RUNNING TITLE: Environmental viromics for the detection of pathogens

Viromic analysis of wastewater input to a river catchment reveals a diverse assemblage of RNA viruses

Evelien M. Adriaenssens[1,*], Kata Farkas[2], Christian Harrison[1], David Jones[2], Heather E. Allison[1], Alan J. McCarthy[1]

[1] Microbiology Research Group, Institute of Integrative Biology, University of Liverpool, UK

[2] School of Environment, Natural Resources and Geography, Bangor University, Bangor, LL57 2UW, UK

* Corresponding author: evelien.adriaenssens@liv.ac.uk

Abstract

Detection of viruses in the environment is heavily dependent on PCR-based approaches that require reference sequences for primer design. While this strategy can accurately detect known viruses, it will not find novel genotypes, nor emerging and invasive viral species. In this study, we investigated the use of viromics, i.e. high-throughput sequencing of the biosphere viral fraction, to detect human/animal pathogenic RNA viruses in the Conwy river catchment area in Wales, UK. Using a combination of filtering and nuclease treatment, we extracted the viral fraction from wastewater, estuarine river water and sediment, followed by RNASeq analysis on the Illumina HiSeq platform for the discovery of RNA virus genomes. We found a higher richness of RNA viruses in wastewater samples than in river water and sediment, and assembled a complete norovirus GI.2 genome from wastewater effluent, which was not contemporaneously detected by conventional qRT-PCR. To our knowledge, this is the first environmentally-derived norovirus genome sequence to be available from a public database. The simultaneous presence of diverse rotavirus signatures in wastewater indicated the potential for zoonotic infections in the area and suggested run-off from pig farms as the origin of these viruses. Our results show that viromics can be an important tool in the discovery of pathogenic viruses in the environment and can be used to inform and optimize reference-based detection methods provided appropriate and rigorous controls are included.

Importance

Enteric viruses cause gastro-intestinal illness and are commonly transmitted through the faecal-oral route. When wastewater is released into river systems, these viruses can contaminate the environment. Our results show that we can use viromics to find

37 the range of potentially pathogenic viruses that are present in the environment and
38 identify prevalent genotypes. The ultimate goal is to trace the fate of these
39 pathogenic viruses from origin to the point where they are a threat to human health,
40 informing reference-based detection methods and water quality management.

41 Introduction

42 Pathogenic viruses in water sources are likely to originate primarily from
43 contamination with sewage. Classic marker bacteria used for faecal contamination
44 monitoring, such as *Escherichia coli* and *Enterococcus* spp., are not, however, good
45 indicators for the presence of human enteric viruses (1). The virus component is
46 often monitored using qPCR approaches, which can give information on relative
47 abundance of specific viruses and their genotype, but only those that are both known
48 and characterised (2). Viruses commonly targeted in sewage contamination assays
49 include noroviruses (3), hepatitis viruses (4), enteroviruses (5), and various
50 adenoviruses (6, 7). Viral monitoring in sewage has previously yielded positive
51 results for norovirus, sapovirus, astrovirus, and adenovirus, indicating that people
52 are shedding viruses that are not necessarily detected in a clinical setting (8). This
53 same study found a spike in norovirus genogroup GII sequence signatures in
54 sewage two to three weeks before the outbreak of associated disease was reported
55 in hospitals and nursing homes. The suggestion, therefore, is that environmental
56 viromics can provide an early warning of disease outbreaks, in addition to the
57 monitoring of virus dissemination in watercourses.

58 Recent reviews have proposed the use of viral metagenomics or viromic approaches
59 as an alternative method to test for the presence of pathogenic viruses in the
60 environment (2, 9, 10). Provided the entire viral community is sampled and

61 sequenced, novel genotypes or even entirely novel viruses can be detected.
62 Potential new viral markers for faecal contamination have already been revealed,
63 such as pepper mild mottle virus and crAssphage (11, 12), among the huge diversity
64 of human viruses found in sludge samples (13).

65 In this pilot study, we have used viromics to investigate the presence of human
66 pathogenic RNA viruses in wastewater, estuarine surface water and sediment in a
67 single catchment. The water and sediment samples were collected at, and
68 downstream of, the wastewater treatment plant (Lanrwst, Wales, UK), at the estuary
69 of the river Conwy (Wales, UK) near Morfa beach (Figure 1). To our knowledge, this
70 is the first study to use unamplified environmental viral RNA for sequencing library
71 construction, sequence dataset production and subsequent analysis. Because we
72 used a directional library sequencing protocol on RNA, rather than amplifying to
73 cDNA, we were able to distinguish single-stranded from double-stranded RNA
74 genome fragments.

75 Results

76 Sample overview

77 Wastewater influent and effluent samples were collected from the Llanrwst
78 wastewater treatment plant (53°08'24.4"N 3°48'12.8"W; Figure 1) in September and
79 October 2016, resulting in four different samples, LI_13-9 (Llanrwst influent Sep
80 2016), LE_13-9 (Llanrwst effluent Sep 2016), LI_11-10 (Llanrwst influent Oct 2016),
81 LE_11-10 (Llanrwst effluent Oct 2016). Estuarine surface water (SW) was collected
82 from Morfa beach (53°17'37.7"N 3°50'22.2"W; Conwy, Wales, Figure 1) in November

2016 and sediment from the same site in October and November 2016 (Sed1, Sed2, respectively).

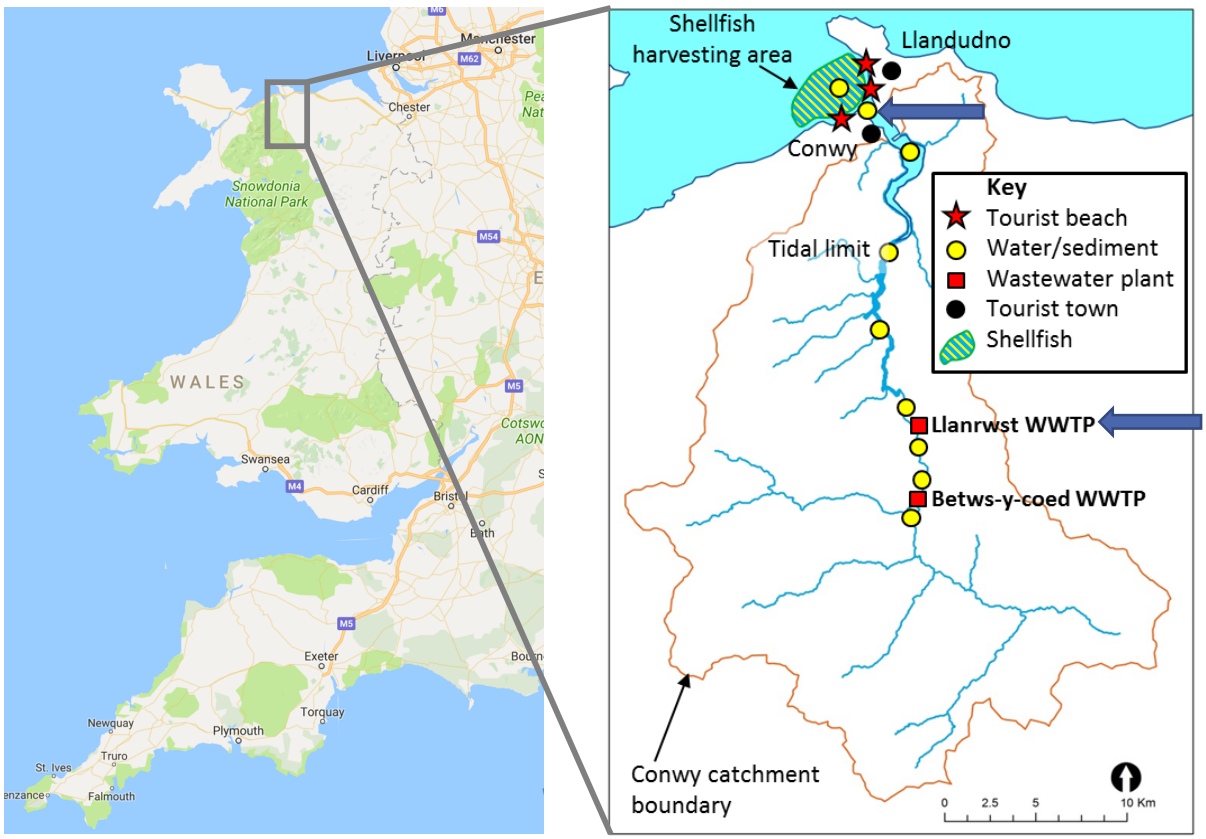


Figure 1: Map of the sampling locations, indicated with blue arrows. Data in the left panel was taken from Google Maps.

As an initial assessment, samples were tested for the presence of a subset of locally occurring enteric RNA viruses using qRT-PCR (Table 1). Only norovirus (NoV) genogroup GII signatures were detected in the wastewater samples. In the samples collected in September 2016, 10^3 genome copies (gc)/l of norovirus GII were observed in both the influent (LI_13-9) and in the effluent (LE_13-9). In the samples collected in October 2016, approx. 10^2 gc/l (below the limit of quantification which was approx. 200 gc/l) were observed in the influent (LI_11-10) and a considerably

96 higher concentration of $5x10^4$ gc/l was noted in the effluent (LE_11-10). All qRT-
97 PCRs were negative for the presence of sapoviruses (SaV) and hepatitis A/E viruses
98 (HAV/HEV). None of the target enteric viruses were found in the surface water and
99 sediment samples.

100 **Table 1: Summary of viromic and qRT-PCR detection of the presence of**
101 **specific RNA viruses across the samples (sewage, estuarine water and**
102 **sediment).**

Sample name[a]	Sample volume/ mass	Location	# contigs (curated)	Target RNA viruses detected in contigs[b]	qRT-PCR results (gc/l)[c]
LI_13-9	1 l	Llanrwst WWTP	5721	RVA, RVC, PBV, SaV	NoVGII 1,457
LE_13-9	1 l	Llanrwst WWTP	2201	RVA, RVC, PBV	NoVGII 1,251
LI_11-10	1 l	Llanrwst WWTP	859	PBV	NoVGII detected
LE_11-10	1 l	Llanrwst WWTP	5433	NoVGI, RVA, RVC, PBV, AsV	NoVGII 50,180
SW	50 l	Morfa beach	243	-	-
Sed1	60 g	Morfa beach	550[d]	-	-
Sed2	60 g	Morfa beach	550[d]	-	-

103 [a] LI: sewage influent; LE: sewage effluent; SW: estuarine surface water; Sed: estuarine sediment
104 [b] RVA: rotavirus A; RVB: rotavirus B; PBV: picobirnavirus; SaV: sapovirus; NoVGI: norovirus genogroup I; AsV:
105 astrovirus
106 [c] Samples were tested with qRT-PCR for the following targets: NoVGI, NoVGII, SaV, HAV, HEV. Results
107 reported in genome copies per liter (gc/l), NoVGII was detected below limit of quantification (approx. 200 gc/l) in
108 sample LI_11-10. Nov GII was the only target virus detected by qRT-PCR.
109 [d] Samples Sed1 and Sed2 were assembled together into the contig dataset Sed.
110

111 Summary of viral diversity

112 The virus taxonomic diversity present in each sample was assessed by comparison
113 of curated read and contig datasets with both the RefSeq Viral protein database and
114 the non-redundant protein database of NCBI, using Diamond blastx (14) and lowest
115 common ancestor taxon assignment with Megan 6 (15). For wastewater samples

116 LI_13-9, LE_13-9 and LE_11-10, two libraries were processed (indicated with _1 and
117 2 in the dataset names) and one each for the wastewater influent sample LI_11-10,
118 the surface water sample (SW) and two sediment samples (Sed1 & Sed2). This
119 section focuses on those reads and contigs that have been assigned to the viral
120 fraction exclusively, disregarding sequences of cellular or unknown origin.

121 The wastewater samples showed a greater richness of known viruses and had a
122 larger number of curated contigs than the surface water and sediment samples
123 (Figure 2). At the viral family level, between 14 and 34 groups were observed for
124 wastewater influent and effluent samples, including the unclassified levels, 12 for the
125 surface estuarine water sample, and 11 and 5 for the sediment samples Sed1 and
126 Sed2, respectively. The unclassified viruses and unassigned bins are indicated in
127 red in Figure 2 and made up the majority of known reads in the estuarine sediment
128 samples. In most of the viromes, dsDNA and ssDNA virus families were present,
129 despite having performed a DNase treatment after viral nucleic acid extraction (Table
130 S1). These families represented only a minor (<5%) proportion of the total assigned
131 reads with a few exceptions. In wastewater influent sample LI_11-10, reads assigned
132 to the dsDNA family *Papillomaviridae* accounted for 61% of the total, while in the
133 surface water sample reads assigned to the ssDNA families *Circoviridae* and
134 *Microviridae* represented 50% and 12% of the total, respectively. This is most likely
135 to be due to incomplete digestion of the viral DNA with the DNase Max kit than to
136 corresponding mRNA transcripts actually being present in the viromes.

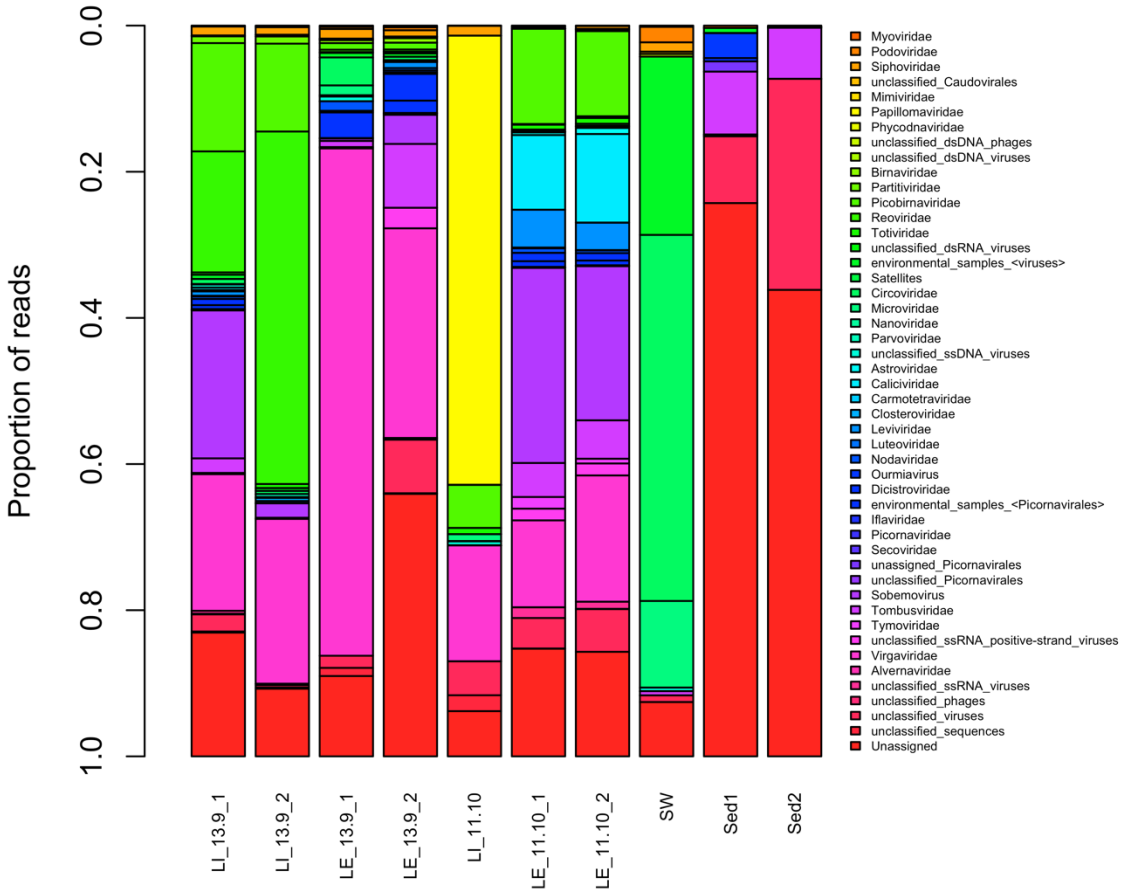


Figure 2: Taxonomic distribution of curated read data at the virus family level. Reads were assigned to a family or equivalent group by Megan6 using a lowest common ancestor algorithm, based on blastx-based homology using the program Diamond with the RefSeq Viral protein database (version January 2017) and the non-redundant protein database (version May 2017). Only viral groupings are shown. LI: sewage influent; LE: sewage effluent; SW: estuarine surface water; Sed: estuarine sediment.

146 The families of dsRNA viruses present in these datasets were *Totiviridae* (fungi and
147 protist hosts), *Reoviridae* (invertebrate, vertebrate & plant hosts), *Picobirnaviridae*
148 (mammals), *Partitiviridae* (fungi & protists) and *Birnaviridae* (vertebrates and
149 invertebrates), with a small number of reads recognized as unclassified dsRNA
150 viruses (Figure 2). None of these groups were present in all libraries, but totivirus
151 and picobirnavirus signatures were present in all wastewater samples and reoviruses
152 were found in three out of the four wastewater samples. *Partitiviridae* signatures
153 were only found in the wastewater LE_11-10 and LI_13-9 samples, while
154 *Birnaviridae* reads were only present in the wastewater LE_13-9 libraries. The
155 sediment and surface water samples did not have detectable levels of dsRNA virus
156 sequences.

157 Positive sense ssRNA viruses were the most diverse class of viruses present in
158 these datasets. The family *Tombusviridae*, which groups plant viruses with
159 monopartite or bipartite linear genomes (16), was present in all samples with the sole
160 exception of the wastewater influent sample LI_11-10 (Figure 2, Table S1). Virus
161 signatures belonging to the family *Virgaviridae*, representing plant viruses, were
162 present in all wastewater samples at comparable levels. Other highly represented
163 families or groupings were the families *Dicistroviridae* (invertebrate hosts),
164 *Nodaviridae* (invertebrate & vertebrate hosts) and the bacteriophage family
165 *Leviviridae*, the plant virus genus *Sobemovirus*, and the groupings of "unclassified
166 ssRNA positive-strand viruses" and several unclassified/unassigned/environmental
167 members of the order *Picornavirales*. Sediment sample Sed1 was the only sample
168 with signatures of the family *Alvernaviridae*, which has as its sole member the
169 dinoflagellate virus Heterocapsa circularisquama RNA virus 01. The wastewater
170 effluent sample LE_11-10 and influent sample LI_13-9_1 were the only samples with

171 calicivirus signatures, and sample LE_11-10_1 and LE_1-10_2 were the only
172 samples with *Astroviridae* reads (vertebrate host). Several families of the order
173 *Picornavirales* were detected in the wastewater samples at different levels in
174 different samples, and a small number of unassigned picornaviruses was detected in
175 the surface water sample (SW).

176 We did not observe any known negative sense ssRNA viruses in any of the
177 sequencing libraries, but it is possible that some of the unaffiliated viral contigs
178 belong to this class. These types of viruses are enveloped and predicted to degrade
179 more rapidly than the non-enveloped enteric viruses in wastewater treatment plants
180 and the environment (17). Given that the known families of negative sense ssRNA
181 viruses consist of potentially deadly pathogens, such as Influenza A virus
182 (orthomyxovirus), Lassa virus (arenavirus), Zaire ebolavirus (filovirus) or Rabies
183 virus (rhabdovirus), the lack of signatures of these viruses in the datasets most likely
184 suggests reassuringly that they were not present in the investigated samples.

185

186 Potential human pathogenic viruses

187 An important aim of this study was to investigate the presence and genomic diversity
188 of potential human pathogenic RNA viruses in different sample types within the river
189 catchment area. To minimize miss-assignments of short sequences to taxa, we used
190 the assembled, curated contig dataset and looked for contigs representing near-
191 complete viral genomes.

192 Presence of a norovirus GI.2 genome

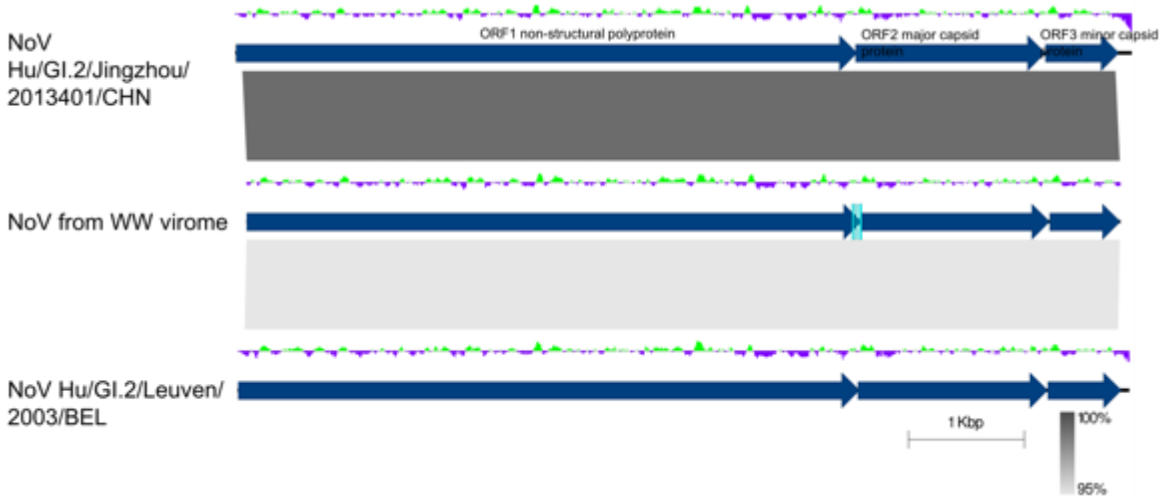
193 We were particularly interested in finding norovirus genomes to explore the genomic
194 diversity of these important and potentially abundant pathogens originating from

195 sewage and disseminated in watercourses, with implications for shellfisheries and
196 recreational waters. This is of relevance due to known issues of sewage
197 contamination in the region (18). Members of the genus *Norovirus* (family
198 *Caliciviridae*) are non-enveloped, icosahedral (+)ssRNA viruses with a linear,
199 unsegmented ~7.6 kb genome encoding three ORFs (16). These viruses are divided
200 into different genogroups of which GI and GII are associated with human
201 gastroenteritis (19, 20). Noroviruses are identified routinely by qRT-PCR, providing
202 an opportunity to examine correlations between qRT-PCR and metaviromic results.

203 We only found norovirus signatures in the libraries of wastewater effluent sample
204 LE_11-10. These reads assembled into a single contig of 7,542 bases, representing
205 a near-complete norovirus genome (GenBank accession number MG599789). Read
206 mapping showed an uneven coverage over the genome length between 18x and
207 745x (13,165 reads of library 1 and 8986 reads of library 2). This confirmed that our
208 contig was derived from a single-stranded genome, as all forward reads in the pairs
209 were oriented in the same direction. Based on this mapping, we performed variant
210 calling and the consensus sequence was corrected in cases where the variant was
211 present in more than 85% of the reads. To our knowledge, this is the only
212 metagenome-derived, environment-associated (i.e. non-host associated) near-
213 complete norovirus genome sequence deposited in a public database (INSDC
214 nuccore database was searched for norovirus, txid142786 sequences > 5000 nt).

215 A BLASTN search revealed two close relatives to our wastewater-associated
216 norovirus genome, norovirus Hu/GI.2/Jingzhou/2013401/CHN (KF306212) which is
217 7740 bases in length (21), displaying a nucleotide sequence identity of 99% over
218 99% of the genome length, and norovirus Hu/GI.2/Leuven/2003/BEL (FJ515294) at
219 95% sequence identity over 99% of the genome (Figure 3). From the 5' end of our

220 norovirus contig, 62 bases were missing compared with
221 Hu/GI.2/Jingzhou/2013401/CHN and from the 3' end 165 bases and the polyA tail
222 were not present. We compared the sequence of our norovirus with
223 Hu/GI.2/Jingzhou/2013401/CHN base by base and observed 81 SNPs and no other
224 forms of variation. Of the SNPs, only eight were non-synonymous resulting in five
225 different amino acids incorporated in the non-structural polyprotein (ORF1); one in
226 the major capsid protein (ORF2) and two in the minor structural protein (ORF3)
227 (Table S2). According to the current classification criteria, this level of similarity
228 places our assembled genome in genogroup GI, genotype GI.2, with only a single
229 amino acid different between the major capsid protein (MCP) of
230 Hu/GI.2/Jingzhou/2013401/CHN and the genome assembled here.



231

232 **Figure 3: Pairwise genome comparison between the virome norovirus genome**
233 **(middle) and its closest relatives, Norovirus Hu/GI.2/Jingzhou/2013401/CHN**
234 **and Norovirus Hu/GI.2/Leuven/2003/BEL.** BLASTN similarity is indicated in shades
235 of grey. ORFs are delineated by dark blue arrows. The deviation from the average
236 GC content is indicated above the genomes in a green and purple graph. The qRT-

237 PCR primer binding sites for the wastewater-associated genome are indicated by
238 light blue rectangles. The figure was created with Easyfig (22).

239

240 We tested the genotype grouping of our genome in a whole genome phylogeny with
241 all complete genome sequences of genogroup I available in GenBank. The
242 phylogenomic tree clearly delineated the different genotypes within genogroup GI,
243 placing the newly-assembled genome within genotype GI.2, with the reference
244 isolate for GII used as an outgroup (Figure 4).

245 For further validation, the full genome of the novel norovirus GI was recovered using
246 RT-PCR. However, the amplicon could not be ligated into a plasmid and hence was
247 not fully sequenced.

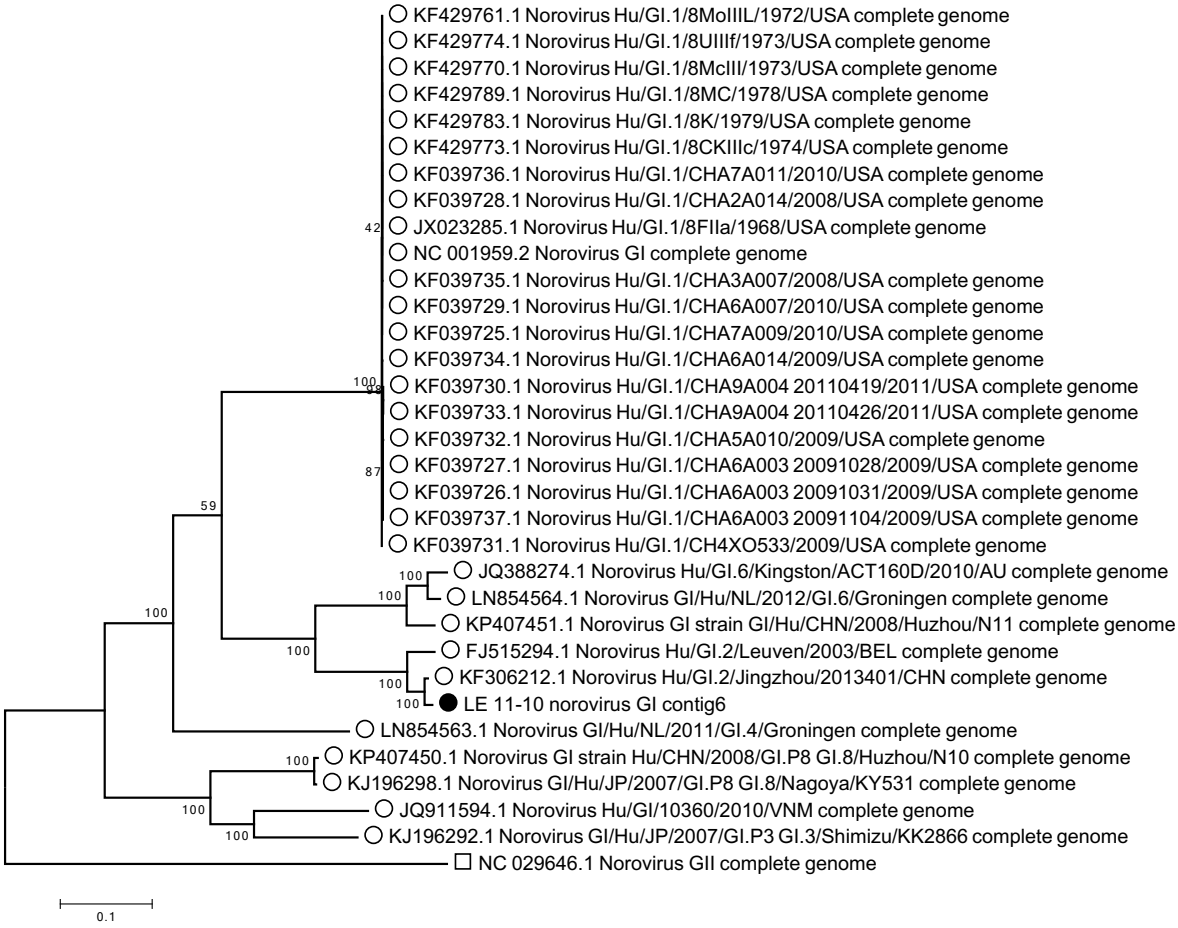


Figure 4: Maximum Likelihood phylogenetic tree of norovirus genomes belonging to genogroup GI, with the norovirus GII reference genome as outlier. The nucleotide sequences were aligned with MUSCLE and the alignment was trimmed to the length of the virome sequence LE_11-10 contig 6, resulting in 7758 positions analyzed for tree building. The Maximum Likelihood method was used with a Tamura Nei model for nucleic acid substitution. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The scale bar represents the number of substitutions per site.

258 **Presence of diverse rotavirus segments in wastewater samples**

259 Rotaviruses are segmented dsRNA viruses belonging to the family *Reoviridae,*
260 causing gastroenteric illness in vertebrates and are transmitted through the faecal-
261 oral route (16). Read signatures assigned to the genus *Rotavirus* were found in three
262 of the four wastewater samples (all but LI_11-10). Wastewater influent sample
263 LI_13-9 contained the most signatures with approximately 75,000 reads, assembled
264 into 120 contigs, representing genome fragments of 10 out of the 11 rotavirus
265 segments. At the species level, these genome fragments were assigned to either the
266 species *Rotavirus A* or *Rotavirus C*. Comparing the amino acid sequences of the
267 predicted proteins, some contigs showed high levels of identity (>88%) with either
268 the segments of rotavirus A (RVA) or rotavirus C (RVC) reference genomes as
269 available in the RefSeq database (23, 24), while others showed a lower identity with
270 a variety of RVC isolates only. The segmented genome nature and the possibility of
271 segment exchange make it difficult to confidently identify the number of rotavirus
272 types present in this sample. Given the amino acid similarities with both RVA and
273 RVC types (Supplementary Table 1), we suggest there are at least two, and possibly
274 three types present here.

275 Using the RotaC 2.0 typing tool for RVA, and blast-based similarity to known
276 genotypes, we have typed the rotavirus genome segments found here (Table 2). The
277 combined genomic make-up of the RV community in sample LI_13-9 was
278 G8/G10/Gx-P[1]/P[14]/P[41]/P[x]-I2/Ix-R2/Rx-C2/Cx-M2/Mx-A3/A11/Ax-Nx-T6/Tx-
279 E2/Ex (25, 26). The potential hosts for each segment were derived from the hosts of
280 the closest relatives. This analysis showed that the RVA viruses were possibly
281 infecting humans or cattle, while the RVC viruses were most likely porcine (Table 2).

282 **Table 2: Rotavirus A and C genome information and its detection in the LI_13-9**

283 **sample dataset.**

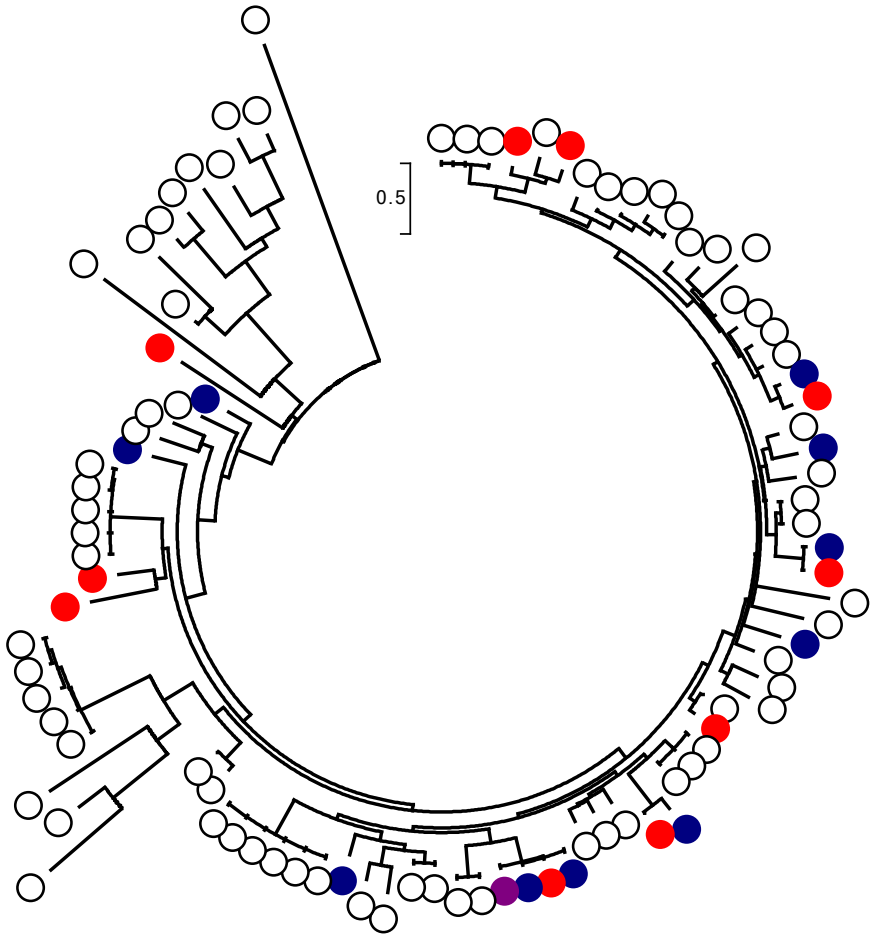
Genome segment	Length (nt)	Protein	Predicted function	# contigs	Putative genotypes	Potential hosts[a]
Rotavirus A						
Segment 1	3302	VP1	RNA-dependent RNA polymerase	7	R2	Human, cow
Segment 2	2693	VP2	core capsid protein	1	C2	Human
Segment 3	2591	VP3	RNA capping protein	1	M2	Human, sheep
Segment 4	2363	VP4	outer capsid spike protein	3	P[1], P[41], P[14]	Human, pig, alpaca, monkey
Segment 5	1614	NSP1	interferon antagonist protein	6	A3, A11	Human, cow, pig, deer
Segment 6	1356	VP6	inner capsid protein	1	I2	Human
Segment 7	1105	NSP3	translation effector protein	4	T6	Human, dog, cow
Segment 8	1059	NSP2	viroplasm RNA binding protein	0	-	-
Segment 9	1062	VP7	outer capsid glycoprotein	2	G10, G8	Cow, Human
Segment 10	751	NSP4	enterotoxin	1	E2	Human, cow
Segment 11	667	NSP5;6	phosphoprotein; non-structural protein	0	-	-
Rotavirus C				(contigs RVCX)		
Segment 1	3309	VP1	RNA-dependent RNA polymerase	7 (0)	Rx	Pig, cow
Segment 2	2736	VP2	core capsid protein	4(2)	Cx	Pig, dog
Segment 3	2283	VP4	outer capsid protein	2 (4)	P[x]	Pig
Segment 4	2166	VP3	guanylyltransferase	6 (0)	Mx	Pig
Segment 5	1353	VP6	inner capsid protein	1 (0)	Ix	Pig
Segment 6	1350	NSP3		0 (1)	Tx	Human
Segment 7	1270	NSP1		0 (2)	Ax	Pig, dog
Segment 8	1063	VP7	outer capsid glycoprotein	0 (2)	Gx	Pig
Segment 9	1037	NSP2		2 (0)	Nx	Pig
Segment 10	730	NSP5		0 (0)	-	-
Segment 11	613	NSP4	enterotoxin	0 (4)	Ex	Pig

284 [a] Potential hosts are defined as the hosts of the reference rotavirus sequence with the highest similarity to the
285 contigs found in the virome sample LI_13-9.

286

287 **Picobirnaviruses showed a high prevalence in wastewater**

288 All the wastewater virome libraries contained signatures assigned to the dsRNA
289 family *Picobirnaviridae*, genus *Picobirnavirus* (Figure 2) and these reads assembled
290 into between 42 (LE_13-9) and 510 (LI_13-9) contigs. Both picobirnavirus genome
291 segments, segment 1 containing two hypothetical proteins and segment 2 on which
292 the RNA-dependent RNA polymerase (RdRP) is encoded, were observed in the
293 samples. The contigs showed little sequence similarity with the reference genome
294 *Human picobirnavirus* (RefSeq segment accession numbers NC_007026.1 and
295 NC_007027.1). Phylogenetic analysis of a partial region of the predicted RdRPs in
296 the virome contigs was not able to resolve any cluster or evolutionary origin (Figure
297 5). Picobirnavirus RdRPs from human, animal and environmental isolates, as well as
298 the majority of the virome sequences were grouped in one large, unsupported cluster
299 that showed relatively little genomic diversity. While many picobirnaviruses have
300 been isolated from humans with gastroenteritis, a review of the known cases
301 suggested that picobirnaviruses are probably not the main cause of acute diarrhea
302 and are secondary pathogens with potential synergistic effects (27). A qRT-PCR-
303 based investigation into the suitability of human picobirnaviruses as indicators of
304 human faecal contamination, showed that they were not present in a sufficient
305 proportions of tested samples to be good water quality indicators (28), but their high
306 diversity in our sample set warrants further investigation using metaviromic methods.



307

308 **Figure 5: Maximum likelihood phylogenetic tree of RdRP amino acid**
309 **sequences of isolated and virome picobirnaviruses.** Sequences from isolates are
310 indicated with open dots, virome-derived sequences with closed dots, sample LI_11-
311 10 in purple, sample LE_11-10 in blue, and sample LI_13-9 in red. Sequences were
312 aligned using MUSCLE providing 114 amino acid positions for tree generation. The
313 Maximum Likelihood was used with a JTT matrix-based model [1]. The scale bar
314 represents the number of substitutions per site. Bootstrap values of all branches
315 were low.

316

317 ### Partial genomes of other potentially pathogenic RNA viruses

318 In sample LI_13-9, a small contig of 347 bases was found that was 94% identical at
319 the nucleotide level to the Sapovirus Mc2 ORF1 (AY237419), in the family
320 *Caliciviridae*. We have also identified four contigs of approximately 500 bases in
321 sample LE_11-10 that resembled most closely the Astrovirus MLB2 isolates
322 MLB2/human/Geneva/2014 (KT224358) and MLB2-LIHT (KX022687) at 99%
323 nucleotide identity. In addition, we identified several reads and contigs assigned to
324 the family *Picornaviridae* which comprises a diverse set of enteric viruses, but the
325 closest relatives in the databases were metagenomically assembled or unidentified
326 picornaviruses.

327

328 ## Discussion

329 We set out to explore the possibility of using viromics to find human pathogenic RNA
330 viruses in the environment. We have been successful in identifying several
331 potentially human pathogenic or zoonotic viral genomes from the wastewater
332 samples, but did not find any in the surface estuarine water and sediment samples.
333 The absence of signatures does not necessarily mean that there are no pathogenic
334 viruses present in water or sediment, but possibly that their levels are below our limit
335 of quantification (approximately 200 gc/l).

336 It is important to note here that during the RNA extraction process, many biases
337 could have been introduced leading to a lower recovery of input viruses. Samples
338 were first concentrated from volumes of 1 l (wastewater) or 50 l (surface water) down
339 to 50 ml using tangential flow filtration (TFF) at a molecular weight cut-off of 100
340 kDa, followed by PEG 6000 precipitation. These samples were diluted in fresh buffer,

341 filtered through syringe filters of 0.22 μm pore size and then treated with nuclease to
342 remove free DNA and RNA. Previous research has shown that while any enrichment
343 method aimed at fractionating the viral and cellular components will decrease the
344 total quantity of viruses, a combination of centrifugation, filtration and nuclease
345 treatment increases the proportion of viral reads in sequencing datasets (29). After
346 implementing these steps, we used the MO BIO PowerViral® Environmental
347 DNA/RNA extraction kit for viral RNA extraction, which has previously been shown to
348 perform best overall in spiking experiments with murine norovirus, in terms of
349 extraction efficiency and removal of inhibitors (30). The kit has, however, given low
350 recoveries of viruses from sediment (31).

351 We did not perform an amplification step before library construction with the
352 NEBNext Ultra Directional RNA Library Prep Kit for Illumina, to retain the genome
353 sense and strand information. Instead, we increased the number of cycles of random
354 PCR during library preparation from 12 to 15 to counteract the low input quantity of
355 RNA (< 1 ng). The random amplification during library construction led to a trade-off
356 in which genome strand information was gained for a loss of quantitative power,
357 making it difficult to compare abundances of viral types within and across libraries.
358 This random PCR-based bias has been highlighted before, but the proposed solution
359 of using library preparation protocols which limit the use of PCR are only feasible
360 with high amounts of input nucleic acid (32), which we have found to be impossible
361 when processing environmental/wastewater samples to generate RNA metaviromes.

362 A critical issue to highlight here, is the inclusion of controls in our sequencing
363 libraries in order to identify potential contaminants and their origins, as has been
364 suggested previously (33, 34). There have been multiple reports of false positive
365 genome discoveries, in particular the reported discovery of a novel parvovirus-like

366 hybrid in hepatitis patients that was later revealed to originate from the silica-based
367 nucleic acid extraction columns (35–37). In this study, we included a positive control
368 that comprised bacterial cells (*Salmonella* Typhimurium isolate D23580 RefSeq
369 accession number NC_016854) and mengovirus (36), an RNA virus that serves as a
370 process control, as well as two negative controls, an extraction control and a library
371 preparation control. Analysis of the control libraries showed that while the *Salmonella*
372 cells and DNA were successfully removed from the positive control sample by the
373 enrichment protocol, the mengovirus was not recovered. Subsequent qRT-PCR
374 analysis revealed that the mengovirus remained detectable in the pre-processing
375 stages of the extraction, but was lost after RNase treatment (data not shown).
376 Inclusion of an inactivation step of the DNase at 75°C potentially exacerbated the
377 effect of the RNase step. Consequently, it is likely that we have missed several viral
378 types during the extraction process despite having still managed to recover an RNA
379 metavirome harbouring substantial diversity.

380 Further examination of the HiSeq and MiSeq control datasets revealed a wide range
381 of contaminant signatures of prokaryotic, eukaryotic and viral origin, making up 45M
382 read pairs per control on the HiSeq platform and 1M read pairs for the MiSeq, even
383 though the 16S and 18S rRNA PCR and RT-PCR reactions showed no visible bands
384 on an agarose gel. Most bacterial contaminant reads belonged to the phyla
385 *Proteobacteria, Actinobacteria* and *Firmicutes*. The most abundant genera included
386 *Corynebacterium, Propionibacterium, Sphingomonas, Ralstonia, Pseudomonas,*
387 *Streptomyces, Staphylococcus* and *Streptococcus* which have in the past been
388 identified as common lab contaminants (38). Within the eukaryotic signatures,
389 human-derived reads, *Beta vulgaris* and *Anopheles* reads were the most prevalent,
390 pointing towards potential cross-contamination of the sequencing libraries. A small

391 number of virus signatures were also identified, with the most prominent being *Feline*
392 *calicivirus* and *Dengue virus*. The presence of the calicivirus was traced back to the
393 library preparation kits after the libraries were reconstructed and resequenced. The
394 dengue virus signature was a <100 nt sequence which was co-extracted in all the
395 samples and potentially originated in one of the reagents or spin extraction column.
396 All sequences present in the controls were carefully removed from the sample
397 datasets during the quality control stage of the bioinformatics processing before
398 further analysis. For future experiments, we will omit the RNase treatment step
399 during extraction and filter out any contaminating ribosomal RNA or cellular-derived
400 mRNA sequences as part of the bioinformatic quality control workflow.

401 Our results show that while contamination is an issue when dealing with low biomass
402 samples, the combination of increased random PCR cycles during library
403 preparation, deep sequencing (i.e. HiSeq rather than MiSeq) and computational
404 subtraction of control sequences provides data of sufficient quantity and quality to
405 assemble near-complete RNA virus genomes *de novo*.

406

407 **Norovirus**

408 Noroviruses are one of the most common causes of gastrointestinal disease in the
409 developed world, with an incidence in the UK estimated as approaching 4 million
410 cases per annum (39). The genotype most commonly associated with disease is
411 GII.4 (40–42) which was not detected in the metaviromes generated here.

412 We retrieved one norovirus GI genome, assembled from 22,151 reads, in
413 wastewater effluent sample LE_11-10. This finding was in direct conflict with the
414 qRT-PCR analysis of this sample which did not detect any NoV GI signatures (Table

415 1). In contrast, NoV GII signatures were detected by qRT-PCR, but no NoV GII
416 genomes or genome fragments were observed in the virome libraries. One
417 hypothesis to explain the discrepancy between PCR and viromics approaches lies in
418 the differences in extraction protocol. For qRT-PCR, no viral enrichment step was
419 performed and RNA was not extracted with the PowerViral kit. Therefore, NoV GII
420 could have been lost before virome sequencing, as was the process control virus. An
421 alternative hypothesis is that the NoV GII signatures detected during qRT-PCR were
422 derived from fragmented RNA or from particles with a compromised capsid. In both
423 these cases, the RNA will not be detected in the virome data because of the RNase
424 preprocessing steps implemented in the enrichment/extraction protocol. This calls
425 into question the reliance of qRT-PCR for NoV detection and whether the detected
426 viruses are infectious or merely remnants of previous infections. Further research
427 using, for example, capsid integrity assays combined with infectious particle counts
428 will need to be conducted to assess the validity of qRT-PCR protocols for norovirus
429 detection.

430 The inability to identify NoV GI with qRT-PCR might be related to the mismatched
431 base present in the forward primer sequence used for detection. We subsequently
432 conducted a normal, long-range PCR to validate the detection of this genotype, and
433 this yielded a fragment of the correct size, but we were unable to clone and
434 sequence this fragment. While the known NoV GI.2 genotypes do not have a
435 mismatch in the qRT-PCR probe sequence, it is possible that the genome recovered
436 in this study fell below the limit of detection using the ISO standard primer/probe
437 combination (ISO/TS 15216-2:2013). In a recent study, researchers designed an
438 improved probe and observed lower Ct values and a lower limit of detection for GI.2
439 strains from waterborne samples (43). In general, viromics as a means of

440 investigating water samples for the presence of norovirus, does have the advantage
441 of unequivocally demonstrating the presence of intact genomes, provided the sample
442 processing requirements do not lead to excessive loss of viruses resulting in false
443 negatives. Certainly, time and money permitting, viromics is a useful adjunct to
444 qPCR for samples that are deemed particularly important or critical for determination
445 of intact viral genome presence.

446 Due to the difficulty of culturing noroviruses in the lab, many studies have used male-
447 specific coliphages such as MS2 and GA, which are ssRNA phages belonging to the
448 family *Leviviridae*, as alternative model systems (44, 45). Interestingly, while some
449 levivirus signatures were present in all wastewater samples (< 500 reads), we
450 observed a striking co-occurrence of these viruses with norovirus signatures in both
451 libraries of sample LE_11-10 (> 2500 reads). The most commonly observed viruses
452 in this sample were Pseudomonas phage PRR1, an unclassified levivirus, and
453 Escherichia phages FI and M11 in the genus *Allolevivirus*. Further studies with more
454 samples and replicates will indicate whether there is a significant correlation between
455 the presence of leviviruses and noroviruses in water samples. Furthermore, the
456 higher abundance of allolevivirus compared with MS2-like viruses could indicate
457 that the former might be more relevant as model systems for noroviruses.

458

459 **Rotavirus**

460 Rotaviruses are, like noroviruses, agents of gastroenteritis, but the disease is
461 commonly associated with children under the age of 5 where severe diarrhea and
462 vomiting can lead to over 10,000 hospitalizations per year in England and Wales
463 (46). Since the introduction of the live-attenuated vaccine Rotarix, the incidence of

464 gastroenteritis in England has declined, specifically for children aged <2 and during
465 peak rotavirus seasons (47–49). Therefore, the discovery of a diverse assemblage of
466 rotavirus genome segments in the wastewater samples here was less expected than
467 the norovirus discovery. While we were unable to recover the genome of the vaccine
468 strain, our genomic evidence suggests that at least one RVA and one RVC
469 population were circulating in the Llanrwst region in September 2016.

470 The genome constellation for the RVA segments in sample LI_13-9, G8/G10-
471 P[1]/P[14]/P[41]-I2-R2-C2-M2-A3/A11-(N)-T6-E2-(H), is distinctly bovine in origin
472 (25) (N and H segments not recovered in this study). The closest genome segment
473 relatives based on nucleic acid similarity, however, have been isolated from humans
474 (Table 2), likely pointing towards a bovine-human zoonotic transmission of this virus
475 (50). The same genomic constellation has been found recently when unusual
476 G8P[14] RVA isolates were recovered from human strain collections in Hungary (51)
477 and Guatemala (52), and isolated from children in Slovenia (53) and Italy (54). Cook
478 and colleagues calculated that there would be approximately 5000 zoonotic human
479 infections per year in the UK from livestock transmission, but many would be
480 asymptomatic (55).

481 The origins of the RVC genome segments are more difficult to trace, because of
482 lower similarity scores with known RVC isolates. The majority of the segments were
483 similar to porcine RVC genomes, while others showed no nucleotide similarity at all,
484 only amino acid similarity. An explanation for the presence of pig-derived rotavirus
485 signatures can be farm run-off. While farm waste is not supposed to end up in the
486 sewage treatment plant, it is likely that the RVC segments originate directly from
487 pigs, not through zoonotic transfer. Run-off from fields onto public roads, broken
488 farm sewer pipes or polluted small streams might lead to porcine viruses entering the

489 human sewerage network, but we cannot provide formal proof from the data
490 available. Based on the evidence, we hypothesize that there is one, possibly two,
491 divergent strains of RVC circulating in the pig farms in the Llanrwst area.

492 ## Conclusion

493 In this study, we investigated the use of metagenomics for the discovery of RNA
494 viruses circulating in watercourses. We have found RNA viruses in all samples
495 tested, but potential human pathogenic viruses were only identified in wastewater.
496 The recovery of plant viruses in most samples points towards potential applications
497 in crop protection, for example the use of metaviromics in phytopathogen
498 diagnostics. However, technical limitations, including the amount of input material
499 necessary and contamination of essential laboratory consumables and reagents, are
500 currently the main bottleneck for the adoption of fine scale metagenomics in routine
501 monitoring and diagnostics. The discovery of a norovirus GI and a diverse set of
502 rotavirus segments in the corresponding metaviromes indicates that qPCR-based
503 approaches can miss a significant portion of relevant pathogenic RNA viruses
504 present in water samples. Therefore, metagenomics can, at this time, best be used
505 for exploration, to design new diagnostic markers/primers targeting novel genotypes
506 and to inform diagnostic surveys on the inclusion of specific additional target viruses.

507

508 ## Materials & Methods

509 ### Sample collection and processing

510 Wastewater samples were collected as part of a viral surveillance study described
511 elsewhere (Farkas et al, in submission). Wastewater influent and effluent, 1 L each,
512 was collected at the Llanrwst wastewater treatment plant by Welsh Water (Wales,
513 UK, Figure 1) on 12th September (processed on 13-9, sample designations LI_13-9
514 and LE_13-9) and 10th October 2016 (processed on 11-10, sample designations
515 LI_11-10 and LE_11-10). The wastewater treatment plant uses filter beds for
516 secondary treatment and serves approx. 4000 inhabitants. The estuarine surface
517 water (50 L) sample (SW) was collected at Morfa Beach (Conwy, Wales, Figure 1)
518 approx. 22 km downstream of the Llanrwst wastewater treatment plant on 19th
519 October and 2nd of November 2016 at low tide (only the sample from November was
520 used for sequencing as the October sample extract failed quality control). Together
521 with the surface water sample, 90 g of the top 1-2 cm layer of the sediment was also
522 collected (sample designations Sed1 for the October sample and Sed2 for the
523 November sample).

524 The wastewater and surface water samples were processed using a two-step
525 concentration method as described elsewhere (Farkas et al, in submission). In brief,
526 the 1l (wastewater) and 50l (surface water) samples were first concentrated down to
527 50 ml using a KrosFlo® Research IIi Tangential Flow Filtration System
528 (Spectrumlabs, USA) with a 100 PEWS membrane. Particulate matter was then
529 eluted from solid matter in the concentrates using beef extract buffer and then
530 viruses were precipitated using polyethylene glycol (PEG) 6000. The viruses from
531 the sediment samples were eluted and concentrated using beef extract elution and

532 PEG precipitation as described elsewhere (31). The precipitates were eluted in 2-10
533 mL phosphate saline buffer, (PBS, pH 7.4) and stored at -80°C.

534 Detection and quantification of enteric viruses with qRT-PCR

535 Total nucleic acids were extracted from a 0.5 mL aliquot of the concentrates using
536 the MiniMag NucliSENS® MiniMag® Nucleic Acid Purification System (bioMérieux
537 SA, France). The final volume of the nucleic acid solution was 0.05 mL (surface
538 water and sediment) and 0.1 mL (wastewater samples). Norovirus GI and GII,
539 sapovirus GI, and hepatitis A and E viruses were targeted in qRT-PCR assays as
540 described elsewhere (56).

541 Viral RNA extraction for metaviromic sequencing

542 Viral particles were extracted from the concentrated samples by filtration. In a first
543 step, the samples were diluted in 10 ml of sterile 0.5 M NaCl buffer and incubated at
544 room temperature with gentle shaking for 30 min to disaggregate particles. The
545 suspension was then filtered through a sterile, 0.22 µm pore size syringe filter
546 (Millex, PES membrane). The sample was desalted by centrifugation (3200 x g,
547 between 1 and 6h for different samples) in a sterilized spin filter (Vivaspin 20, 100
548 kDa molecular weight cut-off) and replacement of the buffer solution with 5 ml of a
549 Tris-based buffer (10 mM TrisHCl, 10 mM $MgSO_4$, 150 mM NaCl, pH 7.5). The buffer
550 exchange was performed twice and the volume retained after the final spin was <
551 500 µl. The samples were then treated with Turbo DNase (20 Units; Ambion) and
552 incubated for 30 minutes at 37°C, followed by inactivation at 75°C for 10 minutes. In
553 a next step, all samples were treated with 80 µg RNase A (Thermo Fisher Scientific)
554 and incubated at 37°C for 30 minutes. The RNase was inactivated with RiboLock
555 RNase Inhibitor (Thermo Fisher Scientific) and the inactivated complex was removed

556 by spin filtration (Vivaspin 500, 100 kDa molecular weight cut-off) and the samples
557 centrifuged until the volume was approximately 200 µl. Viral DNA and RNA were co-
558 extracted using the PowerViral Environmental DNA/RNA kit (MOBIO Laboratories)
559 according to the manufacturer's instructions. In this protocol, buffer PV1 was
560 supplemented with 20 µl/ml betamercaptoethanol to further reduce RNase activity.
561 The nucleic acid was eluted in 100 µl RNase-free water. The extracted viral DNA
562 was degraded using the DNase Max kit (Mobio) according to the manufacturer's
563 instructions. The remaining viral RNA was further purified and concentrated by
564 ethanol precipitation using 2.5 x sample volume of 100% ethanol and 1/10 volume of
565 DEPC-treated Na-acetate (3 M). The quantity and quality of RNA was determined
566 with Bioanalyzer Pico RNA 6000 capillary electrophoresis (Agilent Technologies). A
567 positive and negative extraction control sample were processed alongside the main
568 samples. The positive control samples contained *Salmonella enterica* strain D23580
569 which is not found in the UK (57) and a process control virus mengovirus (56, 58).

570 The viral RNA extracts were tested for bacterial and eukaryotic cellular
571 contamination using 16S and 18S rRNA gene PCR and RT-PCR, with primers e9F
572 (59) and 519R (60), and primers 1389F and 1510R (61), for the 16S and 18S rRNA
573 gene, respectively. Complimentary DNA was created using the SuperScript III
574 Reverse Transcriptase (Invitrogen) with random hexamer primers according to the
575 manufacturer's instructions. (RT)-PCR was performed with the MyTaq Red Mix
576 (Bioline) for 35 cycles (95°C for 45 sec, 50°C for 30 sec, 72°C 1 min 40 sec) and
577 visualized on a 1% agarose gel. Samples were considered suitable for sequencing if
578 no DNA bands were visible on the gel.

579 ## Library preparation and sequencing

580 The library preparation and sequencing were performed at the University of Liverpool
581 Centre for Genomics Research (CGR). Twelve dual indexed, strand-specific libraries
582 were created using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina,
583 according to the manufacturer's instructions. These libraries were pooled and
584 sequenced at 2 x 150 bp read lengths on the Illumina HiSeq 4000 platform. This
585 generated between 10 and 110 million paired reads per sample.

586 To confirm our results, a second set of libraries was constructed from new kits and a
587 milliQ water samples was included as a library prep control. The thirteen resulting
588 libraries were sequenced on the Illumina MiSeq platform at CGR, at 2 x 150 bp read
589 lengths. These data were used for verification and control purposes only as
590 sequencing depth was insufficient for the bioinformatics analyses described in the
591 rest of the study.

592 ## Bioinformatics

593 All command line programs for data analysis were run on the bioinformatics cluster
594 of CGR (University of Liverpool) in a Debian 5 or 7 environment.

595 Raw fastq files were trimmed to remove Illumina adapters using Cutadapt version
596 1.2.1 using option -O 3 (62) and Sickle version 1.200 with a minimum quality score of
597 20 (63). Further quality control was performed with Prinseq-lite (64) with the following
598 parameters: minimum read length 35, GC percentage between 5-95%, minimum
599 mean quality 25, dereplication (removal of identical reads, leaving 1 copy), removal
600 of tails of minimum 5 polyN sequences from 3' and 5' ends of reads.

601 The positive and negative control libraries described earlier were used for
602 contaminant removal. The reads of the control samples were analyzed using

603 Diamond blastx (14) against the non-redundant protein database of NCBI (nr version
604 November 2015). The blast results were visualized using Megan6 Community
605 Edition (15). An extra contaminant file was created with complete genomes of
606 species present at over 1000 reads in the positive and negative control samples.
607 Then, bowtie2 (65) was used for each sample to subtract the reads that mapped to
608 the positive control, negative control or contaminant file. The unmapped reads were
609 used for assembly with SPAdes version 3.9.0 with kmer values 21, 31, 41, 51, 61,
610 71, and the options --careful and a minimum coverage of 5 reads per contig (66).
611 The contig files of each sample were compared with the contigs of the controls
612 (assembled using the same parameters) using blastn of the BLAST+ suite (67).
613 Contigs that showed significant similarity with control contigs were manually
614 removed, creating a curated contig dataset. The unmapped read datasets were then
615 mapped against this curated contig dataset with bowtie2 and only the reads that
616 mapped were retained, resulting in a curated read dataset.

617 The curated contig and read datasets were compared to the Viral RefSeq (release
618 January 2017) and non-redundant protein (nr, release May 2017) reference
619 databases using Diamond blastx at an e value of 1e-5 for significant hits (14, 68, 69).
620 Taxon assignments were made with Megan6 Community Edition according to the
621 lowest common ancestor algorithm at default settings (15). The taxon abundance
622 data were extracted from Megan6 and imported into RStudio for visualization (70).
623 Genes were predicted on the assembled contigs with Prokka (71) using the settings -
624 -kingdom Viruses and an e value of 1e-5. Multiple alignments of genes and genomes
625 were made in MEGA7 using the MUSCLE algorithm at default settings (72, 73). The
626 alignments were manually trimmed and phylogenetic trees were built using the
627 Maximum Likelihood method in MEGA7 at the default settings.

628 ## Accession numbers

629 Read and contig datasets are available from NCBI under the following BioProject
630 accession numbers, PRNJA421889 (wastewater data), PRNJA421892 (sediment
631 data) and PRJNA421894 (estuarine water data). The NoV GI genome isolate was
632 deposited in GenBank under accession number MG599789.

633

634 ## Author contributions

635 EMA, KF, DJ, HA and AJM designed the experiments, EMA, KF, CH, performed the
636 experiments, EMA analysed the data, EMA and KF wrote the manuscript and EMA
637 prepared the manuscript for submission. All authors critically reviewed and edited the
638 manuscript.

639 ## Acknowledgements

640 This study was funded by the Natural Environment Research Council (NERC) and
641 the Food Standards Agency (FSA) under the Environmental Microbiology and
642 Human Health (EMHH) Programme (NE/M010996/1). The authors gratefully
643 acknowledge Dr James Lowther (Centre for Environment, Fisheries and Aquaculture
644 Science; CEFAS) for providing the mengovirus sample. We also thank Gordon
645 Steffen and Dr Nick Barcock (Dŵr Cymru Cyf - Welsh Water Ltd, UK) for facilitating
646 sample collection at the wastewater treatment plants and Dr Julie Webb (Bangor
647 University, UK) for assistance in sampling.

648 ## References

649 1. Lin J, Ganesh A. 2013. Water quality indicators: bacteria, coliphages, enteric
650 viruses. Int J Environ Health Res 23:484–506.

651 2. Girones R, Ferrús MA, Alonso JL, Rodriguez-Manzano J, Calgua B, de Abreu
652 Corrêa A, Hundesa A, Carratala A, Bofill-Mas S. 2010. Molecular detection of
653 pathogens in water - The pros and cons of molecular techniques. Water Res
654 44:4325–4339.

655 3. Laverick MA, Wyn-Jones AP, Carter MJ. 2004. Quantitative RT-PCR for the
656 enumeration of noroviruses (Norwalk-like viruses) in water and sewage. Lett
657 Appl Microbiol 39:127–136.

658 4. Rodriguez-Manzano J, Miagostovich M, Hundesa A, Clemente-Casares P,
659 Carratala A, Buti M, Jardi R, Girones R. 2010. Analysis of the evolution in the
660 circulation of HAV and HEV in Eastern Spain by testing urban sewage
661 samples. J Water Health 8:346–354.

662 5. Schvoerer E, Ventura M, Dubos O, Cazaux G, Serceau R, Gournier N, Dubois
663 V, Caminade P, Fleury HJA, Lafon ME. 2001. Qualitative and quantitative
664 molecular detection of enteroviruses in water from bathing areas and from a
665 sewage treatment plant. Res Microbiol 152:179–186.

666 6. Fong TT, Phanikumar MS, Xagoraraki I, Rose JB. 2010. Quantitative detection
667 of human adenoviruses in wastewater and combined sewer overflows
668 influencing a Michigan river. Appl Environ Microbiol 76:715–723.

669 7. Bofill-Mas S, Albinana-Gimenez N, Clemente-Casares P, Hundesa A,
670 Rodriguez-Manzano J, Allard A, Calvo M, Girones R. 2006. Quantification and
671 stability of human adenoviruses and polyomavirus JCPyV in wastewater
672 matrices. Appl Environ Microbiol 72:7894–7896.

673 8. Hellmér M, Paxéus N, Magnius L, Enache L, Arnholm B, Johansson A,
674 Bergström T, Norder H. 2014. Detection of pathogenic viruses in sewage
675 provided early warnings of hepatitis A virus and norovirus outbreaks. Appl
676 Environ Microbiol 80:6771–6781.

677 9. Nieuwenhuijse DF, Koopmans MPG. 2017. Metagenomic Sequencing for
678 Surveillance of Food- and Waterborne Viral Diseases. Front Microbiol 8:1–11.

679 10. Symonds EM, Breitbart M. 2015. Affordable enteric virus detection techniques
680 are needed to support changing paradigms in water quality management.
681 Clean - Soil, Air, Water 43:8–12.

682 11. Rosario K, Symonds EM, Sinigalliano C, Stewart J, Breitbart M. 2009. Pepper
683 mild mottle virus as an indicator of fecal pollution. Appl Environ Microbiol
684 75:7261–7267.

685 12. Stachler E, Bibby K. 2014. Metagenomic evaluation of the highly abundant
686 human gut bacteriophage CrAssphage for source tracking of human fecal
687 pollution. Environ Sci Technol Lett 1:405–409.

688 13. Bibby K. 2013. Metagenomic identification of viral pathogens. Trends
689 Biotechnol 31:275–9.

690 14. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment
691 using DIAMOND. Nat Methods 12:59–60.

692 15. Huson DH, Weber N. 2013. Microbial community analysis using MEGAN.
693 Methods Enzymol 531:465–85.

694 16. 2012. Virus taxonomy, 9th ed. Elsevier Inc., London, UK.

695 17. Ye Y, Ellenberg RM, Graham KE, Wigginton KR. 2016. Survivability,
696 partitioning, and recovery of enveloped viruses in untreated municipal w
697 astewater. Environ Sci Technol 50:5077–5085.

698 18. Winterbourn JB, Clements K, Lowther JA, Malham SK, McDonald JE, Jones
699 DL. 2016. Use of Mytilus edulis biosentinels to investigate spatial patterns of
700 norovirus and faecal indicator organism contamination around coastal sewage
701 discharges. Water Res 105:241–250.

702 19. Patel MM, Hall AJ, Vinjé J, Parashar UD. 2009. Noroviruses: A comprehensive
703 review. J Clin Virol 44:1–8.

704 20. Zheng DP, Ando T, Fankhauser RL, Beard RS, Glass RI, Monroe SS. 2006.
705 Norovirus classification and proposed strain nomenclature. Virology 346:312–
706 323.

707 21. Huo Y, Cai A, Yang H, Zhou M, Yan J, Liu D, Shen S. 2014. Complete
708 nucleotide sequence of a norovirus GII.4 genotype: Evidence for the spread of
709 the newly emerged pandemic Sydney 2012 strain to China. Virus Genes
710 48:356–360.

711 22. Sullivan MJ, Petty NK, Beatson SA. 2011. Easyfig: a genome comparison
712 visualizer. Bioinformatics 27:1009–1010.

713 23. Small C, Barro M, Brown TL, Patton JT. 2007. Genome heterogeneity of SA11
714 rotavirus due to reassortment with “ O ” agent. Virology 359:415–424.

715 24. Chen Z, Lambden PR, Lau J, Caul EO, Clarke IN. 2002. Human group C
716 rotavirus : completion of the genome sequence and gene coding assignments
717 of a non-cultivatable rotavirus. Virus Res 83:179–187.

718 25. Matthijnssens J, Ciarlet M, Heiman E, Arijs I, Delbeke T, McDonald SM,
719 Palombo EA, Iturriza-Gomara M, Maes P, Patton JT, Rahman M, Van Ranst
720 M. 2008. Full genome-based classification of rotaviruses reveals a common
721 origin between human Wa-Like and porcine rotavirus strains and human DS-1-
722 like and bovine rotavirus strains. J Virol 82:3204–3219.

723 26. Matthijnssens J, Ciarlet M, Rahman M, Attoui H, Bányai K, Estes MK, Gentsch
724 JR, Iturriza-Gómara M, Kirkwood CD, Martella V, Mertens PPC, Nakagomi O,
725 Patton JT, Ruggeri FM, Saif LJ, Santos N, Steyer A, Taniguchi K,
726 Desselberger U, Van Ranst M. 2008. Recommendations for the classification
727 of group a rotaviruses using all 11 genomic RNA segments. Arch Virol
728 153:1621–1629.

729 27. Ganesh B, Bányai K, Martella V, Jakab F, Masachessi G, Kobayashi N. 2012.
730 Picobirnavirus infections: viral persistence and zoonotic potential. Rev Med
731 Virol 22:245–256.

732 28. Hamza IA, Jurzik L, Überla K, Wilhelm M. 2011. Evaluation of pepper mild
733 mottle virus, human picobirnavirus and Torque teno virus as indicators of fecal
734 contamination in river water. Water Res 45:1358–1368.

735 29. Hall RJ, Wang J, Todd AK, Bissielo AB, Yen S, Strydom H, Moore NE, Ren X,
736 Huang QS, Carter PE, Peacey M. 2014. Evaluation of rapid and simple
737 techniques for the enrichment of viruses prior to metagenomic virus discovery.
738 J Virol Methods 195:194–204.

739 30. Iker BC, Bright KR, Pepper IL, Gerba CP, Kitajima M. 2013. Evaluation of
740 commercial kits for the extraction and purification of viral nucleic acids from
741 environmental and fecal samples. J Virol Methods 191:24–30.

742 31. Farkas K, Hassard F, McDonald JE, Malham SK, Jones DL. 2017. Evaluation
743 of molecular methods for the detection and quantification of pathogen-derived
744 nucleic acids in sediment. Front Microbiol 8:53.

745 32. Van Dijk EL, Jaszczyszyn Y, Thermes C. 2014. Library preparation methods
746 for next-generation sequencing: Tone down the bias. Exp Cell Res 322:12–20.

747 33. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. 2014. Tracking
748 down the sources of experimental contamination in microbiome studies.
749 Genome Biol 15:564.

750 34. Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, Fewell C,
751 Taylor CM, Flemington EK. 2014. Microbial contamination in next generation
752 sequencing: Implications for sequence-based analysis of clinical samples.
753 PLoS Pathog 10:e1004437.

754 35. Zhi N, Hu G, Wan Z, Zheng X, Liu X, Wong S, Kajigaya S, Zhao K, Young NS,
755 Africa S. 2014. Correction for Xu et al., Hybrid DNA virus in Chinese patients
756 with seronegative hepatitis discovered by deep sequencing. Proc Natl Acad
757 Sci 111:4344–4345.

758 36. Zhi N, Hu G, Wong S, Zhao K, Mao Q, Young NS. 2014. Reply to Naccache et
759 al: Viral sequences of NIH-CQV virus, a contamination of DNA extraction
760 method. Proc Natl Acad Sci 111:E977–E977.

761 37. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A,
762 Aronsohn A, Hackett J, Delwart EL, Chiu CY. 2013. The perils of pathogen
763 discovery: Origin of a novel Parvovirus-like hybrid genome traced to nucleic
764 acid extraction spin columns. J Virol 87:11966–11977.

765 38. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P,
766 Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory
767 contamination can critically impact sequence-based microbiome analyses.
768 BMC Biol 12:87.

769 39. Harris JP, Iturriza-Gomara M, O'Brien SJ. 2017. Re-assessing the total burden

770 of norovirus circulating in the United Kingdom population. Vaccine 35:853–
771 855.

772 40. Siebenga JJ, Vennema H, Zheng D, Vinjé J, Lee BE, Pang X, Ho ECM, Lim
773 W, Choudekar A, Broor S, Halperin T, Rasool NBG, Hewitt J, Greening GE, Jin
774 M, Duan Z, Lucero Y, O'Ryan M, Hoehne M, Schreier E, Ratcliff RM, White
775 PA, Iritani N, Reuter G, Koopmans M. 2009. Norovirus Illness Is a Global
776 Problem: Emergence and Spread of Norovirus GII.4 Variants, 2001–2007. J
777 Infect Dis 200:802–812.

778 41. Eden J-S, Tanaka MM, Boni MF, Rawlinson WD, White PA. 2013.
779 Recombination within the Pandemic Norovirus GII.4 Lineage. J Virol 87:6270–
780 6282.

781 42. Cannon JL, Barclay L, Collins NR, Wikswo ME, Castro CJ, Magaña LC,
782 Gregoricus N, Marine RL, Chhabra P, Vinjé J. 2017. Genetic and
783 Epidemiologic Trends of Norovirus Outbreaks in the United States from 2013
784 to 2016 Demonstrated Emergence of Novel GII.4 Recombinant Viruses. J Clin
785 Microbiol 55:2208–2221.

786 43. Cho H-G, Lee S-G, Mun S-K, Lee M-J, Park P-H, Jheong W-H, Yoon M-H,
787 Paik S-Y. 2017. Detection of waterborne norovirus genogroup I strains using
788 an improved real time RT-PCR assay. Arch Virol 162:3389–3396.

789 44. Dunkin N, Weng S, Coulter CG, Jacangelo JG, Schwab KJ. 2017. Reduction
790 of Human Norovirus GI, GII, and Surrogates by Peracetic Acid and
791 Monochloramine in Municipal Secondary Wastewater Effluent. Environ Sci
792 Technol 51:11918–11927.

793 45. Arredondo-Hernandez LJR, Diaz-Avalos C, Lopez-Vidal Y, Castillo-Rojas G,
794 Mazari-Hiriart M. 2017. FRNA Bacteriophages as Viral Indicators of Faecal
795 Contamination in Mexican Tropical Aquatic Systems. PLoS One 12:e0170399.

796 46. Harris JP, Jit M, Cooper D, Edmunds WJ. 2007. Evaluating rotavirus
797 vaccination in England and Wales. Part I. Estimating the burden of disease.
798 Vaccine 25:3962–3970.

799 47. Bawa Z, Elliot AJ, Morbey RA, Ladhani S, Cunliffe NA, O'Brien SJ, Regan M,
800 Smith GE, Weinstein RA. 2015. Assessing the likely impact of a rotavirus
801 vaccination program in England: The contribution of syndromic surveillance.
802 Clin Infect Dis 61:77–85.

803 48. Thomas SL, Walker JL, Fenty J, Atkins KE, Elliot AJ, Hughes HE, Stowe J,
804 Ladhani S, Andrews NJ. 2017. Impact of the national rotavirus vaccination
805 programme on acute gastroenteritis in England and associated costs averted.
806 Vaccine 35:680–686.

807 49. Hungerford D, Read JM, Cooke RPD, Vivancos R, Iturriza-G??mara M, Allen
808 DJ, French N, Cunliffe N. 2016. Early impact of rotavirus vaccination in a large
809 paediatric hospital in the UK. J Hosp Infect 93:117–120.

810 50. Wilhelm B, Waddell L, Greig J, Rajić A, Houde A, McEwen SA. 2015. A
811 scoping review of the evidence for public health risks of three emerging

812 potentially zoonotic viruses: Hepatitis E virus, norovirus, and rotavirus. Prev
813 Vet Med 119:61–79.

814 51. Marton S, Doro R, Feher E, Forro B, Ihasz K, Varga-Kugler R, Farkas SL,
815 Banyai K. 2017. Whole genome sequencing of a rare rotavirus from archived
816 stool sample demonstrates independent zoonotic origin of human G8P[14]
817 strains in Hungary. Virus Res 227:96–103.

818 52. Gautam R, Mijatovic-Rustempasic S, Roy S, Esona MD, Lopez B, Mencos Y,
819 Rey-Benito G, Bowen MD. 2015. Full genomic characterization and
820 phylogenetic analysis of a zoonotic human G8P[14] rotavirus strain detected in
821 a sample from Guatemala. Infect Genet Evol 33:206–211.

822 53. Steyer A, Naglič T, Jamnikar-Ciglenečki U, Kuhar U. 2017. Detection and
823 Whole-Genome Analysis of a Zoonotic G8P[14] Rotavirus Strain Isolated from
824 a Child with Diarrhea. Genome Announc 5:e01053-17.

825 54. Medici MC, Tummolo F, Bonica MB, Heylen E, Zeller M, Calderaro A,
826 Matthijnssens J. 2015. Genetic diversity in three bovine-like human G8P[14]
827 and G10P[14] rotaviruses suggests independent interspecies transmission
828 events. J Gen Virol 96:1161–1168.

829 55. Cook N, Bridger J, Kendall K, Gomara MI, El-Attar L, Gray J. 2004. The
830 zoonotic potential of rotavirus. J Infect 48:289–302.

831 56. Farkas K, Peters DE, McDonald JE, de Rougemont A, Malham SK, Jones DL.
832 2017. Evaluation of Two Triplex One-Step qRT-PCR Assays for the
833 Quantification of Human Enteric Viruses in Environmental Samples. Food
834 Environ Virol 9:342–349.

835 57. Kingsley RA, Msefula CL, Thomson NR, Kariuki S, Holt KE, Gordon MA, Harris
836 D, Clarke L, Whitehead S, Sangal V, Marsh K, Achtman M, Molyneux ME,
837 Cormican M, Parkhill J, MacLennan CA, Heyderman RS, Dougan G. 2009.
838 Epidemic multiple drug resistant Salmonella Typhimurium causing invasive
839 disease in sub-Saharan Africa have a distinct genotype. Genome Res
840 19:2279–2287.

841 58. Hennechart-Collette C, Martin-Latil S, Guillier L, Perelle S. 2015.
842 Determination of which virus to use as a process control when testing for the
843 presence of hepatitis A virus and norovirus in food and water. Int J Food
844 Microbiol 202:57–65.

845 59. Reysenbach A, Pace N. 1995. Reliable amplification of hyperthermophilic
846 archaeal 16S rRNA genes by the polymerase chain reaction, p. 101–107. *In*
847 Robb, F, Place, A (eds.), Archaea: a laboratory manual. Cold Spring Harbor
848 Laboratory Press, New York, NY, USA.

849 60. Turner S, Pryer KM, Miao VP, Palmer JD. 1999. Investigating deep
850 phylogenetic relationships among cyanobacteria and plastids by small subunit
851 rRNA sequence analysis. J Eukaryot Microbiol 46:327–338.

852 61. Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. 2009. A method for
853 studying protistan diversity using massively parallel sequencing of V9

854 hypervariable regions of small-subunit ribosomal RNA Genes. PLoS One 4:1–
855 9.

856 62. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput
857 sequencing reads. EMBnet.journal 17:10–12.

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

860 64. Schmieder R, Edwards R. 2011. Quality control and preprocessing of
861 metagenomic datasets. Bioinformatics 27:863–864.

862 65. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2.
863 Nat Methods 9:357–359.

864 66. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A,
865 Prjibelsky A, Pyshkin A, Sirotkin A, Sirotkin Y, Stepanauskas R, McLean J,
866 Lasken R, Clingenpeel SR, Woyke T, Tesler G, Alekseyev MA, Pevzner PA.
867 2013. Assembling genomes and mini-metagenomes from highly chimeric
868 reads, p. 158–170. *In* Deng, M, Jiang, R, Sun, F, Zhang, X (eds.), Research in
869 Computational Molecular Biology. RECOMB 2013. Lecture Notes in Computer
870 Science. Springer, Berlin, Heidelberg.

871 67. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K,
872 Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics
873 10:421.

874 68. Brister JR, Ako-adjei D, Bao Y, Blinkova O. 2015. NCBI Viral Genomes
875 Resource. Nucleic Acids Res 43:D571–D577.

876 69. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B,
877 Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y,
878 Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM,
879 Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li
880 W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S,
881 Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H,
882 Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W,
883 Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt
884 KD. 2016. Reference sequence (RefSeq) database at NCBI: current status,
885 taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–
886 D745.

887 70. Racine JS. 2012. RStudio: A platform-independent IDE for R and Sweave. J
888 Appl Econom 27:167–172.

889 71. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation.
890 Bioinformatics 30:2068–2069.

891 72. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy
892 and high throughput. Nucleic Acids Res 32:1792–1797.

893 73. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary
894 Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol 33:msw054.

895