	2	SW5, TI
	UviG Human fecal virus Jorvi3	Human	2	2	SW3, TI, LI, BI, BE
	UviG Giant panda circovirus 1	Mammals	7	2	TI, TE
Parvoviridae	*Ambidensovirus* - Densovirus	Bat	1	6	LE (SW3)
ssDNA	SC444				

204 * This assignment was based on low similarity scores.
205 ** These UViGs were partial genomes, not near-complete genomes.
206 [a] The number of UViGs clustered at 95% ANI (cd-hit-est) represented by one UViG in the dataset.
207 [b] Category as defined in Fig. 2 and Table 1.

[c] Sample codes as in Table 1. DM = Deganwy mussels (shellfishery 1); CM = Conwy mussels (shellfishery 2). Present indicates that over half of the contig length was covered with reads of at least 3-fold coverage, traces indicates that at least two reads mapped to the UViG.

With respect to the family *Astroviridae*, we recovered one UViG related to bat-infecting astroviruses in mussel (*Mytilus edulis*) tissue from the Deganwy 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 [23] (Extended Data 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 [24–26] in contrast to our pilot study performed in autumn, where we assembled a norovirus GI.2 genome [22]. We did, however, find two near-complete sapovirus UViGs and six shorter contigs grouped with the near-complete genomes (Fig. 3, Extended Data 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) [27]. 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 [28]. 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 [29,30].

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 (Extended Data 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 [31,32]. 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 (Extended Data 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)[33], however, the primers used in that study (SR80 [34], JV33 [35]) 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 [36–38].

While the phylogenetic analysis does not provide enough evidence for the presence of plant-pathogenic picorna-like viruses in the Conwy 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.

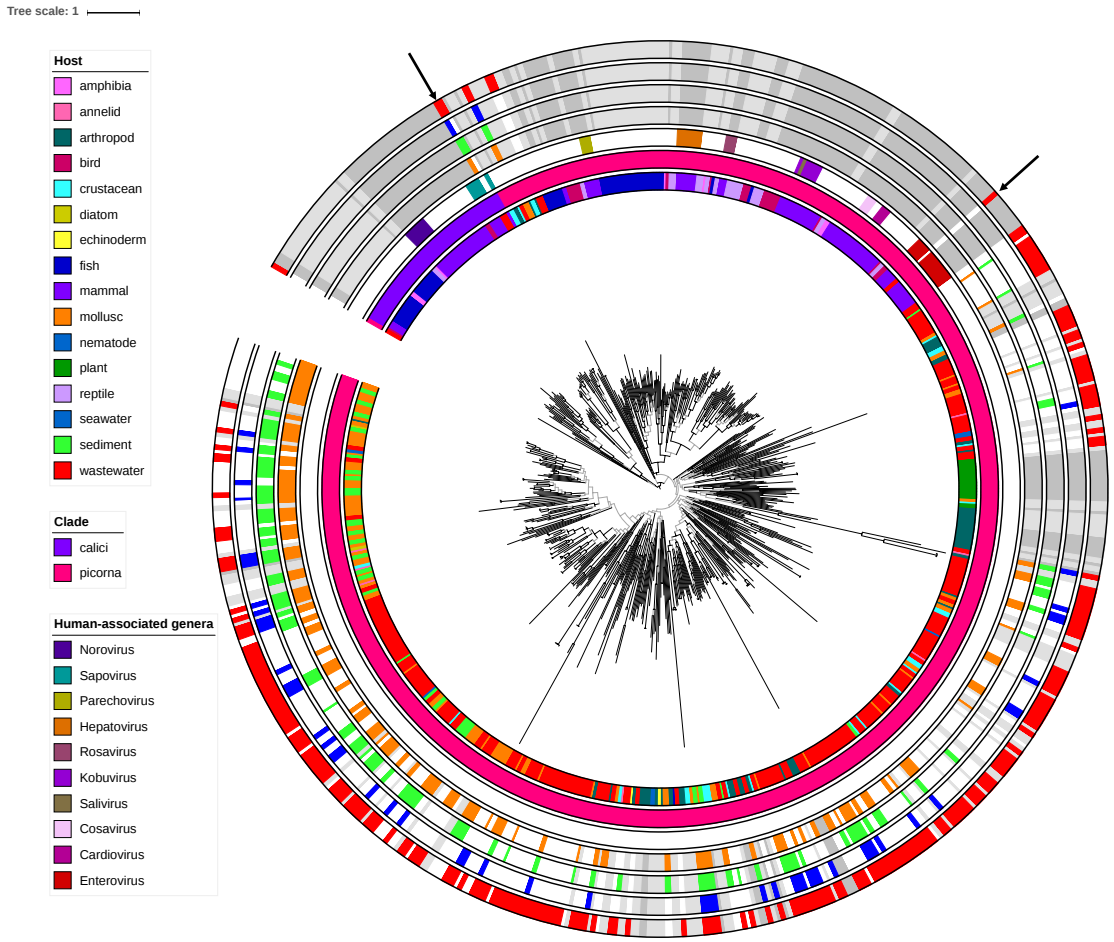
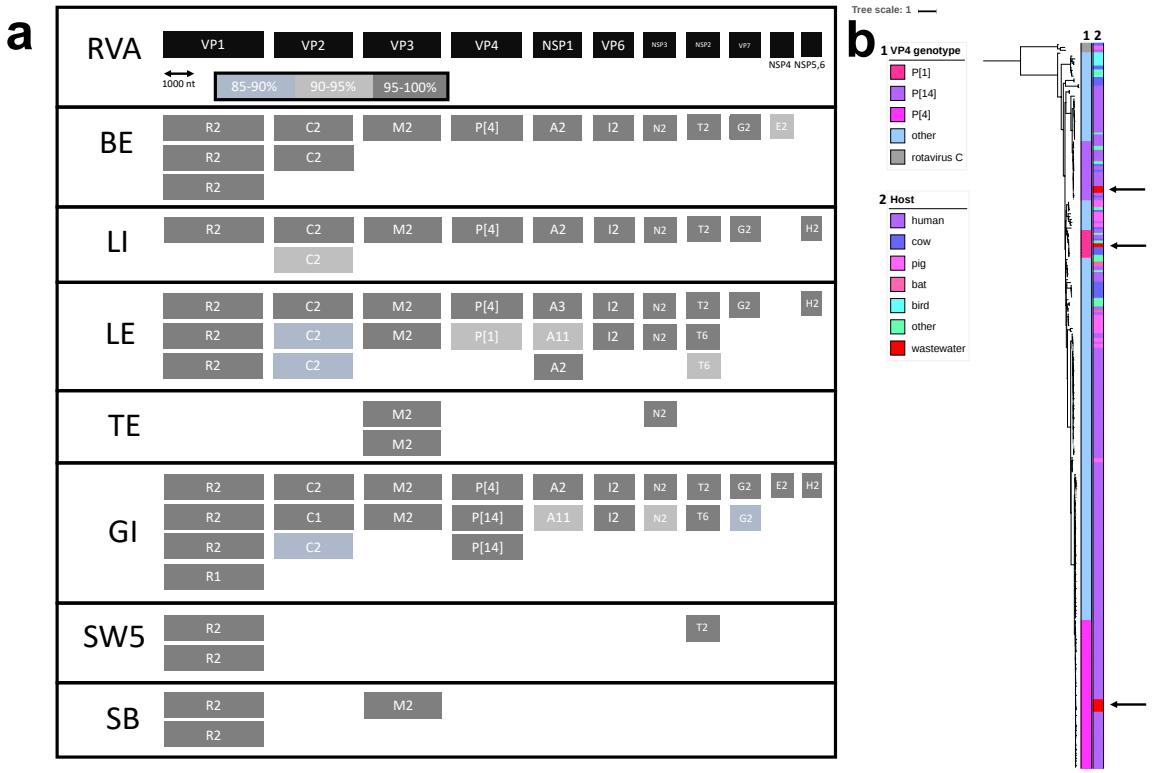


Fig. 4| Maximum likelihood phylogenetic tree of the RdRP amino acid sequences of viruses/genomes assigned to the family *Caliciviridae* and the order *Picornavirales* built with IQ-TREE [39] and visualized with ITOL [40]. The multiple alignment consisted of 622 sequences and 695 amino acid sites, aligned using MAFFT and trimmed with Trimal [41,42]. The best fit model was LG+F+R10 as determined with ModelFinder [43]. Branch support was calculated using the Shimodaira Hasegawa – approximate Likelihood Ratio Test (SH-aLRT) and the UFBoot (ultrafast bootstrap) algorithm on 1000 replications with nodes below 80% (SH-aLRT) and 95% (UFBoot) indicated in grey [44,45]. The three inner colour strips from inside to outside indicate respectively: viral host or metagenome the RdRP was extracted from, predicted clade, human-associated genera (only reference genomes from human pathogenic viruses coloured). The four outside colours strips indicate detection in shellfish samples (orange), beach/river sediment samples (green), river/estuarine water samples (blue) and wastewater samples (red), with other virome-derived UViGs in light grey and reference virus sequences in middle grey. The black arrows indicate the UViGs found in this study that are likely human pathogens.

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. Analyses of reoviruses 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-LI) in our pilot study 10 months previously [22], 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 can infer that rotaviruses 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 [46], who are less likely to engage with such activities which may be the reason for the lack of reported illnesses.



289

290 **Fig. 5**| **Rotavirus A (RVA) in the virome datasets. a,** The 11 segments of the reference genome of RVA ranked by size in
291 black. RVA segments recovered per sample below showing the predicted genotype of the segment and the percentage of
292 nucleotide identity with a representative of that genotype as calculated by the RotaC 2.0 tool. **b,** Maximum likelihood
293 phylogenetic tree of the VP4 amino acid sequences of selected representatives of all RVA genotypes build with IQ-TREE [39]
294 and visualized with ITOL [40]. The multiple alignment consisted of 253 sequences and 774 amino acid sites, aligned using MAFFT
295 and trimmed with Trimal [41,42]. The best fit model was FLU+F+R8 as determined with ModelFinder [43]. Branch support was
296 calculated using the UFBoot (ultrafast bootstrap) algorithm on 1000 bootstraps and is indicated with branch colours in
297 shades of grey, with support values higher than 95% in black [44]. Colour strip 1 indicates the genotype clustering, using RVC
298 isolates as outgroup. Colour strip 2 shows the host of the isolates with arrows indicating the virome-derived sequences.

299

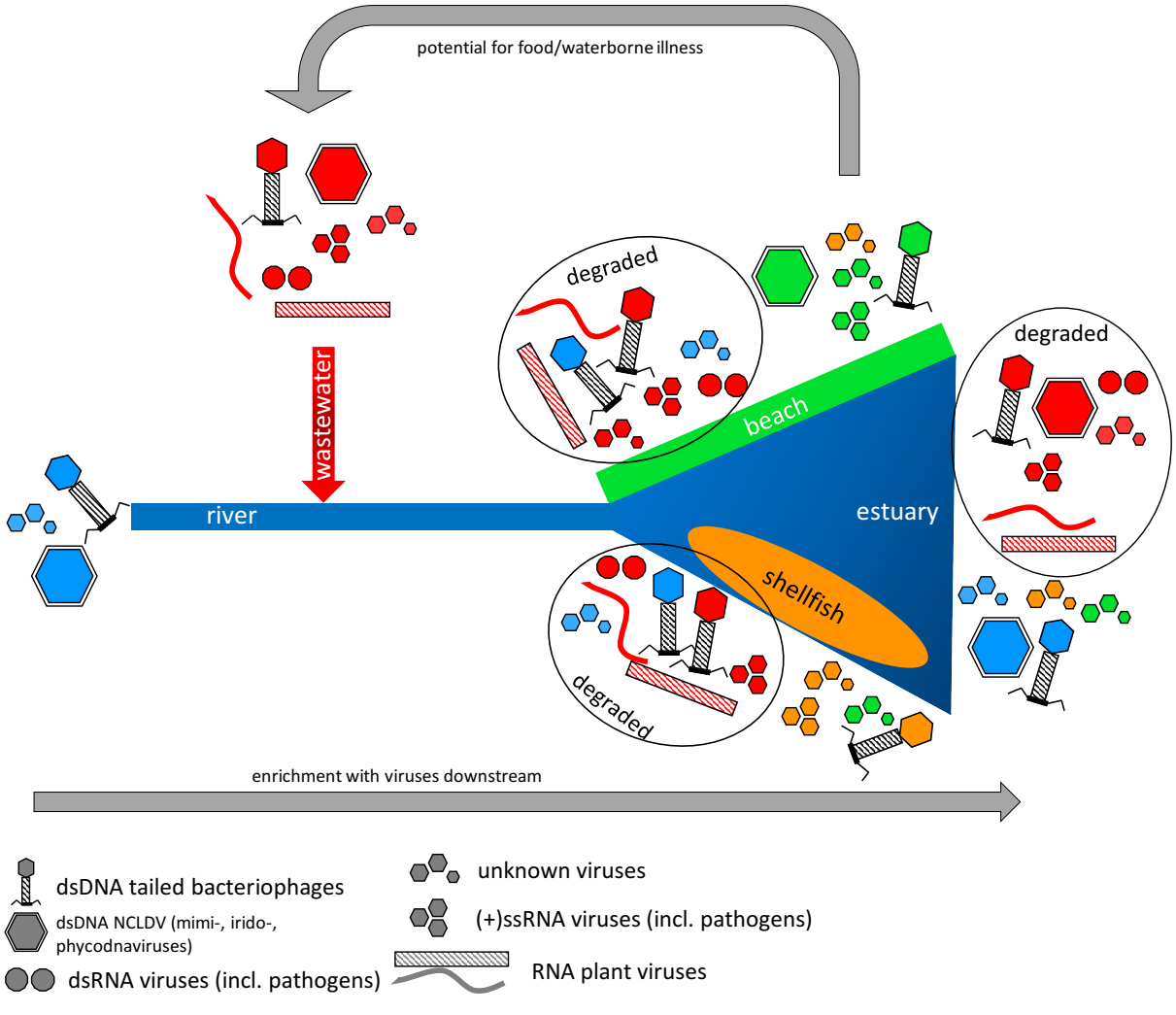
300 We identified a number of small contigs related to ssDNA circular circoviruses and
301 parvoviruses, that were originally recovered from environmental or host-associated metagenomes [47–
302 49]. Of these, four circovirus-associated vOTUs, representing 18 contigs, showed significant sequence
303 similarity to previously described UViGs from animal or wastewater metagenomes. One parvovirus
304 contig, assigned to the genus *Ambidensovirus*, was related to a bat metagenome sequence. However,
305 for these types of ssDNA virus UViGs, any causative links with disease syndromes would be very
306 tenuous.

A model for virus circulation in a freshwater catchment area. The data presented in this study support the following model of virus circulation in the river system (Fig. 6).

Upstream, in 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 discharges enter the water. This effluent is much less rich in viruses than untreated wastewater (influent) but can still contain over a 1000 different viral species per litre. The entire spectrum of viral diversity detected in this study is represented in effluent, with DNA and RNA bacteriophages (predicted to infect members of the human gut microbiome) the most commonly detected groups (caudoviruses, leviviruses). Nucleo-cytoplasmic large DNA virus (NCLDVs; phycodnaviruses, 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 [50,51]) and groups of enteric viruses such as sapovirus, rotavirus and astrovirus within a wider collection of unclassified RNA viruses are also well represented. 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 [52,53]. 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 [21], thus excluding technical bias as the explanantion 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. This model is supported by the results of our previous year-long q(RT)-PCR study on a subset of enteric viruses, which showed that they were still detected at high titres in wastewater post-treatment, followed by lower titres in river water, shellfish and sediment, and ultimately undergoing capsid degradation in environmental matrices [28]. Our model of viral circulation is also consistent with theorectical simulations of viral discharge from wastewater treatment plants into the coastal zone [15]. Importantly, these models have indicated that tidal movement allows viruses in estuarine water to come into contact with shellfisheries and beaches on numerous occasions over a period of days to weeks depending on the lunar tidal cycle.

In 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, and subject to cycles of dilution, enrichment and virion degradation influenced by local geography, weather events and tidal effects.



354

355 **Fig. 6| Model for the circulation of viruses in a river catchment and coastal zone system with wastewater discharge.**
356 Viruses specific to river water are depicted in blue, wastewater in red, beach sediment in green and shellfish in orange.

357

358 **Methods**

359 This work builds on our pilot study of a single wastewater treatment plant, and a downstream water
360 and sediment sampling site, in which we optimised methods and showed that we could reconstruct
361 RNA virus genomes from environmental samples [22].

362 **Sample collection.** We collected and processed four different types of samples for this study:
363 wastewater (influent and effluent), river and estuarine water, sediment and shellfish in June 2017
364 from the Conwy river catchment area located in North Wales (UK) (Fig. 1, Extended Data Table 1).

Wastewater influent was collected from the four major treatment plants in the catchment and the corresponding effluent from three (the Ganol 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. Finally, mussels (*Mytilus edulis*) were collected from the two main commercial shellfishery locations in the estuary and divided into eight pseudoreplicates, as described below.

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 [54]. 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). The system was decontaminated with Virkon® solution (Lanxess, Cologne, Germany) between each sample. Virus particles were then eluted from the 50 mL suspension (containing all particulate matter larger than 100 kDa) using beef extract and $NaNO_3$ to a final concentration of 3% and 2 M (pH 5.5), respectively. After incubation on ice for 30 minutes and centrifugation at 2500 x g for 10 minutes, the supernatant containing the eluted viruses was retained. The viruses were further concentrated by polyethylene glycol (PEG) precipitation. 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 [55].

For the mussel samples, approximately 200 mussels (*Mytilus edulis*) were collected from each location and stored on ice. Each mussel was dissected and the digestive tissue extracted and minced with a scalpel. The tissue was pooled per location and then divided into four replicates. Two replicates

per location were mixed with SM buffer (0.1 M NaCl, 50 mM Tris/HCl–pH 7.4, 10 mM $MgSO_4$) 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.

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 TrisHCl, 5 mM $MgSO_4$, 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 extra-viral 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 x 1 mL aliquots per replicate were mixed with 0.1 mm glass beads (MoBio) and lysed in a PowerLyser (MoBio) shaker (2 x 30 seconds at 3400 rpm). Debris was removed by centrifugation (5 min at 3200 x 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 et al. (2012)[56]. In short, 500 µL of sample was mixed with 1 mL of TRIzol reagent and 200 µl of molecular-grade chloroform in Phasemaker™ tubes (Invitrogen). These were shaken vigorously by hand for 10 seconds and centrifuged for 15 minutes at 13,000 x g in a benchtop microcentrifuge. 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 then precipitated by

adding an equal volume isopropanol and centrifugation at 13,000 x g for 30 minutes, followed by a wash with 70% ethanol. The pellet was air dried 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 (~ 10^5 particles/ml) was extracted, as was a negative Tris buffer control.

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.

Mussel digestive tissue was processed exactly as during RNA extraction (mixed with 0.1 mm glass beads and lysed in PowerLyzer) 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. The samples were then left to cool to

443 room temperature. Samples were transferred to PhaseLock tubes (VWR) for extraction.
444 Phenol/chloroform/isoamylalcohol (25:24:1) was added to each sample at equal volume, inverted to
445 mix and centrifuged for 5 minutes at 13000 x g in a benchtop microcentrifuge to separate the phases.
446 The aqueous phase was transferred to a new tube and the process was repeated at least once (twice
447 for turbid samples), followed by one round of chloroform phase separation. Finally, samples were
448 further cleaned and concentrated with ethanol precipitation (2.5 x volume 100% ethanol; 1/10 volume
449 3 M NaAc pH 5; incubation at -20°C for 30 minutes; precipitation 30 min at 15,000 x g, 4°C), washed
450 with 70% ethanol and air-dried in a laminar flow cabinet.
451 In tandem with the whole process, control samples were extracted, starting with the dilution
452 in NaCl. We used a negative control consisting of Tris buffer and a positive control consisting of 500 µl
453 stationary culture *Escherichia coli* MG1655 cells (RefSeq acc NC_000913), 2.2 x 10^8 pfu of Escherichia
454 phage T5 (RefSeq acc NC_005859) and 1.3 x 10^5 pfu of Escherichia phage vB_EcoP_phi24B (GenBank
455 acc HM208303).

456 **Sequencing.** Sequence library preparation and sequencing was performed by the Centre for Genomics
457 Research (CGR) NBAF facilities at the University of Liverpool, UK. RNA libraries were prepared as in
458 the pilot study [22] using the NEBNext Ultra directional RNA library preparation kit of Illumina with dual
459 indexes. During library preparation, the number of PCR cycles was increased to 30 to account for the
460 low amounts of input RNA (< 1 ng). Dual-indexed DNA libraries were generated using the NEBNext
461 Ultra II DNA Library Prep kit according to the manufacturer's instructions. Libraries were pooled and
462 sequenced on six lanes of the Illumina HiSeq 4000 generating paired-end 2 x 150 bp reads, three lanes
463 for the RNA libraries in July 2017 and three lanes for the DNA libraries in March 2018.
464 The RNA libraries gave a median number of paired reads of 50 million, with library RNA_TyCa
465 the lowest number of reads pairs (23 M) and RNA_LI the highest number (142 M). The DNA libraries
466 yielded at a median 33 M read pairs, ranging from 0 (the sequencing run failed for libraries DNA_SW2a,
467 DNA_TyCa/b, DNA_SBa/b) to 61 M (DNA_CS2). Unfortunately, we were unable to reconstruct the
468 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.

***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 [57,58]. 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 [59] 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 [60]. 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 [61], with the exception of the mussel 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 [62].

From these contigs, an Anvi'o contig database was created according to the instructions of the metagenomics workflow [63]. To be included in the database, contigs needed to meet the following criteria: RNA library assemblies (i) contig length min 1000 nucleotides (nt); (ii) amino acid similarity with any known virus; (iii) recruit no reads from control libraries; DNA library assemblies (i) contig length min 10,000 nt, (ii) identified by VirSorter as viral in categories 1 or 2 [64], (iii) recruit no reads from control libraries. VirSorter was run on all DNA contig sets using the microbiome decontamination mode on the iVirus Cyverse infrastructure [65]. 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 [14,66–69]. 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 [17] with the nr database downloaded from the NCBI in January 2018. The final database contained 10,149 UViGs (Uncultivated Viral Genomes, [14]) 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 [70].

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 [71]. The index was generated with "kallisto index" and the reads were mapped with "kallisto quant" using the --pseudobam flag to generate mapping files. 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 [72]. The resulting 10,149 by 58 matrix was visualised with Phantasus [73]. The pseudobam alignment files generated by kallisto were then transformed into Anvi'o profiles according to the metagenomics workflow instructions and investigated using the anvi-interactive interface [63]. Numbers of species detected per library, sample or sample type were calculated as the number of UViGs having a TPM value of minimum 10. Venn diagrams were produced on the online webserver http://bioinformatics.psb.ugent.be/webtools/Venn/ hosted by the VIB-UGent Center for Plant Systems Biology.

The taxonomic classifications by Kaiju as part of the Anvi'o platform left over 5000 UViGs unclassified. We then used diamond BLASTx against the viral RefSeq protein database (version 200, May 2020) and Megan 6 Community Edition to assign all UViGs to their most reliable taxonomic rank using the Megan 6 "long read" lowest common ancestor algorithm at the default settings [17,18]. The

taxonomic bin information was added to the Phantasus heatmaps by matching the UViG names and exported to R Studio to create graphs.

To generate phylogenetic trees of taxonomic groups of interest, we used the Megan 6 taxonomic bins. All UViGs assigned to a bin were annotated with Prokka [74] using the -kingdom Viruses setting and the predicted CDSs were manually curated in UGene [75] to adjust for the presence of polyproteins and missing start or stop codons from incomplete genomes. Per RNA virus taxonomic group, the RNA-dependent RNA polymerase (RdRP) amino acid sequences were extracted and aligned together with RdRP sequences from reference databases using MAFFT with maximum 5 iterations [42]. The resulting alignments were trimmed with TrimAl [41] using the -gappyout setting, followed by manual inspection in the UGene alignment viewer. Sequences missing the conserved structural motifs present in RdRPs [76] were removed, as were sequences missing more than 50% of the trimmed sites. Trees were computed using the IQ-Tree suite [39] including calculation of the best substitution model with ModelFinder [43], calculation of the approximate likelihood ratio test (1000 repetitions) [77] and ultrafast bootstrap approximation with UFBOOT2 (1000 repetitions) [44]. The resulting trees were analysed and annotated in iTOL [78]. For the picorna-calici tree, the alignments generated by Shi and colleagues were additionally used as references [21,79].

Rotavirus segment genotyping was performed on the RotaC 2.0 webserver of the Rega Institute (KU Leuven, Belgium) [80].

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).

Author contributions

E.M.A., K.F., D.L.J., J.E.M., H.E.A., and A.J.M. designed the study; E.M.A, K.F, J.E.M. and D.L.J collected samples; E.M.A. and K.F. performed the experiments; E.M.A. analyzed the data; E.M.A. wrote the manuscript and prepared the manuscript for submission. All authors critically reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Sidhu, J. P. S., Sena, K., Hodgers, L., Palmer, A. & Toze, S. Comparative enteric viruses and coliphage removal during wastewater treatment processes in a sub-tropical environment. *Sci. Total Environ.* **616**, 669–677 (2017).

2. Girones, R. *et al.* Molecular detection of pathogens in water - The pros and cons of molecular techniques. *Water Res.* **44**, 4325–4339 (2010).

3. Da Silva, A. K. *et al.* Evaluation of removal of noroviruses during wastewater treatment, using real-time reverse transcription-PCR: Different behaviors of genogroups I and II. *Appl. Environ. Microbiol.* **73**, 7891–7897 (2007).

4. Qiu, Y. *et al.* Assessment of human virus removal during municipal wastewater treatment in Edmonton, Canada. *J. Appl. Microbiol.* **119**, 1729–1739 (2015).

5. Hellmér, M. *et al.* Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. *Appl. Environ. Microbiol.* **80**, 6771–6781 (2014).

6. Farkas, K. *et al.* Seasonal and diurnal surveillance of treated and untreated wastewater for human enteric viruses. *Environ. Sci. Pollut. Res.* 33391–33401 (2018). doi:10.1007/s11356-018-3261-y

7. Fong, T. T., Phanikumar, M. S., Xagoraraki, I. & Rose, J. B. Quantitative detection of human adenoviruses in wastewater and combined sewer overflows influencing a Michigan river. *Appl. Environ. Microbiol.* **76**, 715–723 (2010).

8. Kitajima, M., Iker, B. C., Pepper, I. L. & Gerba, C. P. Relative abundance and treatment reduction of viruses during wastewater treatment processes—Identification of potential viral indicators. *Sci. Total Environ.* **488**, 290–296 (2014).

9. Prado, T. *et al.* Performance of wastewater reclamation systems in enteric virus removal. *Sci. Total Environ.* **678**, 33–42 (2019).

10. Gomes, J., Frasson, D., Quinta-Ferreira, R., Matos, A. & Martins, R. Removal of Enteric Pathogens from Real Wastewater Using Single and Catalytic Ozonation. *Water* **11**, 127 (2019).

11. Farkas, K. *et al.* Evaluation of Two Triplex One-Step qRT-PCR Assays for the Quantification of Human Enteric Viruses in Environmental Samples. *Food Environ. Virol.* **9**, 342–349 (2017).

12. Farkas, K., Mannion, F., Hillary, L. S., Malham, S. K. & Walker, D. I. Emerging technologies for the rapid detection of enteric viruses in the aquatic environment. *Current Opinion in Environmental Science and Health* **16**, 1–6 (2020).

13. DiCaprio, E. Recent advances in human norovirus detection and cultivation methods. *Curr. Opin. Food Sci.* **14**, 93–97 (2017).

14. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29–37 (2019).

15. Robins, P. E., Farkas, K., Cooper, D., Malham, S. K. & Jones, D. L. Viral dispersal in the coastal zone: A method to quantify water quality risk. *Environ. Int.* **126**, 430–442 (2019).

16. Perkins, T. L. *et al.* Sediment composition influences spatial variation in the abundance of human pathogen indicator bacteria within an estuarine environment. *PLoS One* **9**, (2014).

17. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

18. Huson, D. H. & Weber, N. Microbial Community Analysis Using MEGAN. in *Methods in enzymology* **531**, 465–485 (2013).

19. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol. Mol. Biol. Rev.* **84**, 1–33 (2020).

20. Gorbalenya, A. E. *et al.* The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.* **5**, 668–674 (2020).

21. Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature* **540**, 1–12 (2016).

22. Adriaenssens, E. M. *et al.* Viromic Analysis of Wastewater Input to a River Catchment Reveals a Diverse Assemblage of RNA Viruses. *mSystems* **3**, e00025-18 (2018).

23. Finkbeiner, S. R., Kirkwood, C. D. & Wang, D. Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea. *Virol. J.* **5**, 117 (2008).

24. FSA. *Estimating Quality Adjusted Life Years and Willingness to Pay Values for Microbiological Foodborne Disease (Phase 2)*. (2017).

611 25. Kirk, M. D. *et al.* World Health Organization Estimates of the Global and Regional Disease
612 Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010: A Data Synthesis. *PLoS*
613 *Med.* **12**, 1–21 (2015).

614 26. Ahmed, S. M. *et al.* Global prevalence of norovirus in cases of gastroenteritis: A systematic
615 review and meta-analysis. *Lancet Infect. Dis.* **14**, 725–730 (2014).

616 27. Diez-Valcarce, M. *et al.* Genetic diversity of human sapovirus across the Americas. *J. Clin. Virol.*
617 **104**, 65–72 (2018).

618 28. Farkas, K. *et al.* Seasonal and spatial dynamics of enteric viruses in wastewater and in riverine
619 and estuarine receiving waters. *Sci. Total Environ.* **634**, 1174–1183 (2018).

620 29. Inns, T. *et al.* What proportion of care home outbreaks are caused by norovirus? An analysis of
621 viral causes of gastroenteritis outbreaks in care homes, North East England, 2016-2018. *BMC*
622 *Infect. Dis.* **20**, 1–8 (2019).

623 30. Brown, J. R., Shah, D. & Breuer, J. Viral gastrointestinal infections and norovirus genotypes in a
624 paediatric UK hospital, 2014–2015. *J. Clin. Virol.* **84**, 1–6 (2016).

625 31. Zell, R. *et al.* ICTV Virus Taxonomy Profile: Picornaviridae. *J. Gen. Virol.* **98**, 2421–2422 (2017).

626 32. Tapparel, C., Siegrist, F., Petty, T. J. & Kaiser, L. Picornavirus and enterovirus diversity with
627 associated human diseases. *Infect. Genet. Evol.* **14**, 282–293 (2013).

628 33. Amar, C. F. L. *et al.* Detection by PCR of eight groups of enteric pathogens in 4,627 faecal
629 samples: Re-examination of the English case-control Infectious Intestinal Disease Study (1993-
630 1996). *Eur. J. Clin. Microbiol. Infect. Dis.* **26**, 311–323 (2007).

631 34. Noel, J. S. *et al.* Parkville virus: A novel genetic variant of human calicivirus in the Sapporo virus
632 clade, associated with an outbreak of gastroenteritis in adults. *J. Med. Virol.* **52**, 173–178

(1997).

35. Vinjé, J. *et al.* Molecular detection and epidemiology of Sapporo-like viruses. *J. Clin. Microbiol.* **38**, 530–536 (2000).

36. Varela, M. F. *et al.* Sapovirus in wastewater treatment plants in Tunisia: Prevalence, removal, and genetic characterization. *Appl. Environ. Microbiol.* **84**, (2018).

37. Pang, X. *et al.* Prevalence, levels and seasonal variations of human enteric viruses in six major rivers in Alberta, Canada. *Water Res.* **153**, 349–356 (2019).

38. Mann, Pietsch & Liebert. Genetic Diversity of Sapoviruses among Inpatients in Germany, 2008−2018. *Viruses* **11**, 726 (2019).

39. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

40. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2–5 (2019). doi:10.1093/nar/gkz239

41. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

42. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

43. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

44. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

655 45. Anisomova, M. & Gascuel, O. Approximate Likelihood-Ratio Test for Branches : A Fast ,
656 Accurate ,. *Syst. Biol.* **55**, 539–552 (2006).

657 46. *Epidemiology and prevention of vaccine-preventable diseases*. (Centers for Disease Control and
658 Prevention, 2015).

659 47. Dayaram, A. *et al.* Diverse small circular DNA viruses circulating amongst estuarine molluscs.
660 *Infect. Genet. Evol.* **31**, 284–295 (2015).

661 48. Zawar-Reza, P. *et al.* Diverse small circular single-stranded DNA viruses identified in a
662 freshwater pond on the McMurdo Ice Shelf (Antarctica). *Infect. Genet. Evol.* (2014).
663 doi:10.1016/j.meegid.2014.05.018

664 49. Phan, T. G. *et al.* Small circular single stranded DNA viral genomes in unexplained cases of
665 human encephalitis, diarrhea, and in untreated sewage. *Virology* **482**, 98–104 (2015).

666 50. Zhang, T. *et al.* RNA viral community in human feces: Prevalence of plant pathogenic viruses.
667 *PLoS Biol.* **4**, 0108–0118 (2006).

668 51. Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J. & Breitbart, M. Pepper mild mottle virus
669 as an indicator of fecal pollution. *Appl. Environ. Microbiol.* **75**, 7261–7267 (2009).

670 52. Maalouf, H. *et al.* Distribution in tissue and seasonal variation of norovirus genogroup I and II
671 ligands in oysters. *Appl. Environ. Microbiol.* **76**, 5621–5630 (2010).

672 53. Whitman, R. L. *et al. Microbes in beach sands: Integrating environment, ecology and public*
673 *health*. *Reviews in Environmental Science and Biotechnology* **13**, (2014).

674 54. Farkas, K., McDonald, J. E., Malham, S. K. & Jones, D. L. Two-step concentration of complex
675 water samples for the detection of viruses. *Methods Protoc.* **1**, 35 (2018).

676 55. Farkas, K., Hassard, F., McDonald, J. E., Malham, S. K. & Jones, D. L. Evaluation of molecular

methods for the detection and quantification of pathogen-derived nucleic acids in sediment. *Front. Microbiol.* **8**, 53 (2017).

56. Kroger, C. *et al.* The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc. Natl. Acad. Sci.* **109**, E1277–E1286 (2012).

57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

58. Joshi, N. & Fass, J. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. (2011).

59. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).

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

61. Nurk, S. *et al.* Assembling genomes and mini-metagenomes from highly chimeric reads. in *Research in Computational Molecular Biology. RECOMB 2013. Lecture Notes in Computer Science* (eds. Deng, M., Jiang, R., Sun, F. & Zhang, X.) **7821**, 158–170 (Springer, Berlin, Heidelberg, 2013).

62. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

63. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).

64. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

65. Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B. L. & Sullivan, M. B. iVirus: facilitating new

699 insights in viral ecology with software and community data sets imbedded in a
700 cyberinfrastructure. *ISME J.* **11**, 7–14 (2017).

701 66. Adriaenssens, E. & Brister, J. R. How to name and classify your phage: An informal guide.
702 *Viruses* **9**, 70 (2017).

703 67. Gregory, A. C. *et al.* Genomic differentiation among wild cyanophages despite widespread
704 horizontal gene transfer. *BMC Genomics* **17**, 930 (2016).

705 68. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean
706 viruses. *Nature* **537**, 689–693 (2016).

707 69. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat.*
708 *Microbiol.* (2018). doi:10.1038/s41564-018-0190-y

709 70. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics
710 with Kaiju. *Nat. Commun.* **7**, 1–9 (2016).

711 71. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
712 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

713 72. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level
714 estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2016).

715 73. Zenkova, D., Kamenev, V., Sablina, R., Artyomov, M. & Sergushichev, A. Phantasus: visual and
716 interactive gene expression analysis. (2018). doi:10.18129/B9.bioc.phantasus

717 74. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069
718 (2014).

719 75. Okonechnikov, K. *et al.* Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **28**,
720 1166–1167 (2012).

721 76. Venkataraman, S., Prasad, B. & Selvarajan, R. RNA Dependent RNA Polymerases: Insights from
722 Structure, Function and Evolution. *Viruses* **10**, 76 (2018).

723 77. Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support
724 methods demonstrates accuracy, power, and robustness of fast likelihood-based
725 approximation schemes. *Syst. Biol.* **60**, 685–699 (2011).

726 78. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display
727 and annotation. *Bioinformatics* **23**, 127–128 (2007).

728 79. Shi, M. *et al.* The evolutionary history of vertebrate RNA viruses. *Nature* **556**, 197–202 (2018).

729 80. Maes, P., Matthijnssens, J., Rahman, M. & Ranst, M. Van. RotaC : A web-based tool for the
730 complete genome classification of group A rotaviruses. **4**, 2–5 (2009).

731

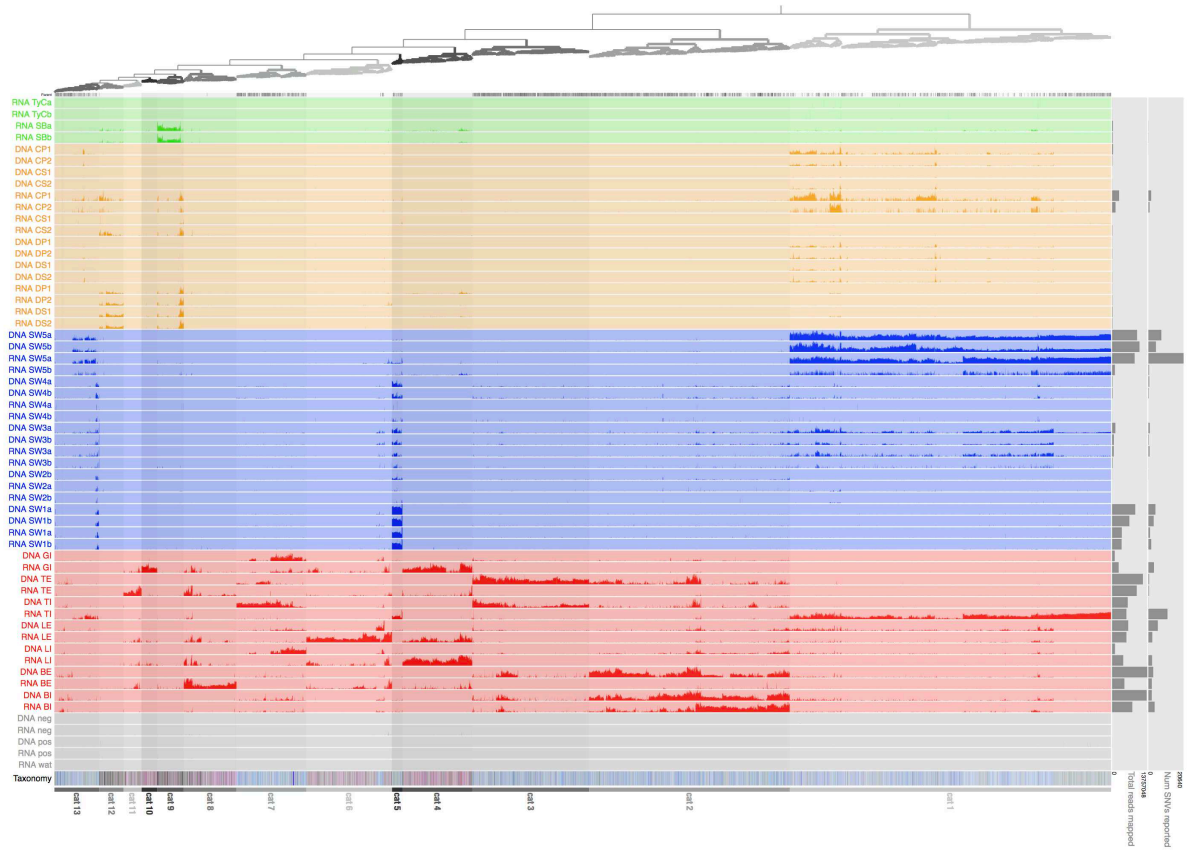
732

Extended data

Extended Data Table 1: Summary of sample distribution in the Conwy water catchment.

Sample code[a]	Sample type	Location	Coordinates	Volume
BI	WW influent	Betws-y-Coed WWTP	53°05'44.0"N 3°48'01.8"W	2 x 0.5 L
BE	WW effluent	Betws-y-Coed WWTP	53°05'44.0"N 3°48'01.8"W	2 x 0.5 L
LI	WW influent	Llanrwst WWTP	53°08'24.4"N 3°48'12.8"W	2 x 0.5 L
LE	WW effluent	Llanrwst WWTP	53°08'24.4"N 3°48'12.8"W	2 x 0.5 L
TE	WW effluent	Tal-y-Bont WWTP	53°12'07.7"N 3°50'20.6"W	2 x 0.5 L
TI	WW influent	Tal-y-Bont WWTP	53°12'07.7"N 3°50'20.6"W	2 x 0.5 L
GI	WW influent	Ganol WWTP	53°16'43.6"N 3°47'32.9"W	2 x 0.5 L
SW1a/b	River water	Upstream Betws WWTP	53°05'32.2"N 3°47'56.9"W	4 x 50 L
SW2a/b	River water	Between Betws and Llanrwst	53°08'13.2"N 3°47'51.2"W	4 x 50 L
SW3a/b	River water - tidal edge	Downstream Llanrwst WWTP	53°08'35.1"N 3°48'24.9"W	4 x 50 L
SW4a/b	River water - within tidal limit	Tal-y-Cafn	53°13'45.1"N 3°49'12.2"W	4 x 50 L
SW5a/b	River water - estuary	Morfa beach	53°17'37.7"N 3°50'22.2"W	4 x 50 L
TyCa/b	Sediment	Tal-y-Cafn	53°13'45.1"N 3°49'12.2"W	4 x 50 g
DS1/2; DP1/2	Shellfish	Deganwy shellfish bed	53°18'29.8"N 3°50'36.0"W	8 x 25 g
CS1/2; CP1/2	Shellfish	Conwy shellfish bed	53°17'50.8"N 3°50'51.1"W	8 x 25 g
SBa/b	Sediment	Morfa bathing beach	53°17'37.7"N 3°50'22.2"W	4 x 75 g

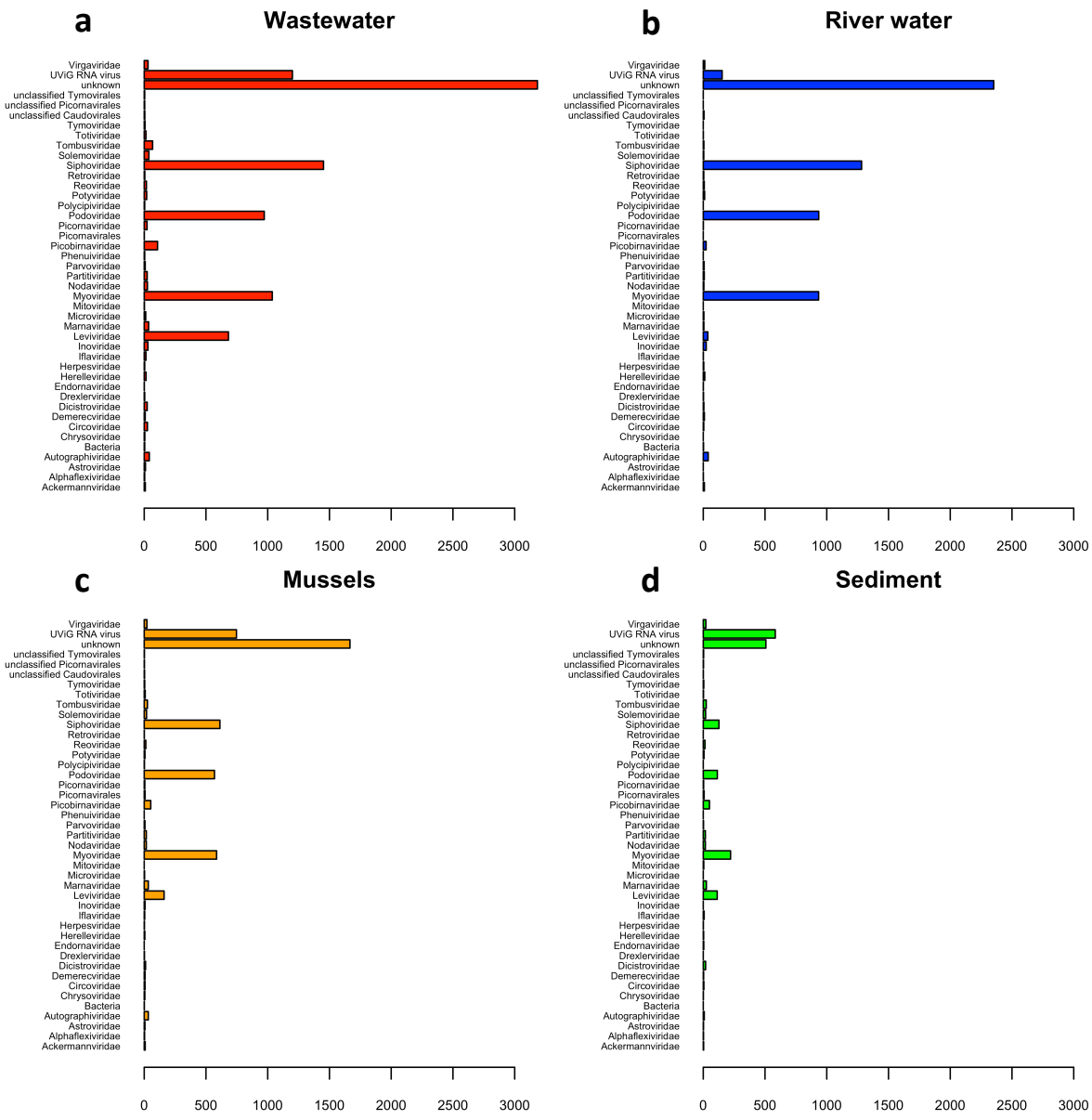
[a] Biological replicates originating from the same sampling location are indicated with small letters a and b, while technical replicates of the pooled mussel samples are indicated as follows: Deganwy or Conwy, extraction with SM or PBS buffer, replicate 1 or 2. For each replicate, two libraries were constructed, an RNA and a DNA library, and indicated as such in the sample code. WW denotes wastewater. WWTP denotes wastewater treatment plant.



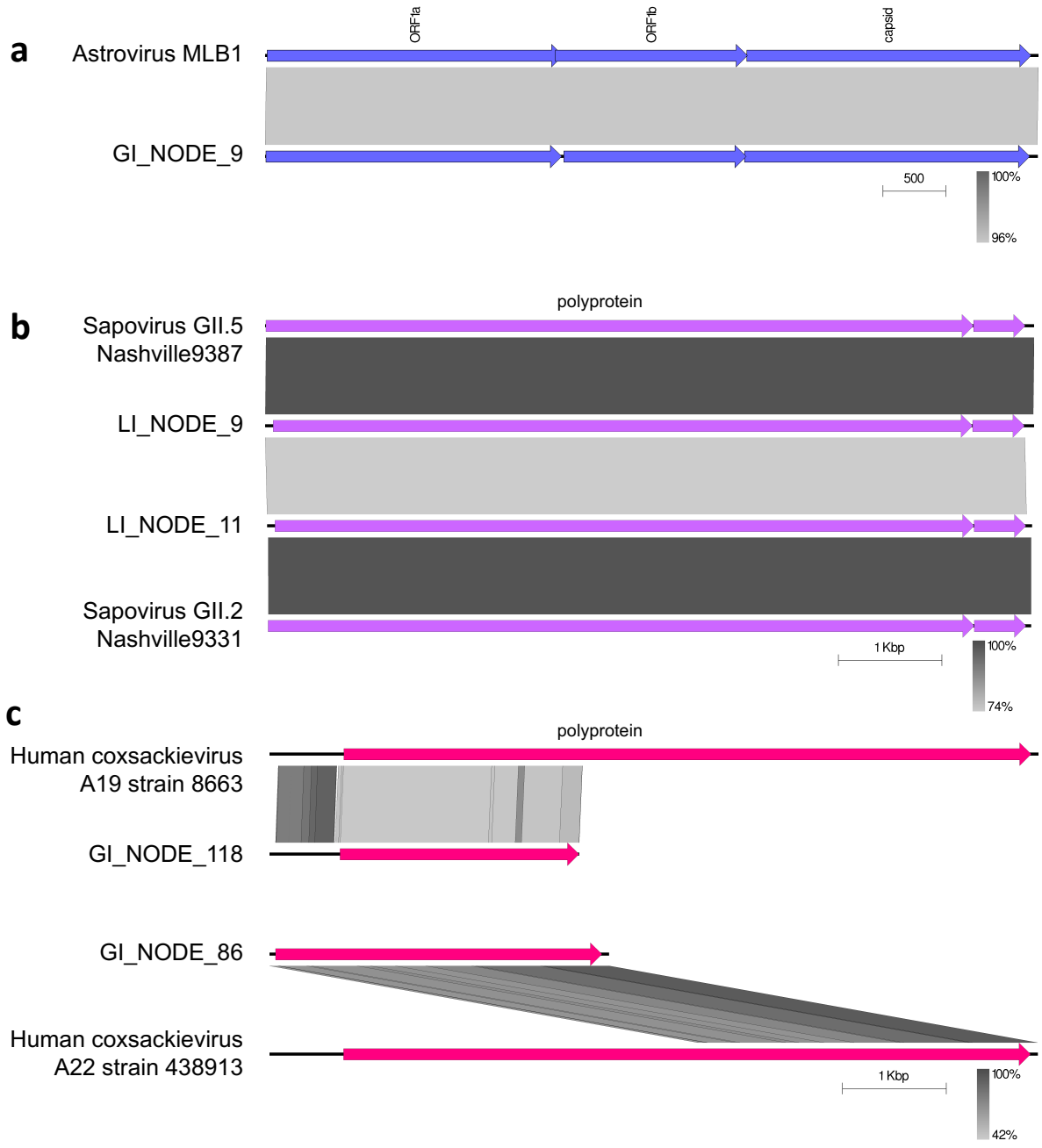
741

742 **Extended Data Figure 1 | Differential patterns of abundance of each viral genome (UViG) along the wastewater**
743 **impacted Conwy river and coastal zone, including positive and negative controls.** Anvi'o - mean coverage per
744 contig (split). Each row is a sequencing library, coloured by its sample type (green = sediment; orange = mussels;
745 blue = river/estuary water; red = wastewater, grey = controls). Each column (leaf in top dendrogram) is a contig
746 or a split of a contig (in cases where contigs were larger than 11 kb). The height of the bar in each row is the log
747 mean coverage across the contig or contig split length. The contigs are clustered (top dendrogram) according to
748 their sequence composition and differential coverage using Euclidean distance and Ward linkage. Based on this
749 clustering, we identified 13 categories of UViGs, indicated by shades of grey in the dendrogram and numbered
750 at the bottom of the plot. The bottom row represents the taxonomy assigned by Kaiju (using its viral database)
751 to the predicted genes in each contig. Contigs without assigned taxonomy are depicted in grey, dsDNA
752 bacteriophages in shades of blue, other dsDNA viruses in shades of green, ssDNA viruses in shades of yellow,
753 RNA (ds, (+)ss, (-)ss) in shades of purple/red. The right hand panels the number of single nucleotide variants
754 (SNVs) found after read mapping (0-20,640) and the total number of reads mapped to contigs (0-13,757,048).

755



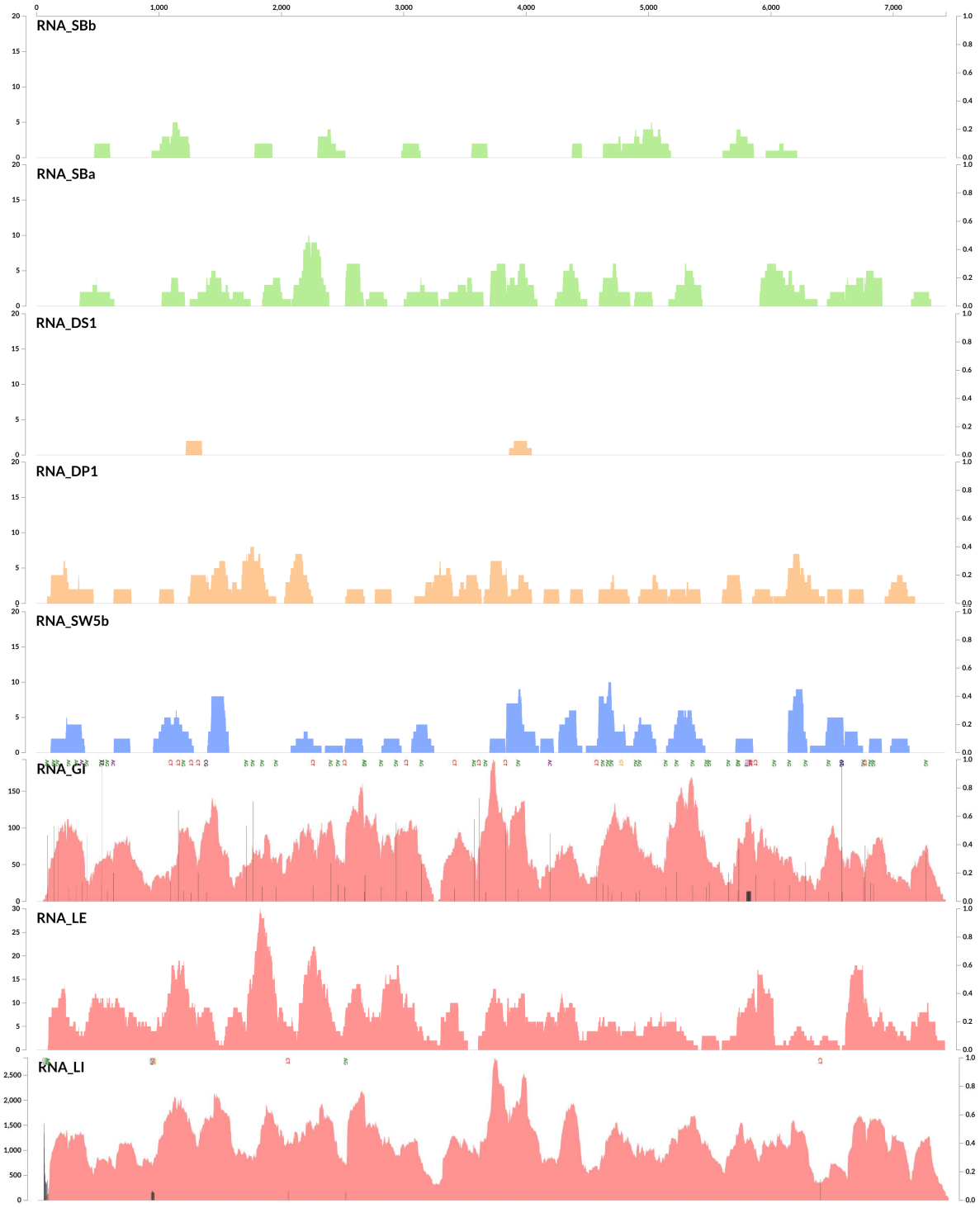
Extended Data Figure 2 | Number of UViGs detected per environment and classified into family-level taxonomic groups. The cut-off for detection was 10 TPM summed per environment, a) wastewater, b) river and estuary river water, c) mussel digestive tissue, d) beach sediment.



Extended Data Figure 3 | UViG genome comparison with closest relative in sequence database. The figures were generated with Easyfig, displaying annotated ORFs as arrows and tBLASTx-based pairwise genome identity in shades of grey. a) UViG GI_NODE_9 compared with astrovirus MLB1; b) UViGs LI_NODE_9 and LI_NODE_11 compared with each other and the sapoviruses GII.5 Nashville9387 and GII.2 Nashville9331, respectively; c) partial UViGs GI_NODE_118 and GI_NODE_86 compared with human coxsackieviruses A19 strain 8663 and A22 strain 43913, respectively.

LI_NODE_9_length_7431_cov_255.051_split_00001 detailed

Sapovirus



Extended Data Figure 4 | Anvi'o read mapping inspection panel for UViG LI_NODE_9, a predicted sapovirus. The genome was detected at low coverage in beach sediment RNA libraries (RNA_SBb, RNA_SBa), shellfish libraries (RNA_DS1, RNA_DP1) and one river water library (RNA_SW5b). It was detected at high coverage in the wastewater libraries RNA_GI, RNA_LE and RNA_LI, with the RNA_GI mapping showing the presence of multiple single nucleotide variants, indicating that a different strain was present in the Ganol wastewater treatment plant than in the Llanrwst plant.

Figures

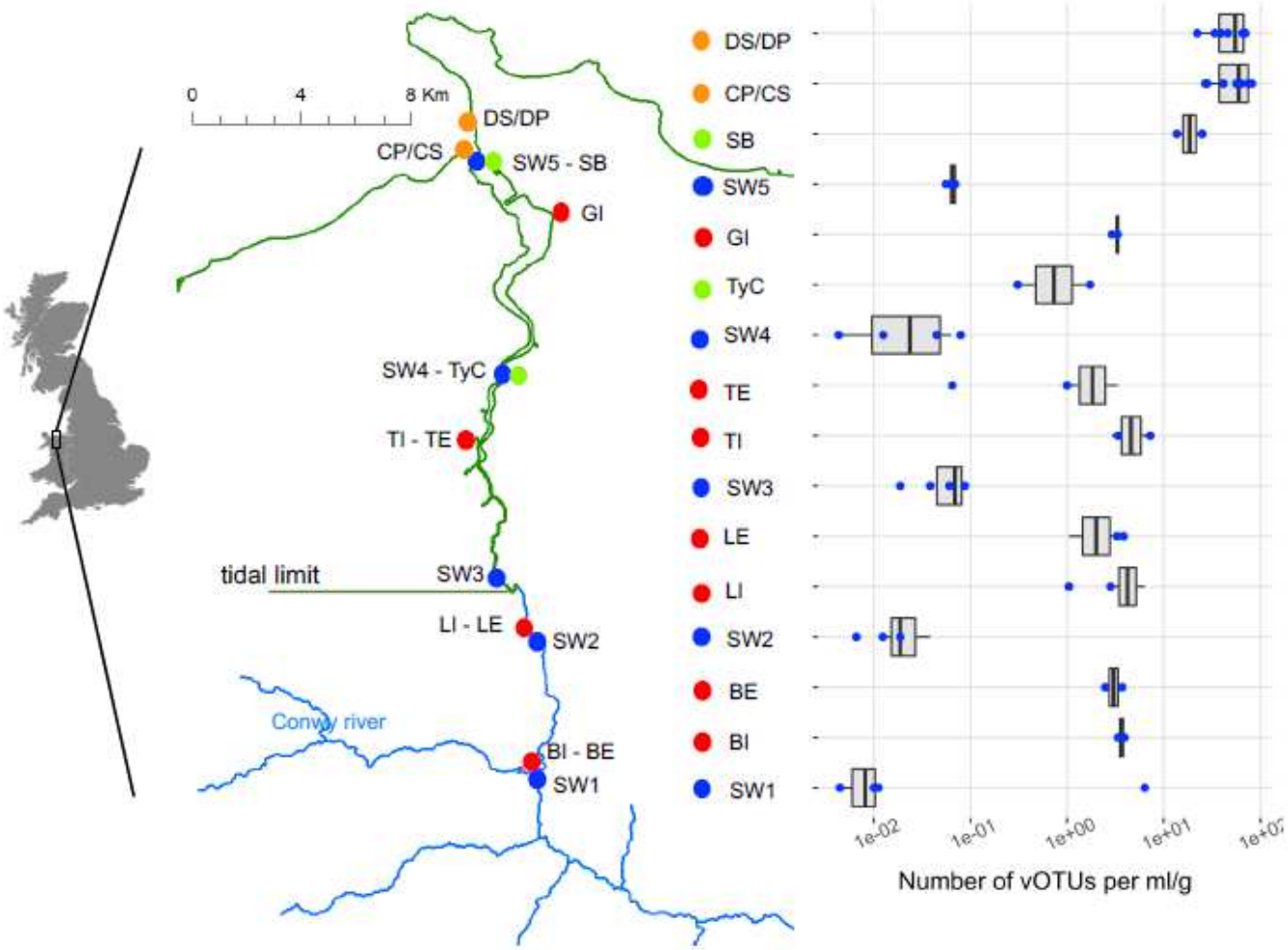


Figure 1

Viral abundance along the wastewater impacted Conwy river catchment and coastal zone. Left: Schematic of the Conwy river catchment with sampling sites designated by colour-coded dots (red – wastewater, blue – surface water, green – sediment, orange – shellfish). The section of the river within the tidal limit is designated in green. Map of Great Britain by Free Vector Maps. Right: Boxplot representation of the number of species (vOTUs) detected in each sample per ml or g of sample extracted, composed of RNA and DNA libraries and biological replicates, species numbers for single libraries in blue dots

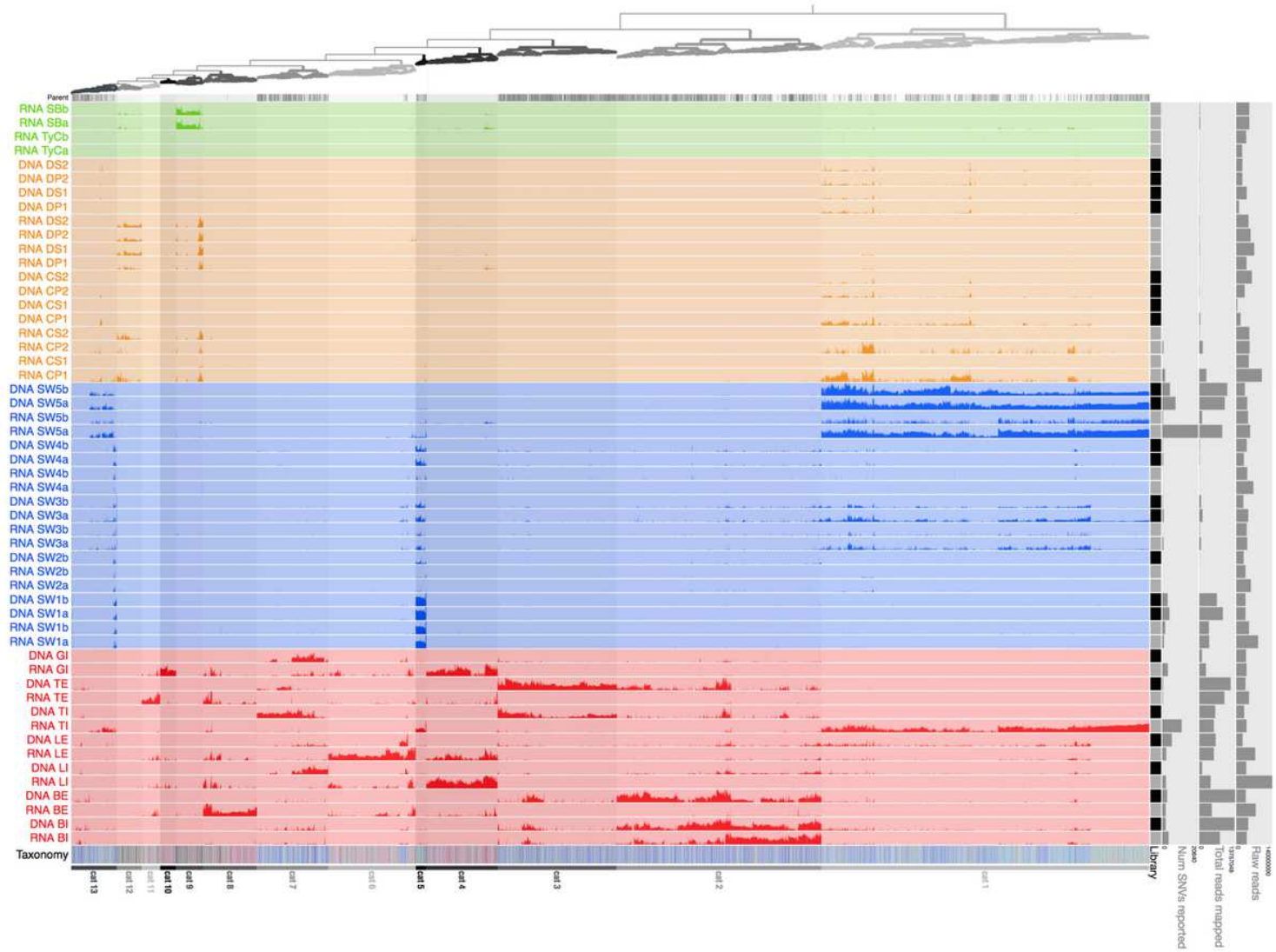


Figure 2

Differential patterns of abundance of each viral genome (UViG) along the wastewater impacted Conwy river and coastal zone. Anvi'o - mean coverage per contig (split). Each row is a sequencing library, coloured by its sample type (green = sediment; orange = mussels; blue = river/estuary water; red = wastewater). Each column (leaf in top dendrogram) is a contig or a split of a contig (in cases where contigs were larger than 11 kb). The height of the bar in each row is the log mean coverage across the contig or contig split length. The contigs are clustered (top dendrogram) according to their sequence composition and differential coverage using Euclidean distance and Ward linkage. Based on this clustering, we identified 13 categories of UViGs, indicated by shades of grey in the dendrogram and numbered at the bottom of the plot. The bottom row represents the taxonomy assigned by Kaiju (using its viral database) to the predicted genes in each contig. Contigs without assigned taxonomy are depicted in grey, dsDNA bacteriophages in shades of blue, other dsDNA viruses in shades of green, ssDNA viruses in shades of yellow, RNA (ds, (+)ss, (-)ss) in shades of purple/red. The right hand panels show the library type (RNA = grey; DNA = black), the number of single nucleotide variants (SNVs) found after read

mapping (0-20,640), the total number of reads mapped to contigs (0-13,757,048) and the total number of raw sequencing reads (before QC and contamination screen; 0-140,000,000).

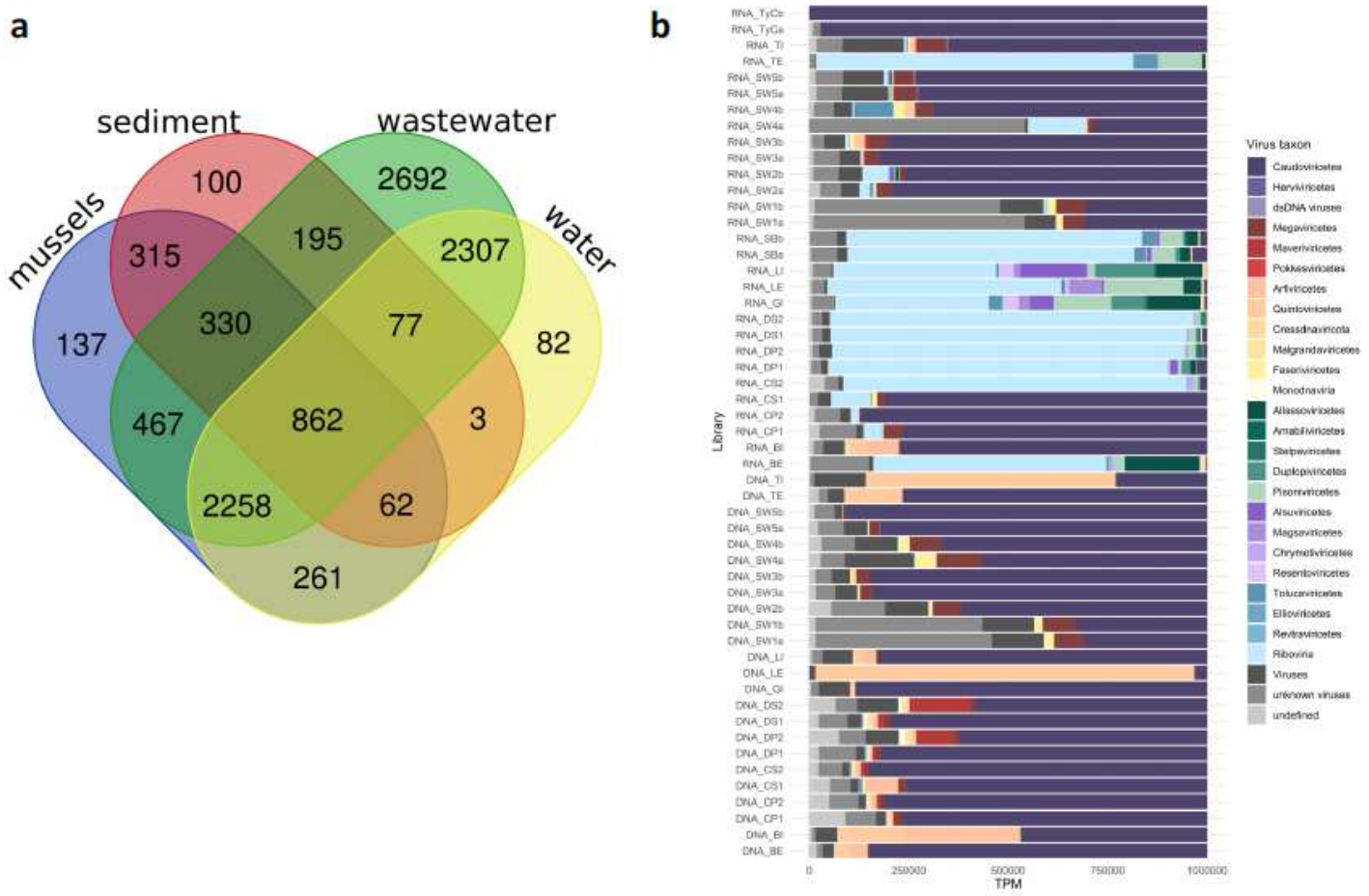


Figure 3

Commonality and taxonomic composition of viral genomes (UViG) in samples types from the wastewater impacted Conwy river and coastal zone. a, Venn diagram representation of the number of UViGs shared between different a b environment types (min 10 TPM for detection). b, Relative abundances of the UViGs at the virus class level per sequencing library, normalized per library as transcipts (=contig) per million (TPM). dsDNA viruses in shades of dark purple and red; ssDNA viruses in shades of pink and yellow; RNA viruses in shades of green, purple and blue; unknown viruses in shades of grey.

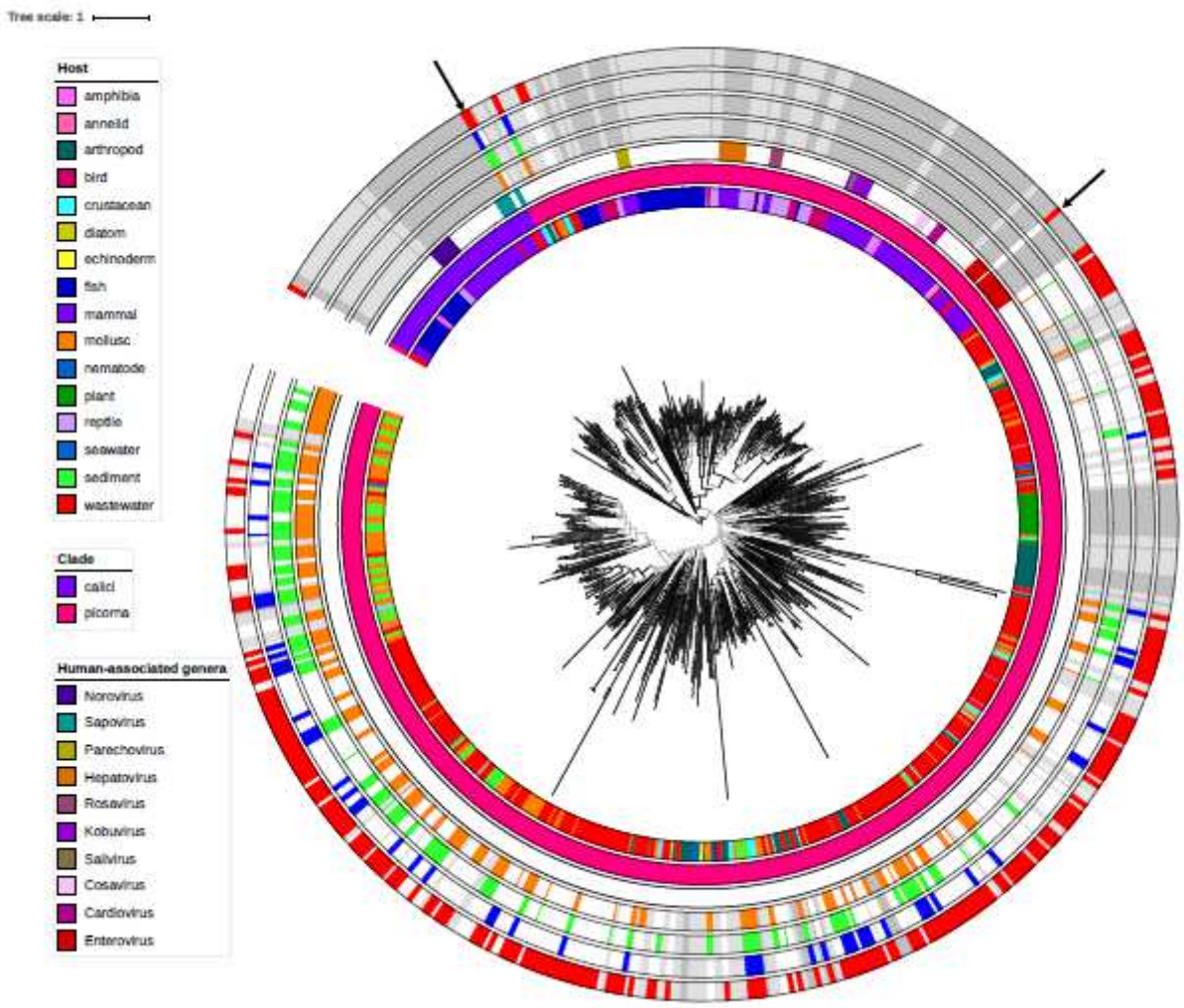


Figure 4

Maximum likelihood phylogenetic tree of the RdRP amino acid sequences of viruses/genomes assigned to the family Caliciviridae and the order Picornavirales built with IQ-TREE 39 and visualized with ITOL 40. The multiple alignment consisted of 622 sequences and 695 amino acid sites, aligned using MAFFT and trimmed with Trimal 41,42. The best fit model was LG+F+R10 as determined with ModelFinder 43. Branch support was calculated using the Shimodaira Hasegawa – approximate Likelihood Ratio Test (SH-aLRT) and the UFBoot (ultrafast bootstrap) algorithm on 1000 replications with nodes below 80% (SH-aLRT) and 95% (UFBoot) indicated in grey 44,45. The three inner colour strips from inside to outside indicate respectively: viral host or metagenome the RdRP was extracted from, predicted clade, human-associated genera (only reference genomes from human pathogenic viruses coloured). The four outside colours strips indicate detection in shellfish samples (orange), beach/river sediment samples (green), river/estuarine water samples (blue) and wastewater samples (red), with other virome-derived UViGs in light grey and reference virus sequences in middle grey. The black arrows indicate the UViGs found in this study that are likely human pathogens.

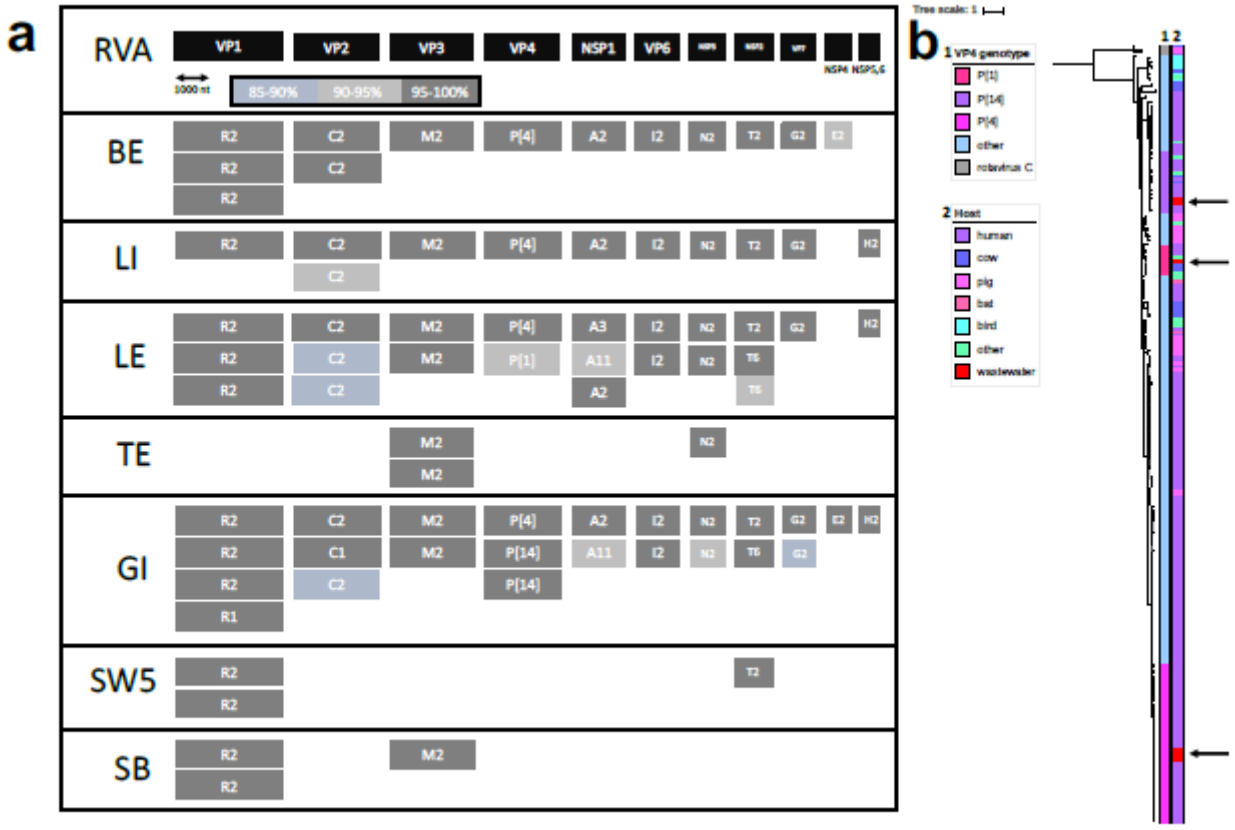


Figure 5

Rotavirus A (RVA) in the virome datasets. a, The 11 segments of the reference genome of RVA ranked by size in black. RVA segments recovered per sample below showing the predicted genotype of the segment and the percentage of nucleotide identity with a representative of that genotype as calculated by the RotaC 2.0 tool. b, Maximum likelihood phylogenetic tree of the VP4 amino acid sequences of selected representatives of all RVA genotypes build with IQ-TREE 39 and visualized with ITOL 40. The multiple alignment consisted of 253 sequences and 774 amino acid sites, aligned using MAFFT and trimmed with Trimal 41,42. The best fit model was FLU+F+R8 as determined with ModelFinder 43. Branch support was calculated using the UFBoot (ultrafast bootstrap) algorithm on 1000 bootstraps and is indicated with branch colours in shades of grey, with support values higher than 95% in black 44. Colour strip 1 indicates the genotype clustering, using RVC isolates as outgroup. Colour strip 2 shows the host of the isolates with arrows indicating the virome-derived sequences.

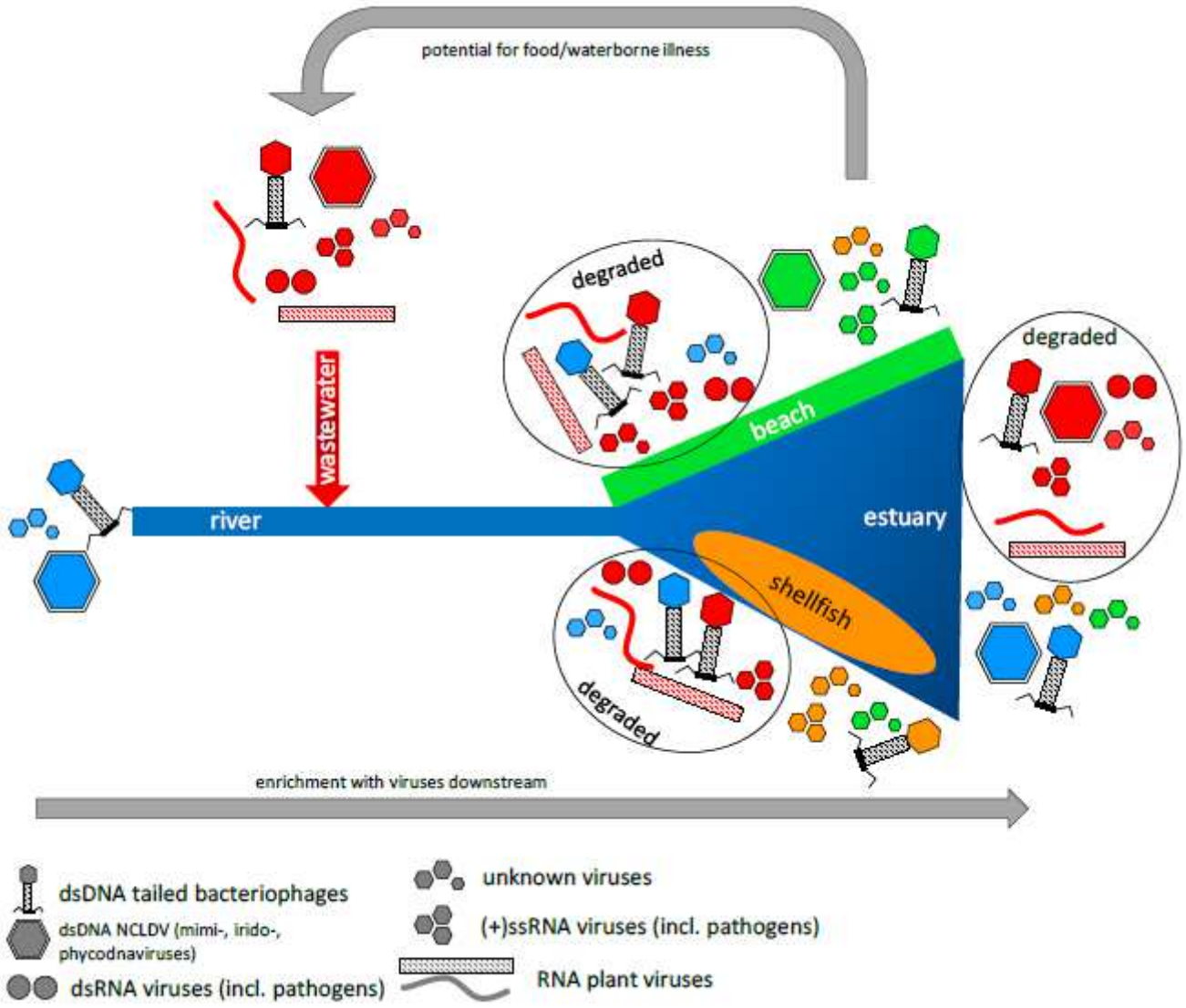


Figure 6

Model for the circulation of viruses in a river catchment and coastal zone system with wastewater discharge. Viruses specific to river water are depicted in blue, wastewater in red, beach sediment in green and shellfish in orange.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1.png
- FigS2.png
- FigS3final.jpg
- FigS4final.jpg

- TableS1final.jpg