

1 Evaluation of the impact of concentration and extraction 2 methods on the targeted sequencing of human viruses from 3 wastewater

4 Minxi Jiang^a, Audrey L.W. Wang^a, Nicholas A. Be^b, Nisha Mulakken^c, Kara L. Nelson^a, Rose S. Kantor^{a*}

5 a. Department of Civil and Environmental Engineering, University of California, Berkeley, CA, USA

6 b. Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA, USA.

7 c. Computing and Global Security Directorates, Lawrence Livermore National Laboratory, Livermore,
8 CA, USA

9 *Corresponding author: Rose S. Kantor rkantor@berkeley.edu

10 Abstract

11 Sequencing human viruses in wastewater is challenging due to their low abundance compared to the
12 total microbial background. This study compared the impact of four virus concentration/extraction
13 methods (Innovaprep, Nanotrap, Promega, Solids extraction) on probe-capture enrichment for human
14 viruses followed by sequencing. Different concentration/extraction methods yielded distinct virus
15 profiles. Innovaprep ultrafiltration (following solids removal) had the highest sequencing sensitivity and
16 richness, resulting in the successful assembly of most near-complete human virus genomes. However, it
17 was less sensitive in detecting SARS-CoV-2 by dPCR compared to Promega and Nanotrap. Across all
18 preparation methods, astroviruses and polyomaviruses were the most highly abundant human viruses,
19 and SARS-CoV-2 was rare. These findings suggest that sequencing success can be increased by using
20 methods that reduce non-target nucleic acids in the extract, though the absolute concentration of total
21 extracted nucleic acid, as indicated by Qubit, and targeted viruses, as indicated by dPCR, may not be
22 directly related to targeted sequencing performance. Further, using broadly targeted sequencing panels
23 may capture viral diversity but risks losing signals for specific low-abundance viruses. Overall, this study
24 highlights the importance of aligning wet lab and bioinformatic methods with specific goals when
25 employing probe-capture enrichment for human virus sequencing from wastewater.

26 Keywords

27 Targeted sequencing, probe-capture enrichment, human virus, wastewater-based surveillance,
28 wastewater-based epidemiology, virus concentration, nucleic acid extraction

29 Synopsis (~ 30 words)

30 Four concentration/extraction methods combined with probe-capture sequencing of human viruses in
31 raw wastewater were compared. Innovaprep ultrafiltration with solids removal had the best
32 performance for human virus detection sensitivity, richness, and recovery of near-complete genomes.

33 1. Introduction

34 Wastewater-based epidemiology (WBE), previously employed for monitoring enteric viruses like polio ¹,
35 has been widely applied during the COVID-19 pandemic. In 2020, the US Centers for Disease Control and
36 Prevention (CDC) launched the National Wastewater Surveillance System (NWSS), to build and
37 coordinate the capacity for WBE as a component of the nationwide monitoring of SARS-CoV-2 ².
38 Subsequently, groups around the world have expanded WBE to include PCR-based monitoring of known
39 seasonal respiratory viruses including RSV and Influenza A and B, and new PCR panels are expected to
40 contribute to CDC NWSS ³.

41 Unlike PCR-based virus quantification, sequencing of viruses in wastewater has the potential to monitor
42 many human viruses at the genome level simultaneously. Reference-based amplicon sequencing using
43 tiled panels such as ARTIC SARS-CoV-2 ⁴, ARTIC HAdV-F41 ⁵, Swift Normalase™ Amplicon Panel ⁶, or
44 targeted amplicons like those for the VP1 or VP4 regions of enterovirus ^{7, 8}, have enabled subtyping and
45 tracking of circulating variants and strains, providing evidence that wastewater data aligns with available
46 clinical data ^{4, 5}. However, amplicon-based sequencing is limited in its ability to detect novel viruses, due
47 to the challenges of degenerate primer design and multiplexing. In contrast, deep untargeted
48 sequencing offers a comprehensive view of viral diversity in wastewater ⁹⁻¹¹, but human viruses
49 constitute a minimal fraction of the microbial nucleic acids present in wastewater, approximately
50 0.011% of unique reads ¹⁰ or 0.1% of the assembled contigs ¹¹. To increase sequencing coverage of
51 human viruses and to allow the detection of divergent or novel viruses in wastewater, probe-capture
52 enrichment panels have been adopted from clinical research ¹². Here, probes hybridize to DNA targets in
53 a sample, allowing downstream separation of targets from background DNA. Because probe
54 hybridization allows more mismatches than primer binding during PCR, more divergent sequences may
55 be enriched by probe capture, potentially including novel relatives of known viruses. Recent studies that
56 have applied virus probe-capture panels to wastewater-derived samples reported an increase in the
57 proportion of viral reads up to 81% compared to untargeted sequencing ¹³. Although probe-capture-
58 based sequencing enriched human viruses, most of the recovered viral content (> 80%) still consisted of
59 bacteriophages and plant viruses ^{14, 15}. These findings indicate that probe capture panels are still limited
60 in their ability to enrich target sequences in samples with large amounts of background/non-target
61 sequences, suggesting that the choice of upstream sample processing method may affect the detection
62 of human viruses.

63 Prior to the COVID-19 pandemic, sequencing-based wastewater virus studies relied on large-volume
64 time-intensive methods that had initially been developed to culture infectious viruses (e.g., polyethylene
65 glycol precipitation, skim milk flocculation, ultracentrifugation, and membrane filtration). Multiple
66 studies reported that the choice of concentration method influenced the resulting virus profiles by
67 untargeted sequencing ^{9, 16} and few studies reported any sequences from enveloped viruses. During the
68 pandemic, the demand for rapid routine monitoring of SARS-CoV-2 led to the development and wider
69 adoption of streamlined concentration/extraction methods with lower sampling volumes, ending with
70 qPCR or digital PCR quantification ^{17, 18}. These methods included size separation (e.g., Innovaprep
71 ultrafiltration pipette, centrifugal ultrafiltration), capture based on virus surface characteristics (e.g.,
72 Nanotrap beads, electron-negative HA membrane), and direct nucleic acid extraction (e.g., Promega
73 Wizard Enviro large-volume extraction, or extraction of wastewater solids after centrifugation). These
74 routine monitoring methods were also used to obtain SARS-CoV-2 RNA for sequencing, with varying
75 success ¹⁹⁻²² and later extended for detection of a wider spectrum of viruses ^{14, 15, 23-25}. To date, few
76 studies have directly compared the effects of different methods on the success of virus probe-capture

77 enrichment sequencing. McCall et al. (2023) compared methods with very different sampling volumes
78 (300 μ L for direct extraction and 50 mL for HA filtration) and suggested that direct extraction may yield a
79 lower equivalent volume of viruses in the final extracted nucleic acid compared to pre-filtered samples
80 ²³. Spurbeck et al (2023) indirectly compared five wastewater virus concentration/extraction methods,
81 but each was applied to wastewater samples from a different location(s). They found that Innovaprep
82 ultrafiltration yielded the highest virus sequence recovery in untargeted RNA sequencing, although most
83 sequences corresponded to bacteriophage ²⁴. These findings highlight the potential impact of
84 concentration/extraction methods on targeted sequencing of diverse viruses, but direct comparisons
85 and analysis of potential biases from concentration methods on sequencing performance are needed,
86 especially for targeted sequencing using probe-capture panels.

87 In this study, four wastewater virus concentration/extraction methods were selected based on their
88 ongoing use in wastewater surveillance efforts, and the success of probe-capture enrichment
89 sequencing was compared for each method. The wastewater input volume was held constant, and the
90 resulting nucleic acids were enriched using the Illumina virus surveillance panel (VSP). The evaluation of
91 methods performance included total nucleic acid quality, unique sequence output, taxonomic
92 composition, richness, recovered genome completeness, and sensitivity comparisons between
93 sequencing and dPCR. Ultimately, these findings improve our understanding of wet lab approaches and
94 their compatibility with virus probe-capture enrichment and sequencing, informing tailored responses to
95 emerging viral threats.

96 2. Materials and Methods

97 2.1 Sample collection

98 Influent wastewater was collected as 24-hour composite samples on three dates: March 1st, April 19th,
99 and April 26th, 2023, from the EBMUD wastewater treatment plant (Alameda County, CA). This facility
100 serves approximately 700,000 people, receiving domestic and industrial wastewater. On each date, the
101 sample was transported to the laboratory on ice, and twelve 40 mL aliquots were prepared. Bovine
102 coronavirus (BCoV) was added to each tube as a sample processing control to assess viral RNA recovery.
103 First, one vial of BCoV (Merck) vaccine powder was resuspended in 2 mL 0.1 mM Tris-
104 Ethylenediaminetetraacetic acid (TE) buffer and diluted 10-fold. Each wastewater aliquot was spiked
105 with 50 μ L BCoV solution and incubated overnight at 4°C.

106 2.2 Concentration and extraction

107 Four concentration and extraction methods (described below) were employed in this study: Innovaprep
108 pipette concentration (**IP method**), Nanotrap bead concentration (**NT method**), Promega large-volume
109 direct extraction (**PMG method**), and pelleted solids direct extraction (**Solids method**). Each method was
110 performed on three 40-mL aliquots of wastewater per sample date, alongside a negative control
111 consisting of 40 mL 1x phosphate-buffered saline (PBS) solution (**Table S1**). All methods resulted in 100
112 μ L purified total nucleic acid (TNA).

113 In the first two methods, separate concentration and extraction steps were applied. In the **IP method**,
114 400 μL of 5% Tween 20 was added to the wastewater sample and mixed by inversion, followed by
115 centrifugation at 7000 x g for 10 min. The supernatant was ultrafiltered using the automatic HF
116 Concentration Pipette (Innovaprep CP-Select™) and eluted with the elution fluid (Innovaprep) to
117 produce the viral concentrate (ranging from 160 to 882 μL , **Table S1**). TNA was then extracted from up
118 to 200 μL of viral concentrate using the Allprep PowerViral DNA/RNA kit (Qiagen) and eluted in 100 μL ,
119 following the manufacturer's liquid sample extraction protocol. The **NT method** followed the Nanotrap®
120 Microbiome A Protocol, compatible with AllPrep PowerViral DNA/RNA Kit (APP-091 December 2022).
121 Briefly, 115 μL of Nanotrap® Enhancement Reagent 2 (ER2) and 600 μL of Nanotrap Microbiome A
122 Particles (Ceres Nanosciences) were sequentially added to each sample, followed by mixing and
123 incubation. The beads were separated from the solution on a magnetic rack and resuspended in 1 mL of
124 molecular-grade water, followed by another separation using the magnetic rack. The beads were then
125 mixed with 600 μL of preheated PM1 + Beta-mercaptoethanol solution from the Allprep PowerViral kit,
126 and the mixture was heated at 95°C for 10 min to release nucleic acids. Beads were removed using the
127 magnetic rack and the supernatant was used for subsequent extraction steps using the Allprep
128 PowerViral DNA/RNA kit liquid protocol, resulting in 100 μL of final TNA.

129 The other two methods were direct extractions from either total wastewater or pelleted solids. The
130 **PMG method** used the commercial kit from Promega (Wizard® Enviro Total Nucleic Acid) following the
131 manufacturer's protocol. Briefly, 0.5 mL of protease was added to each 40 mL wastewater sample and
132 incubated for 30 minutes. After centrifugation at 3000 x g for 10 min, binding buffers and isopropanol
133 were added to the resulting supernatant before passing it through the PureYield™ binding column. The
134 bound nucleic acids were washed and then eluted in 1 mL of nuclease-free water. The eluted samples
135 were further purified, concentrated, and eluted using the PureYield™ Minicolumn, resulting in a final
136 total nucleic acid volume of 100 μL . In the **Solids method**, the 40 mL wastewater sample was centrifuged
137 at 20,000 x g for 10 minutes to pellet the solids. Total nucleic acid was then extracted from 0.2 g (wet
138 weight) of solid pellets using the Allprep PowerViral DNA/RNA extraction kit. This followed the
139 manufacturer's solids extraction protocol, which included a 10-minute bead-beating step after the
140 addition of PM1 and Beta-mercaptoethanol solution. The final extracted TNA were eluted in 100 μL of
141 nuclease-free water.

142 DNA and RNA concentrations were quantified using the Qubit 1X dsDNA HS Assay (Fisher Scientific) and
143 Qubit RNA HS Assay (Fisher Scientific), respectively. Aliquots of all extracts were stored at -20°C and
144 quantified by dPCR within one week and at -80°C for subsequent sequencing library preparation.

145 **2.3 Digital PCR quantification of SARS-CoV-2 in the extracted total nucleic acid**

146 Digital PCR was performed on the QIAcuity Four Platform Digital PCR System (Qiagen). The details of
147 SARS-CoV-2 and BCoV assays' primers, probes, and thermal cycling conditions are summarized in **Table**
148 **S2a**. The reaction mixtures were prepared using the QIAcuity OneStep Advanced Probe Kit (Qiagen) and
149 loaded onto either 8.5k 24-well or 26k 24-well nanoplates (Qiagen). Details of reaction mixture
150 composition and volumes were summarized in **Table S2b**. The positive control was linearized gene
151 plasmids from Integrated DNA Technologies, and the negative control was nuclease-free water. See
152 **Figure S1** for examples of the partition fluorescence plots of positive and negative controls. The number
153 of valid partitions ranged from 7,920 to 8,269 per well for 8.5k plates and 12,548 to 25,493 per well for

154 26k plates. Data were analyzed using the QIAcuity Suite Software V1.1.3 (Qiagen, Germany) with
155 automated settings for threshold and baseline, followed by manual inspection. Results were plotted
156 using a customized Python script. dMIQE checklists²⁶ are provided in **Table S3**. The operational limit of
157 detection was treated as ≥ 3 positive partitions per well.

158 **2.4 Library preparation and targeted sequencing**

159 Before library preparation, DNA and RNA quality were measured by Fragment Analyzer with the default
160 HS NGS Fragment 1-50 kb assay and Bioanalyzer (Agilent 2100) with the Agilent RNA 6000 Pico RNA
161 assay, respectively. Library preparation followed the Illumina RNA Prep with Enrichment kits with
162 modifications to total input (Illumina, San Diego, CA, USA). In brief, the mixture of purified DNA and RNA
163 from samples collected on April 19 and April 26 was diluted with nuclease-free water such that the final
164 concentration of RNA was ≤ 100 ng/ μ l. Dilution was not conducted for IP and NT samples due to the low
165 RNA concentrations. The DNA and RNA from samples collected on March 1 were used directly as the
166 input for library preparation without dilution for all concentration/extraction methods (**Table S1**). Next,
167 8.5- μ L of each sample was denatured followed by first- and second-strand DNA synthesis. Tagmentation
168 of the total enriched double-stranded cDNA was performed using bead-linked transposons (BLT), and
169 adapter sequences were added at the same time. The resulting fragments were purified and amplified
170 to add index sequences. Libraries were quantified using the Qubit dsDNA broad-range Assay Kit.
171 Enrichment was performed with the Illumina VSP Panel by pooling 200 ng of each library from three
172 biological replicates into hybridization reactions. This step was followed by bead-based capture of
173 hybridized probes, amplification, clean-up, and quantification of the final enriched library. After library
174 preparation, all enriched samples were pooled in equimolar ratios and sequenced on one lane of
175 Illumina Novaseq 6000 SP 150PE.

176 **2.5 Bioinformatics Analysis Pipeline**

177 Sequence data were quality trimmed using BBduk²⁷ to remove adaptors and filter out low-quality reads
178 and short reads. Seqkit was used to deduplicate reads and summarize unique reads²⁸ (**Figure 1b**).
179 Before taxonomy classification, human reads were filtered using bowtie2 (v2.5.1)²⁹ by mapping to
180 GRCh38.p14 (RefSeq GCF_000001405.40) and CHM13v2.0 (RefSeq GCF_009914755.1). The remaining
181 non-human unique reads were classified by Centrifuge (1.0.4)³⁰ and Recentrifuge³¹ using a
182 decontaminated version of NCBI-nt database (NCBI release date June 5, 2023). A minimum hit length
183 (MHL) threshold of 15 was employed for Centrifuge. An MHL threshold of 40 was subsequently applied
184 in Recentrifuge for downstream analysis. After classification, all viral reads were extracted from each
185 sample using reextract and viral sequence similarities between samples were compared using MASH³².
186 Pairwise Mash distances were calculated for the construction of the PCoA plot using the
187 sklearn.decomposition PCA package in Python. A PERMANOVA test with 999 permutations was
188 performed using the vegan package (2.6.4) in R³³. One sample (PMG_426_2) displayed distinct
189 sequence properties from the other two biological replicates in the original PCoA (**Figure S2**) and yielded
190 unexpectedly low unique read counts (**Table S1**), likely due to unsuccessful enrichment during the
191 library preparation. This sample was excluded from all sequencing analyses. To precisely identify SARS-
192 CoV-2 reads during the sampling period, unique reads classified by Centrifuge at the species level as
193 severe acute respiratory syndrome-related coronavirus (taxID: 694009) were extracted and mapped to
194 references from the GISAID database³⁴ downloaded on January 2, 2024 (**Table S4**). The references

195 comprised 463 complete genome sequences with high coverage and collection dates ranging from
196 January 1, 2023, to May 31, 2023. The mapped reads were subjected to additional filtering using
197 reformat.sh from BBduk²⁷. Mapped reads with fewer than 5 mismatches were considered as classified
198 SARS-CoV-2 reads.

199 To determine putative virus host assignments for each read (**Figure 2c and 2d**), the NCBI taxonomy
200 database³⁵, which includes virus host information, was queried with the NCBI taxID of the best hit given
201 by Recentrifuge, and the results were manually inspected (see SI methods for details). The comparison
202 of DNA and RNA viruses was conducted at the virus kingdom level based on taxonomic classification
203 results. To focus specifically on human viruses, the identification of all human viruses (section 3.3)
204 occurred at the species level based on NCBI taxonomy³⁵. Species-level richness was determined by
205 counting the unique human virus species, with a cutoff of > 10 classified reads applied to discard low-
206 abundance viruses (**Figure S3a and 3b**). Some reads were assigned to species-level NCBI taxIDs that
207 paradoxically lacked clear species-level taxonomy names in the database. To make this apparent, these
208 species are displayed with “unclassified” appended to the species name (see SI methods for details).

209 To maximize the recovery of near-complete viral genomes for wet lab method comparison, all unique
210 reads were separately assembled using SPAdes with the -meta option (v3.15.5)³⁶. All virus scaffolds
211 identified by VirSorter2³⁷ were further subjected to quality filtering, requiring a length of > 1000 bp and
212 an average coverage of > 10x. These filtered assemblies were then subjected to BLASTn search against
213 the NCBI nt virus database, with stringent quality filters applied: > 80% identity, > 90% alignment/query
214 length, and an e-value < 1E-8. The best hit based on bitscore was retained for each assembled scaffold
215 and information including virus name, taxID, genome completeness, and genome length was retrieved
216 from NCBI via the dataset and dataformat functions³⁸. Hit genomes were retained only if complete, and
217 assembled genomes were used for further analysis if the alignment length was > 70% of the complete
218 hit genome, indicating the assembly of a near-complete genome from wastewater (**Figure S4b**).

219 Scaffolds representing near-complete genomes for JC polyomavirus were collected for phylogenetic
220 analyses. Potential assembly errors were inspected by mapping reads to assembled scaffolds using
221 bowtie2 and visualizing with the Integrative Genomics Viewer (IGV)³⁹. No assembly errors were
222 detected, and representative mappings are shown in **Figure S5**. Given the circularity of the JC
223 polyomavirus genome, assemblies were also examined in Geneious⁴⁰, and repeated regions at the
224 beginning and the end of the sequences were trimmed before the alignment (**Figure S6**). Multiple
225 sequence alignment was performed by MUSCLE⁴¹ with all trimmed scaffolds, all JC polyomavirus
226 reference genomes from NCBI GenBank released within two years (n=39), and the best-hit results from
227 BLASTn for each scaffold. The alignment was inspected in Geneious to identify a common starting point,
228 and all 10 scaffolds were recircularized to this point. The recircularized scaffolds were queried against
229 NCBI again to identify new best-hit reference genomes, which may have changed due to genome
230 curation. The final dataset included these curated genomes, new best-hits, and the 39 JC polyomavirus
231 references. Alignment was performed with MUSCLE followed by GBlocks⁴² to identify informative
232 regions, and MEGA 11.0⁴³ was used to generate and visualize the final maximum likelihood tree using
233 the Tamura Nei model with 100 bootstrap replicates.

234 **2.6 Statistical analysis and data availability**

235 The normality of data was assessed using the Shapiro-Wilk test. Statistical differences between
236 concentration and extraction methods were evaluated using the Kruskal-Wallis test, followed by post-
237 hoc pairwise Dunn's test. All statistical tests were performed using the Python package `scipy.stats`, and
238 significance was determined at a 95% confidence interval ($p < 0.05$). Sequencing data for this project for
239 this project has been deposited in the NCBI Sequence Read Archive (SRA) under accession number:
240 SUB13892842 and Bioproject ID: PRJNA1047067. The processed data, reproducible code, and the
241 analysis workflow are available at <https://github.com/mj2770/Wastewater-virus-surveillance->.

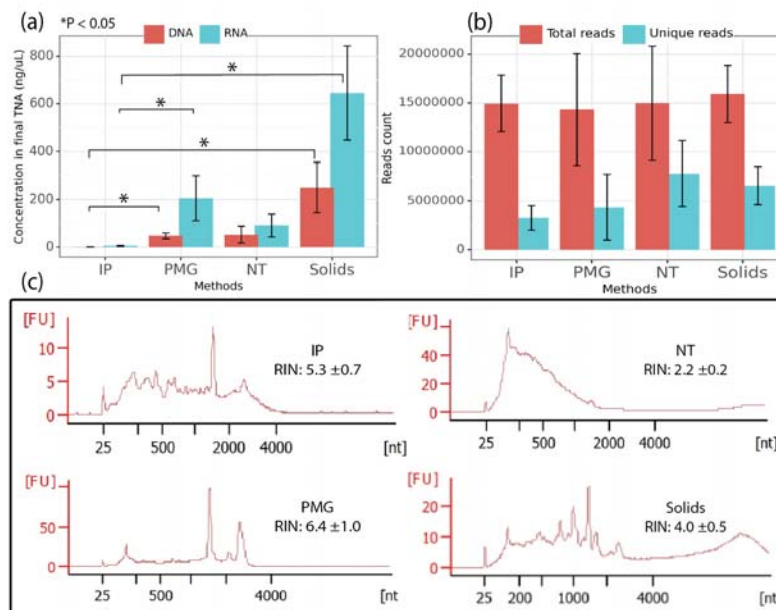
242 3. Results and Discussion

243 In this study, wastewater influent was collected from a single WWTP on three dates, and viruses were
244 concentrated and extracted by four methods: **IP method** (Innovaprep ultrafiltration of liquid portion
245 paired with a small-volume extraction kit), **NT method** (Nanotrap beads-based affinity capture
246 performed on total influent paired with small-volume extraction kit), **PMG method** (Promega large
247 volume direct extraction), and **Solids method** (centrifugation paired with small-volume extraction kit).
248 The resulting 36 samples (12 samples in biological triplicate) were processed using the virus surveillance
249 panel (VSP) from Illumina using probe-capture enrichment. Following the initial analysis, an outlier
250 sample was identified, indicating unsuccessful library preparation (see Methods), and this sample was
251 excluded from all analyses.

252 3.1 Sample quality and sequence data

253 The DNA and RNA generated using the four methods differed in concentration (Kruskal-Wallis test $p =$
254 $2E-6$ and $7E-7$, respectively), fragment size distribution, and RNA integrity (ANOVA test $p = 1E-13$). The
255 Solids method consistently resulted in yields that were higher than other methods for both DNA and
256 RNA (**Figure 1a**), while the IP method, which includes a solids removal step, resulted in significantly
257 lower total DNA and RNA yield compared to Solids and PMG (**Figure 1a**, IP v.s. Solids DNA $p = 3E-7$, IP
258 v.s. PMG DNA $p = 0.02$, IP v.s. Solids RNA $p = 3E-7$, IP v.s. PMG RNA $p = 0.004$). All methods yielded a
259 higher concentration of RNA than DNA, but the resulting ratios of RNA:DNA varied significantly (Kruskal-
260 Wallis test $p = 0.002$) across methods from 2.0 ± 0.7 (for NT) to 4.3 ± 1.6 (for PMG). Unlike the other
261 methods, shorter RNA fragments were observed with the NT method and 16S rRNA and 23S rRNA were
262 absent, perhaps accounting for the low RNA:DNA ratio. The lack of ribosomal RNA may be due to the
263 exclusion of bacteria by the nanotrap hydrogel particle shells, which have specific pore sizes and are
264 chemically modified to prevent the entry and capture of large or non-targeted particles⁴⁴. Although viral
265 RNA integrity is not discernible from the RNA Integrity Number (RIN) alone, the highest RIN was
266 observed with the PMG method (6.4 ± 1.0 , **Figure 1c**), which suggested more intact prokaryotic RNA was
267 preserved with the PMG method.

268 After sequencing 36 samples, a total of 535 million reads were generated, averaging 14.86 ± 4.46 million
269 reads per sample (**Figure 1b**), and the removal of PCR duplicates reduced read counts by over 50% for all
270 samples. As the IP method produced the lowest RNA and DNA input concentrations, it was not
271 surprising that after deduplication these samples also retained significantly fewer unique reads ($3.3 \pm$
272 1.3 million, **Figure 1b**) compared to samples from the Solids and NT methods (IP vs. NT $p = 0.005$, IP vs.
273 Solids $p = 0.04$). Nonetheless, the count of unique reads was not clearly related to the DNA and RNA
274 concentrations, perhaps due to the dilution of nucleic acids (**Table S1**) before library preparation, and
275 the multiple amplification and equimolar pooling steps during library preparation.

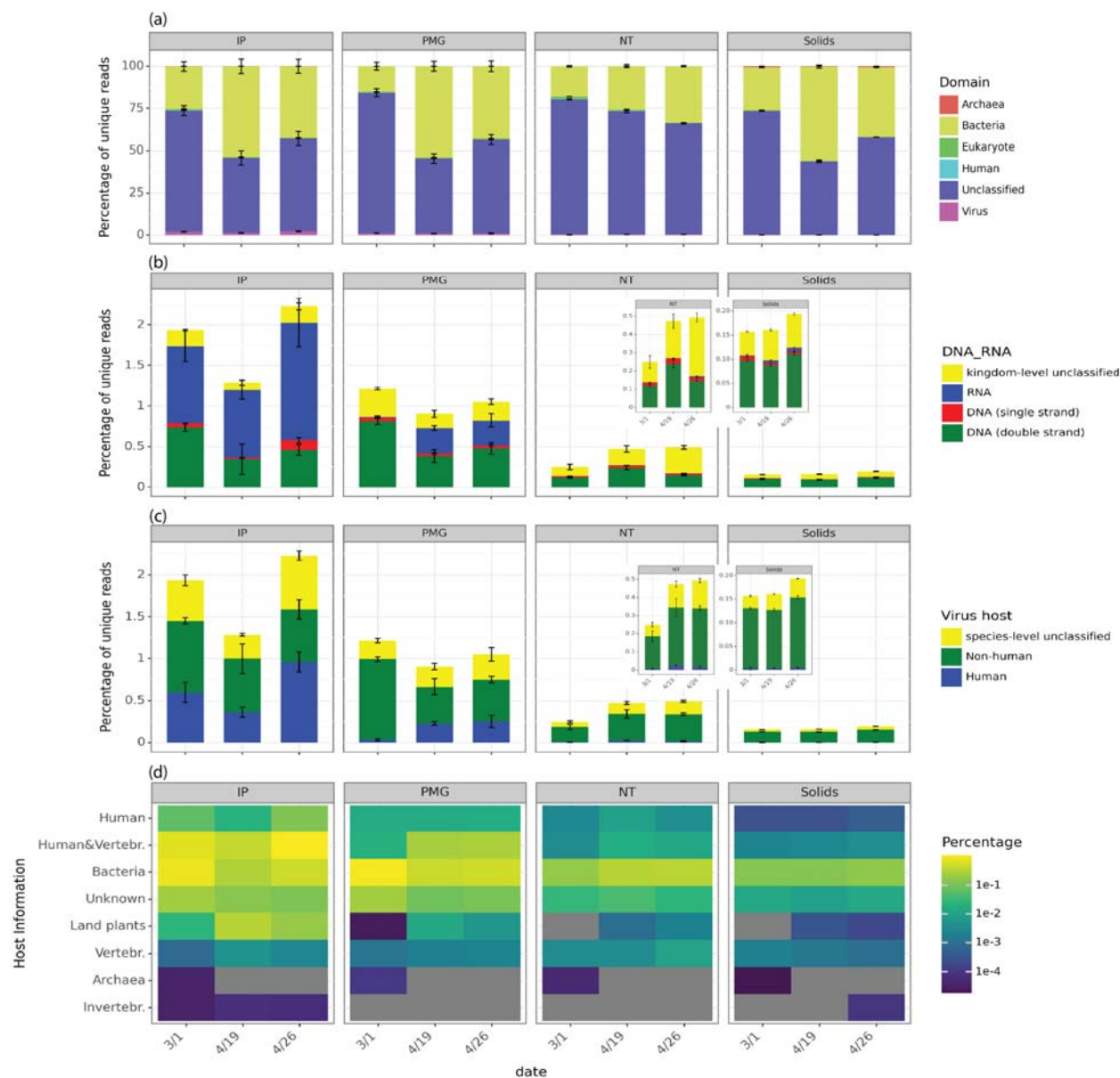


276

277 **Figure 1. Nucleic acids and unique read counts by sample processing method.** (a) Averaged concentrations of
278 extracted DNA and RNA produced by each method (n=9 samples per method); (b) Averaged raw read counts and counts of
279 unique reads after QC trimming and deduplication in each method (n = 9 samples for IP, NT, and Solids, n = 8 for PMG); (c)
280 Representative RNA fragment size distribution and average RNA integrity number (RIN) for each method. Note that samples
281 were diluted before fragment analysis (IP: undiluted, NT: 25x, PMG: 25x, Solids: 200x), so y-axes are not comparable.

282 3.2 Taxonomic classification and virus composition similarity

283 Over 40% of unique reads were not taxonomically classified by Recentrifuge at the Domain level with
284 the selected Minimum Hit Length (MHL) across all methods, and most classified reads were assigned to
285 the domain Bacteria (ranging from $25.84 \pm 6.81\%$ to $40.88 \pm 13.13\%$, **Figure 2a**). It is likely that a larger
286 proportion of unique reads would have received an assigned taxonomy at a lower classification
287 stringency; however, such low-confidence assignments have the potential to introduce substantial noise
288 to downstream assessments. Future functionalization of these platforms will require tuning of these
289 stringency thresholds for the desired application, balancing classification sensitivity with assignment
290 confidence. These findings could also reflect the current limitations of reference-based classifiers⁴⁶ and
291 limited enrichment of targets using probe-capture, irrespective of the concentration and extraction
292 methods employed.



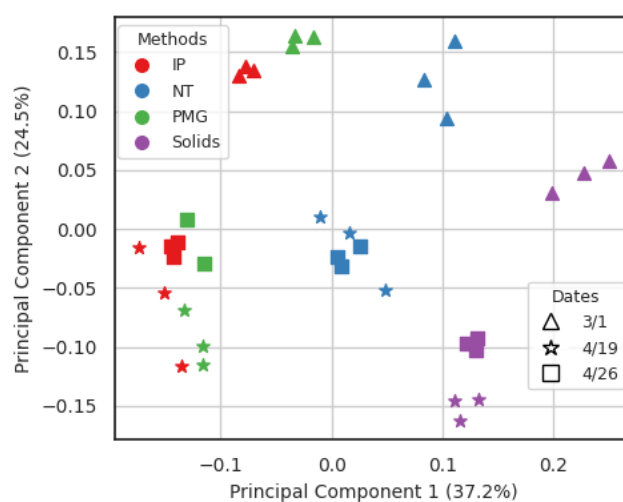
293

294 **Figure 2. Taxonomic profiles of reads and virus hosts differed by method.** (a) Domain-level classification of
 295 unique reads by Recentrifuge, with samples collected on three sampling dates and processed by four methods (n=3, except
 296 Promega 4/26). “unclassified” is the sum of reads discarded by Recentrifuge without taxonomic classification and those
 297 classified as “Root” but without a domain-level classification. “Human” represented unique reads mapped to two downloaded
 298 human genomes (see methods); (b) Percentages of unique reads identified as RNA, double-strand DNA, and single-strand DNA
 299 viruses based on kingdom-level virus classification; (c) Percentages of unique reads identified as virus species linked to human
 300 and non-human hosts in NCBI or for which species-level taxonomy was not determined; (d) Percentages of unique viral reads
 301 associated with different host categories in the NCBI Virus database. Note that “human” in (c) encompasses the categories
 302 “human & vertebrates” and “human” in (d). In (d), reads assigned to BCoV were subtracted from counts of reads assigned to
 303 “human & vertebrates” and are not displayed.

304 The percentage of reads classified as viral ranged from $0.17 \pm 0.02\%$ (Solids) to $1.82 \pm 0.46\%$ (IP) of
 305 unique reads across different methods (**Figure 2b**), surpassing the reported $< 0.011\%$ in untargeted
 306 sequencing⁹. The IP samples yielded significantly higher percentages of viral reads than Solids and NT
 307 ($1.82 \pm 0.46\%$, **Figure 2b**, IP vs. Solids $p = 8E-7$, and IP vs. NT $p = 0.004$), followed by the PMG samples

308 (1.06 ± 0.18%, Figure 2b, PMG vs. Solids p = 0.002). Additionally, the IP method concentrated
309 significantly more RNA viruses (**Figure 2b**) and viruses associated with human and/or vertebrate hosts
310 than NT and Solids methods (0.64 ± 0.27% human viruses in total unique reads from IP, **Figures 2c and**
311 **2d**, IP vs. NT p = 0.002, IP vs. Solids p = 1E-6). The IP and PMG methods incorporated a solids removal
312 step after attempting to release solid-associated viruses by adding 5% Tween 20⁴⁵ or protease,
313 respectively⁴⁶. These steps not only prevent clogging during sample processing but also strike a balance
314 between eliminating solid-associated non-viral microorganisms like bacteria and attempting to retain
315 viruses. As a result, a notably lower ratio of classified bacterial reads to classified viral reads was
316 observed in IP and PMG samples (25 ± 14:1 and 38 ± 24:1, respectively) in comparison to Solids and NT
317 samples (241 ± 83 and 66 ± 12, respectively) (IP vs. NT p = 0.04; IP vs. Solids p = 1E-5; PMG vs. Solids p =
318 0.0006). In NT and Solids samples, most viral reads were associated with bacterial hosts, based on the
319 NCBI taxonomy database (**Figure 2d**). This finding is consistent with the high fraction of DNA viruses in
320 those samples (**Figure 2b**), as most bacteriophages are DNA viruses⁴⁷.

321 To compare virus composition across the four sample preparation methods, reads classified as viral by
322 Recentrifuge were extracted from each sample, and MASH³² was used to assess pairwise sequence
323 similarity. In a principal component analysis using MASH distances, triplicate samples clustered together
324 (PERMANOVA test p = 0.985), while all samples were separated by concentration/extraction methods
325 along PC1 (37.2% of the variation, **Figure 3**, PERMANOVA test p = 0.001). Specifically, IP- and PMG
326 samples clustered together, while NT and Solids samples were distinct (**Figure 3**). The predominance of
327 bacteriophage in both NT and Solids samples likely contributed to their differentiation from the other
328 two methods. Samples were separated by sampling dates along PC2 (24.5% of the variation, **Figure 3**,
329 PERMANOVA test p = 0.001), with samples from March 1, 2023, differing from those collected on April
330 19 and April 26. This differentiation was observed consistently across all four methods. These temporal
331 shifts in virus composition may suggest a temporally variable metavirome composition in wastewater,
332 potentially influenced by changes in circulating viruses^{8, 48, 49} and changing wastewater conditions, such
333 as flow rate, total suspended solids (TSS), total organic compounds (TOC), and the abundance of
334 antagonistic microorganisms^{50, 51}.

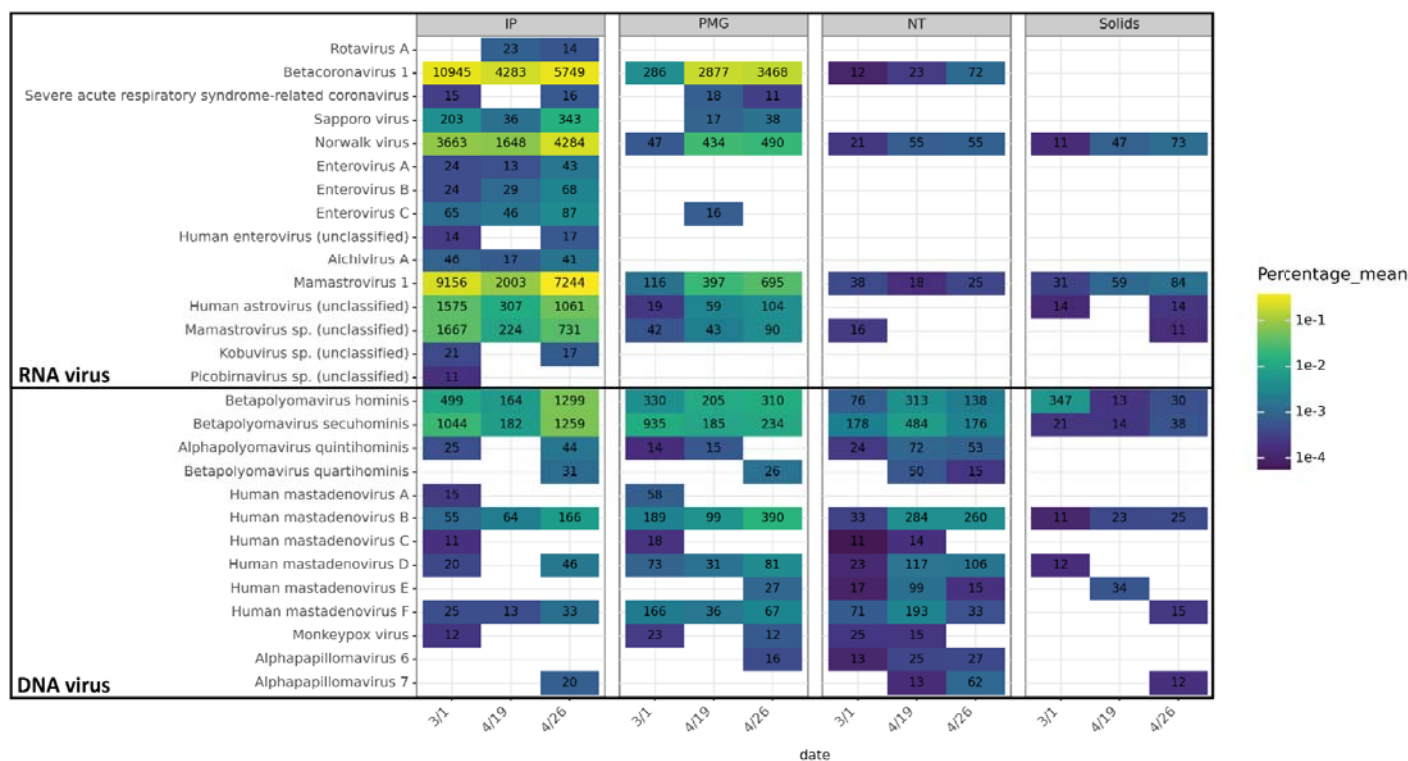


335
336 **Figure 3. Viral sequence composition was influenced by wastewater virus concentration/extraction**
337 **method and sample date.** Principal component analysis (PCoA) plot was generated using the MASH distance, which was

338 calculated based on sequence similarity among all reads classified as viral by Centrifuge. Different methods are represented by
 339 colors, and different sampling dates are represented by shapes.

340 3.3 Human virus species richness and composition

341 PMG and IP methods yielded higher species-level richness of total viruses detected with >10 reads (241
 342 and 176 viruses, respectively) and human viruses (20 and 26 respectively) compared to NT and Solids
 343 (**Figure S3a**), although total read depth was similar for all samples (**Figure 1b**, $p = 0.44$). Thus, removing
 344 solids after releasing solid-associated viruses did not compromise the richness of detected human
 345 viruses. Conversely, including solids produced lower species-level diversity. Of the 66 virus “groups” of
 346 high public health significance listed as targets in the Illumina VSP panel (**Table S5**), IP samples detected
 347 members of 11 (**Figure S3a**). These included human coronavirus-OC43 (hCoV-OC43), adenovirus,
 348 astrovirus, aichivirus, enterovirus, norovirus, coxsackievirus, rotavirus, salivirus, and sapovirus, as well as
 349 mpox (**Figure S3b**), though the exact list of species and strains used by Illumina for probe design is
 350 proprietary; we note that enteroviruses are a diverse group which contains coxsackieviruses, while
 351 hCoV-OC43 is a sub-species level category.



352
 353 **Figure 4. Relative abundance of human virus species in each sample.** Fill indicates the average percent
 354 relative abundance of each virus species in total unique reads across triplicate samples, based on Recentrifuge
 355 read classification. Species with fewer than an average of 10 reads per sample are not shown. Text in each cell
 356 indicates the average read counts assigned to the species for each sample. Viruses are grouped by genome type.
 357 NCBI taxIDs corresponding to species without names (e.g. “sp.”) are appended with “(unclassified)” (see
 358 supplementary methods). Note that Betacoronavirus 1 includes the spike-in bovine coronavirus.

359

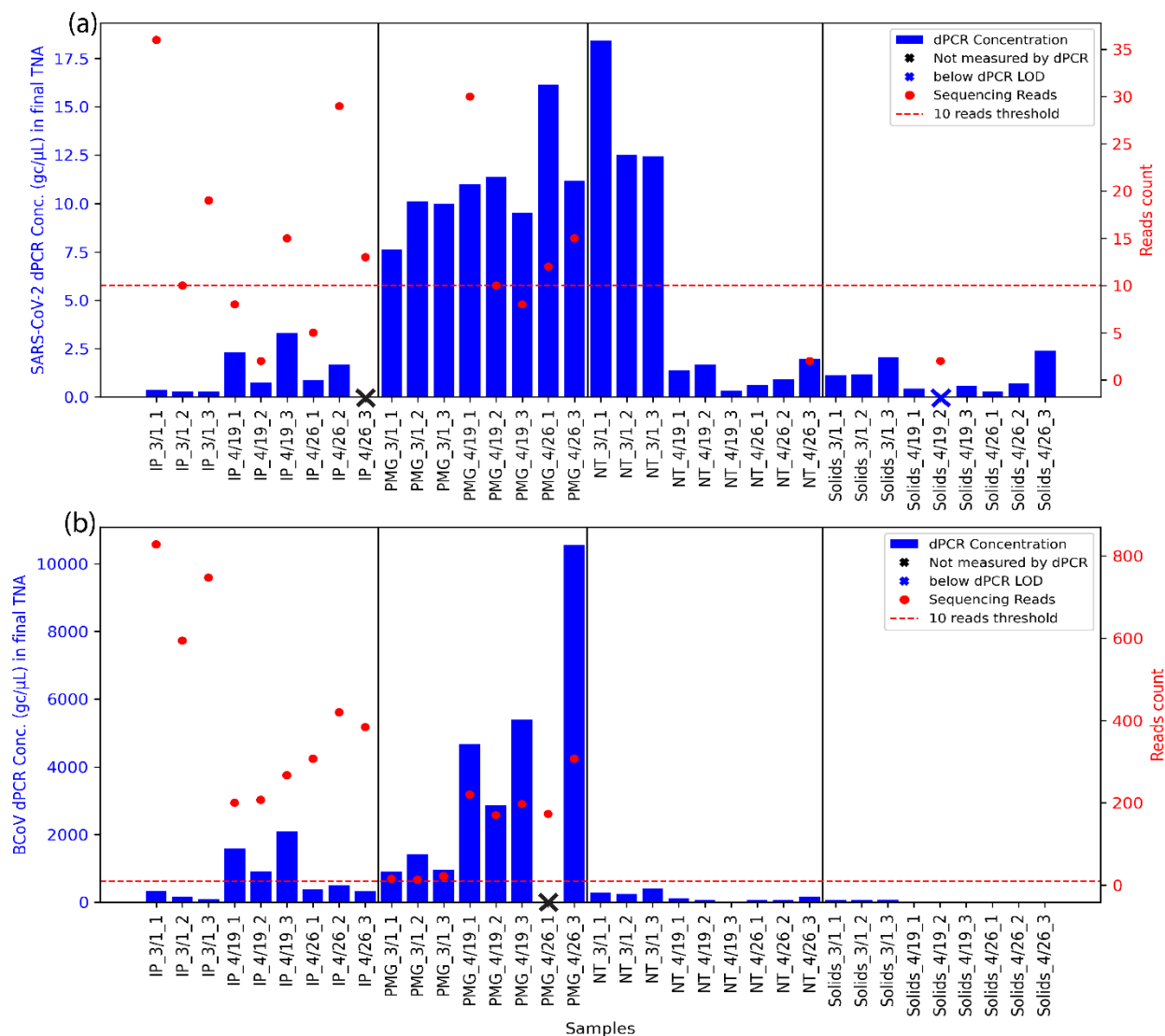
360 All human virus species detected (>10 reads per species) in at least one sample were compared across
361 the four methods (**Figure 4**). Some viruses were consistently detected by all methods, including human
362 polyomavirus, mastadenovirus, mamastrovirus 1, and norwalk virus, which are known to be shed at high
363 concentrations in human waste^{5, 9, 10, 13, 23, 48, 52-55}. RNA virus species, including severe acute respiratory
364 syndrome-related coronavirus, sapporo virus, and enteroviruses were not detected in NT and Solids
365 samples. Different trends were also observed among virus species within the same genus. For instance,
366 human mastadenovirus B, D, and F were detected in all samples, while human mastadenovirus A, C, and
367 E were not detected in certain samples (**Figure 4**). This variability suggested that related virus species
368 may be differentially detected by different concentration methods. No arthropod-transmitted viruses
369 (e.g., Dengue, Chikungunya), bloodborne viruses (e.g., Hepatitis virus and HIV), or hemorrhagic fever-
370 related viruses (e.g., lassa mamarenavirus, junin virus, etc.) were detected, despite their inclusion in the
371 probe panel. Mpox, detected intermittently in wastewater since the outbreak in 2022^{56, 57}, was detected
372 at low levels in IP, PMG, and NT samples. Overall, it seems reasonable to conclude that these results
373 generally reflect a subset of current infectious diseases present in and absent from the San Francisco
374 Bay Area at the time of sample collection.

375 **3.4 Potential of recovering near-complete human virus genomes**

376 Seven near-complete human virus genomes were assembled from IP samples, the most from any
377 concentration/extraction method (**Figure S4b**). This aligned with the high numbers of total virus and
378 human virus reads in these samples ($59,965 \pm 28,180$ and $20,242 \pm 9,294$, respectively, **Table S1**). No
379 near-complete human virus genomes were obtained from Solids-extracted samples (**Figure S4b**) likely
380 due to insufficient reads for total viruses and human viruses ($11,043 \pm 2,720$ and 213 ± 99 , respectively,
381 **Table S1**). These results highlight the need to understand the minimum sequencing depth in relation to
382 the proportion of viral reads required for the assembly of high-quality virus genomes.

383 JC polyomavirus composite genomes were assembled in samples from three concentration/extraction
384 methods (IP, PMG, and NT) and multiple replicates (**Figure S4b**). The recovery of JC polyomavirus
385 genomes is perhaps unsurprising given that approximately 40% of the population sheds the virus
386 through urine⁵⁴. Also, as a non-enveloped DNA virus with a circular genome, JC polyomavirus is highly
387 resistant to environmental stress and exonuclease activity⁹. Ten scaffolds classified as near-complete JC
388 polyomavirus genomes were used for phylogenetic analysis. At least one subtype of JC polyomavirus 3
389 was present (Node 1353 NT_301_1), affiliated with clades from South Africa (**Figure S7**). Although other
390 scaffolds were clustered together, they exhibited relatively low node support values (< 50); likely several
391 of these scaffolds represent the same JC polyomavirus population in replicate wastewater samples, with
392 variations in the composite assembly. These results, and those from other recent studies²⁵,
393 demonstrated that probe capture enrichment can yield whole genomes of high-abundance viruses for
394 phylogenetic analysis, which may be useful for identifying novel virus strains in the future.

395 **3.5 The comparison of SARS-CoV-2 detection between dPCR and targeted sequencing using**
 396 **the VSP panel**



397
 398 **Figure 5. Detection sensitivity comparison between dPCR and reads-based classification (Recentrifuge)**
 399 **of sequencing results.** (a) SARS-CoV-2 detection comparison; (b) BCoV detection comparison. Blue bars on the left y-axis
 400 represent the virus concentration measured by dPCR in the final total nucleic acids (TNA) eluted in 100 μl after each extraction.
 401 Samples with dPCR concentration below the operational limit of detection are shown with a blue “x”, and samples without
 402 measurement were labeled with a black “x”. Red points on the right y-axis represent virus read counts from unique reads. The
 403 dashed red line at 10 reads indicates the operational limit of detection used elsewhere in the analysis.

404 To compare the sensitivity of sequencing to that of digital PCR, endogenous SARS-CoV-2 and the spike-in
 405 BCoV were quantified in the final extracted nucleic acids produced by each concentration method. By
 406 sequencing, both SARS-CoV-2 and BCoV were detected in PMG and IP samples at the employed
 407 alignment stringency and read count threshold (> 10 reads, **Figure 5**), which corresponded with the
 408 higher relative abundances of human viruses in these two methods. However, the absolute
 409 concentrations of SARS-CoV-2 were significantly lower in IP samples than in PMG samples (IP v.s. PMG p
 410 = 0.009, **Figure 5a** and **Table S6**). Target virus concentrations could be increased by increasing the

411 effective volume of wastewater processed. Specifically, the final volume of the ultrafilter concentrate
412 nearly always exceeded the maximum input for nucleic acid extraction, resulting in a lower effective
413 volume (ranging from 16.3 ± 13 mL, **Table S1**). Similarly, only 0.25 g was extracted from 0.60 ± 0.18 g of
414 wet solids due to the limitation of the extraction kit, resulting in a lower effective sample volume
415 processed relative to the PMG and NT methods. The limited input may partially explain the low
416 concentrations of targets observed by dPCR. Notably, although samples from March 1 showed similar
417 SARS-CoV-2 concentrations from both NT and PMG methods (NT: 14.5 ± 3.4 gc/ μ L, PMG: 9.2 ± 1.4 gc/ μ L,
418 $p = 0.18$, **Table S6**), no SARS-CoV-2 reads were detected in the NT samples from this date. Meanwhile,
419 although BCoV was detected by dPCR in NT and Solids samples at low levels, it was absent in the
420 sequencing results. This suggests that in addition to the absolute viral concentration indicated by dPCR,
421 background non-target sequences may also influence target detection by sequencing.

422 **3.6 Implications for genome surveillance of known and novel human viruses**

423 Based on the comparisons reported here, wastewater virus concentration/extraction methods should be
424 chosen carefully and aligned with the specific monitoring endpoint and goal (e.g., sequencing or dPCR,
425 specific targets or broad range of targets). Removing wastewater solids, after treatment with either
426 Tween 20 (IP method) or protease (PMG method) and prior to concentration and extraction, improved
427 the overall detection of human viruses in probe-capture sequencing by minimizing the ratio between
428 off-target sequences and targeted human virus sequences⁵⁸ (**Figure 2**). However, solids removal may
429 also decrease the sensitivity of virus detection by dPCR by decreasing the absolute quantities of the
430 target in the sample (**Figure 5**)¹⁸. As untargeted sequencing was not performed in parallel, the extent to
431 which solids removal improved probe-capture enrichment specifically cannot be directly quantified.
432 Given that methods that included solids showed higher relative abundances of DNA viruses, a DNase
433 treatment may improve the recovery of human RNA viruses with these methods. Additionally, while
434 only two extraction methods were applied here (Qiagen AllPrep PowerViral and Promega Wizard Enviro
435 TNA), the extraction method used for solids and viral concentrates may affect the overall sensitivity of
436 sequencing by influencing the degree of viral lysis and integrity of the resulting nucleic acids¹⁶. Further
437 studies should be performed during periods of higher target concentration in wastewater (e.g., SARS-
438 CoV-2 surges) or using spike-in viruses to quantitatively determine limits of detection for sequencing
439 using different concentration and extraction methods.

440 Finally, the choice of probe panel likely also impacts the sensitivity of virus detection in probe-capture
441 sequencing. When using the RVOP probe set (which contains fewer virus targets than the VSP probe set)
442 several studies found remarkably high coverages of SARS-CoV-2, surpassing that of other human viruses
443 included in the RVOP panel^{14, 15, 49}. However, in the present study and other studies using broad virus
444 capture panels^{23, 48}, sequence data were dominated by enteric viruses such as mamastrovirus, with
445 limited detection of SARS-CoV-2. This points to the inherent challenge of using broad panels as a means
446 of wastewater-based surveillance for early detection of novel virus strains, which may appear at low
447 abundance before going on to cause a larger outbreak.

448 **4. Acknowledgements**

449 Funding was provided by the UCOP Lab Fees CRT Award (L22CR4507). We thank Khi Lai at EBMUD for
450 sample collection and Sanaiya Islam for laboratory management. Library preparation was performed
451 with advice from Justin Choi and Byran Bach at the Functional Genomics Laboratory and sequencing was

452 performed at the Vincent J. Coates Sequencing Laboratory (QB3, UC Berkeley, RRID: SCR_022170). We
453 thank Allie Nguyen and Van Trinh for their assistance with laboratory and bioinformatic analyses.

454 5. References

- 455 1. World Health, O., Guidelines for environmental surveillance of poliovirus circulation. In World
456 Health Organization: Geneva, 2003.
- 457 2. National Academies of Sciences, E.; Medicine; Health; Medicine, D.; Division on, E.; Life, S.;
458 Board on Population, H.; Public Health, P.; Water, S.; Technology, B.; Committee on Community
459 Wastewater-based Infectious Disease, S., In *Wastewater-based Disease Surveillance for Public Health
460 Action*, National Academies Press (US)
- 461 Copyright 2023 by the National Academy of Sciences. All rights reserved.: Washington (DC), 2023.
- 462 3. Boehm, A. B.; Hughes, B.; Duong, D.; Chan-Herur, V.; Buchman, A.; Wolfe, M. K.; White, B. J.,
463 Wastewater concentrations of human influenza, metapneumovirus, parainfluenza, respiratory syncytial
464 virus, rhinovirus, and seasonal coronavirus nucleic-acids during the COVID-19 pandemic: a surveillance
465 study. *Lancet Microbe* **2023**, *4*, (5), e340-e348.
- 466 4. Nemudryi, A.; Nemudraia, A.; Wiegand, T.; Surya, K.; Buyukyoruk, M.; Cicha, C.; Vanderwood, K.
467 K.; Wilkinson, R.; Wiedenheft, B., Temporal Detection and Phylogenetic Assessment of SARS-CoV-2 in
468 Municipal Wastewater. *Cell Rep Med* **2020**, *1*, (6), 100098.
- 469 5. Reyne, M. I.; Allen, D. M.; Levickas, A.; Allingham, P.; Lock, J.; Fitzgerald, A.; McSparron, C.;
470 Nejad, B. F.; McKinley, J.; Lee, A.; Bell, S. H.; Quick, J.; Houldcroft, C. J.; Bamford, C. G. G.; Gilpin, D. F.;
471 McGrath, J. W., Detection of human adenovirus F41 in wastewater and its relationship to clinical cases
472 of acute hepatitis of unknown aetiology. *Sci Total Environ* **2023**, *857*, (Pt 2), 159579.
- 473 6. Spurbeck, R. R.; Minard-Smith, A.; Catlin, L., Feasibility of neighborhood and building scale
474 wastewater-based genomic epidemiology for pathogen surveillance. *Sci Total Environ* **2021**, *789*,
475 147829.
- 476 7. Bisseux, M.; Debroas, D.; Mirand, A.; Archimbaud, C.; Peigue-Lafeuille, H.; Bailly, J. L.; Henquell,
477 C., Monitoring of enterovirus diversity in wastewater by ultra-deep sequencing: An effective
478 complementary tool for clinical enterovirus surveillance. *Water Res* **2020**, *169*, 115246.
- 479 8. Brinkman, N. E.; Fout, G. S.; Keely, S. P., Retrospective Surveillance of Wastewater To Examine
480 Seasonal Dynamics of Enterovirus Infections. *mSphere* **2017**, *2*, (3).
- 481 9. Fernandez-Cassi, X.; Timoneda, N.; Martinez-Puchol, S.; Rusinol, M.; Rodriguez-Manzano, J.;
482 Figuerola, N.; Bofill-Mas, S.; Abril, J. F.; Girones, R., Metagenomics for the study of viruses in urban
483 sewage as a tool for public health surveillance. *Sci Total Environ* **2018**, *618*, 870-880.
- 484 10. Cantalupo, P. G.; Calgua, B.; Zhao, G.; Hundesa, A.; Wier, A. D.; Katz, J. P.; Grabe, M.; Hendrix, R.
485 W.; Girones, R.; Wang, D.; Pipas, J. M., Raw Sewage Harbors Diverse Viral Populations. *mBio* **2011**, *2*, (5),
486 10.1128/mbio.00180-11.
- 487 11. Bibby, K.; Peccia, J., Identification of Viral Pathogen Diversity in Sewage Sludge by Metagenome
488 Analysis. *Environmental Science & Technology* **2013**, *47*, (4), 1945-1951.
- 489 12. Gaudin, M.; Desnues, C., Hybrid Capture-Based Next Generation Sequencing and Its Application
490 to Human Infectious Diseases. *Front Microbiol* **2018**, *9*, 2924.
- 491 13. Martinez-Puchol, S.; Rusinol, M.; Fernandez-Cassi, X.; Timoneda, N.; Itarte, M.; Andres, C.;
492 Anton, A.; Abril, J. F.; Girones, R.; Bofill-Mas, S., Characterisation of the sewage virome: comparison of
493 NGS tools and occurrence of significant pathogens. *Sci Total Environ* **2020**, *713*, 136604.
- 494 14. Crits-Christoph A, K. R., Olm MR, Whitney ON, Al-Shayeb B, Lou YC, Flamholz A, Kennedy LC,
495 Greenwald H, Hinkle A, Hetzel J, Spitzer S, Koble J, Tan A, Hyde F, Schroth G, Kuersten S, Banfield JF,

- 496 Nelson KL., Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *mBio*
497 **2021**, *12*, (1).
- 498 15. Rothman, J. A.; Loveless, T. B.; Kapcia, J., 3rd; Adams, E. D.; Steele, J. A.; Zimmer-Faust, A. G.;
499 Langlois, K.; Wanless, D.; Griffith, M.; Mao, L.; Chokry, J.; Griffith, J. F.; Whiteson, K. L., RNA Viromics of
500 Southern California Wastewater and Detection of SARS-CoV-2 Single-Nucleotide Variants. *Appl Environ*
501 *Microbiol* **2021**, *87*, (23), e0144821.
- 502 16. Hjelmsø, M. H.; Hellmér, M.; Fernandez-Cassi, X.; Timoneda, N.; Lukjancenko, O.; Seidel, M.;
503 Elsässer, D.; Aarestrup, F. M.; Löfström, C.; Bofill-Mas, S.; Abril, J. F.; Girones, R.; Schultz, A. C., Evaluation
504 of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic
505 Sequencing. *PLOS ONE* **2017**, *12*, (1), e0170199.
- 506 17. Ahmed, W.; Bivins, A.; Metcalfe, S.; Smith, W. J. M.; Verbyla, M. E.; Symonds, E. M.; Simpson, S.
507 L., Evaluation of process limit of detection and quantification variation of SARS-CoV-2 RT-qPCR and RT-
508 dPCR assays for wastewater surveillance. *Water Res* **2022**, *213*, 118132.
- 509 18. North, D.; Bibby, K., Comparison of viral concentration techniques for native fecal indicators and
510 pathogens from wastewater. *Sci Total Environ* **2023**, *905*, 167190.
- 511 19. Giron-Guzman, I.; Diaz-Reolid, A.; Cuevas-Ferrando, E.; Falco, I.; Cano-Jimenez, P.; Comas, I.;
512 Perez-Cataluna, A.; Sanchez, G., Evaluation of two different concentration methods for surveillance of
513 human viruses in sewage and their effects on SARS-CoV-2 sequencing. *Sci Total Environ* **2023**, *862*,
514 160914.
- 515 20. Izquierdo-Lara, R.; Elsinga, G.; Heijnen, L.; Munnink, B. B. O.; Schapendonk, C. M. E.;
516 Nieuwenhuijse, D.; Kon, M.; Lu, L.; Aarestrup, F. M.; Lycett, S.; Medema, G.; Koopmans, M. P. G.; de
517 Graaf, M., Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater
518 Sequencing, the Netherlands and Belgium. *Emerg Infect Dis* **2021**, *27*, (5), 1405-1415.
- 519 21. Jahn, K.; Dreifuss, D.; Topolsky, I.; Kull, A.; Ganesanandamoorthy, P.; Fernandez-Cassi, X.;
520 Bänziger, C.; Devaux, A. J.; Stachler, E.; Caduff, L.; Cariti, F.; Corzón, A. T.; Fuhrmann, L.; Chen, C.;
521 Jablonski, K. P.; Nadeau, S.; Feldkamp, M.; Beisel, C.; Aquino, C.; Stadler, T.; Ort, C.; Kohn, T.; Julian, T. R.;
522 Beerenwinkel, N., Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using
523 COJAC. *Nature Microbiology* **2022**, *7*, (8), 1151-1160.
- 524 22. Jahn, K.; Dreifuss, D.; Topolsky, I.; Kull, A.; Ganesanandamoorthy, P.; Fernandez-Cassi, X.;
525 Bänziger, C.; Devaux, A. J.; Stachler, E.; Caduff, L.; Cariti, F.; Corzón, A. T.; Fuhrmann, L.; Chen, C.;
526 Jablonski, K. P.; Nadeau, S.; Feldkamp, M.; Beisel, C.; Aquino, C.; Stadler, T.; Ort, C.; Kohn, T.; Julian, T. R.;
527 Beerenwinkel, N., Detection and surveillance of SARS-CoV-2 genomic variants in wastewater. In Cold
528 Spring Harbor Laboratory: 2021.
- 529 23. McCall, C.; Leo Elworth, R. A.; Wylie, K. M.; Wylie, T. N.; Dyson, K.; Doughty, R.; Treangen, T. J.;
530 Hopkins, L.; Ensor, K.; Stadler, L. B., Targeted Metagenomic Sequencing for Detection of Vertebrate
531 Viruses in Wastewater for Public Health Surveillance. *ACS ES&T Water* **2023**, *3*, (9), 2955-2965.
- 532 24. Spurbeck, R. R.; Catlin, L. A.; Mukherjee, C.; Smith, A. K.; Minard-Smith, A., Analysis of
533 metatranscriptomic methods to enable wastewater-based biosurveillance of all infectious diseases.
534 *Front Public Health* **2023**, *11*, 1145275.
- 535 25. Wyler, E.; Lauber, C.; Manukyan, A.; Deter, A.; Quedenau, C.; Teixeira Alves, L. G.; Seitz, S.;
536 Altmüller, J.; Landthaler, M., Comprehensive profiling of wastewater viromes by genomic sequencing.
537 **2022**.
- 538 26. d, M. G.; Huggett, J. F., The Digital MIQE Guidelines Update: Minimum Information for
539 Publication of Quantitative Digital PCR Experiments for 2020. *Clin Chem* **2020**, *66*, (8), 1012-1029.
- 540 27. B, B., BBTools software packag. In e, 2014.
- 541 28. Shen, W.; Le, S.; Li, Y.; Hu, F., SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File
542 Manipulation. *PLOS ONE* **2016**, *11*, (10), e0163962.

- 543 29. Langmead, B.; Salzberg, S. L., Fast gapped-read alignment with Bowtie 2. *Nature Methods* **2012**,
544 9, (4), 357-359.
- 545 30. Kim, D.; Song, L.; Breitwieser, F. P.; Salzberg, S. L., Centrifuge: rapid and sensitive classification of
546 metagenomic sequences. *Genome Res* **2016**, 26, (12), 1721-1729.
- 547 31. Martí, J. M., Recentrifuge: Robust comparative analysis and contamination removal for
548 metagenomics. *PLOS Computational Biology* **2019**, 15, (4), e1006967.
- 549 32. Ondov, B. D.; Treangen, T. J.; Melsted, P.; Mallonee, A. B.; Bergman, N. H.; Koren, S.; Phillippy, A.
550 M., Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **2016**, 17,
551 (1), 132.
- 552 33. Oksanen J, S. G., Blanchet F, Kindt R, Legendre P.; Minchin P, O. H. R., Solymos P, Stevens M,
553 Szoecs E.; Wagner H, B. M., Bedward M, Bolker B, Borcard D.; Carvalho G, C. M., De Caceres M, Durand
554 S.; Evangelista H, F. R., Friendly M, Furneaux B.; Hannigan G, H. M., Lahti L, McGlinn D, Ouellette M.;
555 Ribeiro Cunha E, S. T., Stier A, Ter Braak C, Weedon; Jablonski, K. P., `_vegan`: Community Ecology
556 Package. R
package version 2.6-4. **2022**.
- 557 34. Khare, S.; Gurry, C.; Freitas, L.; Schultz, M. B.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; Lee, R. T.; Yeo,
558 W.; Curation Team, G. C.; Maurer-Stroh, S., GISAD's Role in Pandemic Response. *China CDC Wkly* **2021**,
559 3, (49), 1049-1051.
- 561 35. Schoch, C. L.; Ciufo, S.; Domrachev, M.; Hotton, C. L.; Kannan, S.; Khovanskaya, R.; Leipe, D.;
562 McVeigh, R.; O'Neill, K.; Robbertse, B.; Sharma, S.; Soussov, V.; Sullivan, J. P.; Sun, L.; Turner, S.; Karsch-
563 Mizrachi, I., NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*
564 (Oxford) **2020**, 2020.
- 565 36. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A. A.; Dvorkin, M.; Kulikov, A. S.; Lesin, V. M.;
566 Nikolenko, S. I.; Pham, S.; Prjibelski, A. D.; Pyshkin, A. V.; Sirotkin, A. V.; Vyahhi, N.; Tesler, G.; Alekseyev,
567 M. A.; Pevzner, P. A., SPAdes: a new genome assembly algorithm and its applications to single-cell
568 sequencing. *J Comput Biol* **2012**, 19, (5), 455-77.
- 569 37. Guo, J.; Bolduc, B.; Zayed, A. A.; Varsani, A.; Dominguez-Huerta, G.; Delmont, T. O.; Pratama, A.
570 A.; Gazitúa, M. C.; Vik, D.; Sullivan, M. B.; Roux, S., VirSorter2: a multi-classifier, expert-guided approach
571 to detect diverse DNA and RNA viruses. *Microbiome* **2021**, 9, (1), 37.
- 572 38. Sayers, E. W.; Bolton, E. E.; Brister, J. R.; Canese, K.; Chan, J.; Comeau, D. C.; Connor, R.; Funk, K.;
573 Kelly, C.; Kim, S.; Madej, T.; Marchler-Bauer, A.; Lanczycki, C.; Lathrop, S.; Lu, Z.; Thibaud-Nissen, F.;
574 Murphy, T.; Phan, L.; Skripchenko, Y.; Tse, T.; Wang, J.; Williams, R.; Trawick, B. W.; Pruitt, K. D.; Sherry,
575 S. T., Database resources of the national center for biotechnology information. *Nucleic Acids Res* **2022**,
576 50, (D1), D20-d26.
- 577 39. Robinson, J. T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E. S.; Getz, G.; Mesirov,
578 J. P., Integrative genomics viewer. *Nat Biotechnol* **2011**, 29, (1), 24-6.
- 579 40. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper,
580 A.; Markowitz, S.; Duran, C.; Thierer, T.; Ashton, B.; Meintjes, P.; Drummond, A., Geneious Basic: an
581 integrated and extendable desktop software platform for the organization and analysis of sequence
582 data. *Bioinformatics* **2012**, 28, (12), 1647-9.
- 583 41. Edgar, R. C., MUSCLE: multiple sequence alignment with high accuracy and high throughput.
584 *Nucleic Acids Res* **2004**, 32, (5), 1792-7.
- 585 42. Castresana, J., Selection of Conserved Blocks from Multiple Alignments for Their Use in
586 Phylogenetic Analysis. *Molecular Biology and Evolution* **2000**, 17, (4), 540-552.
- 587 43. Tamura, K.; Stecher, G.; Kumar, S., MEGA11: Molecular Evolutionary Genetics Analysis Version
588 11. *Mol Biol Evol* **2021**, 38, (7), 3022-3027.

- 589 44. Xu, W.; Xu, N.; Zhang, M.; Wang, Y.; Ling, G.; Yuan, Y.; Zhang, P., Nanotraps based on
590 multifunctional materials for trapping and enrichment. *Acta Biomater* **2022**, *138*, 57-72.
- 591 45. Richter, L.; Ksiezarczyk, K.; Paszkowska, K.; Janczuk-Richter, M.; Niedziolka-Jonsson, J.; Gapinski,
592 J.; Los, M.; Holyst, R.; Paczesny, J., Adsorption of bacteriophages on polypropylene labware affects the
593 reproducibility of phage research. *Sci Rep* **2021**, *11*, (1), 7387.
- 594 46. Mondal, S.; Feirer, N.; Brockman, M.; Preston, M. A.; Teter, S. J.; Ma, D.; Goueli, S. A.; Moorji, S.;
595 Saul, B.; Cali, J. J., A direct capture method for purification and detection of viral nucleic acid enables
596 epidemiological surveillance of SARS-CoV-2. *Sci Total Environ* **2021**, *795*, 148834.
- 597 47. Hatfull, G. F.; Hendrix, R. W., Bacteriophages and their genomes. *Curr Opin Virol* **2011**, *1*, (4),
598 298-303.
- 599 48. Martinez-Puchol, S.; Itarte, M.; Rusinol, M.; Fores, E.; Mejias-Molina, C.; Andres, C.; Anton, A.;
600 Quer, J.; Abril, J. F.; Girones, R.; Bofill-Mas, S., Exploring the diversity of coronavirus in sewage during
601 COVID-19 pandemic: Don't miss the forest for the trees. *Sci Total Environ* **2021**, *800*, 149562.
- 602 49. Khan, M.; Li, L.; Haak, L.; Payen, S. H.; Carine, M.; Adhikari, K.; Uppal, T.; Hartley, P. D.; Vasquez-
603 Gross, H.; Petereit, J.; Verma, S. C.; Pagilla, K., Significance of wastewater surveillance in detecting the
604 prevalence of SARS-CoV-2 variants and other respiratory viruses in the community - A multi-site
605 evaluation. *One Health* **2023**, *16*, 100536.
- 606 50. Paul, D.; Kolar, P.; Hall, S. G., A review of the impact of environmental factors on the fate and
607 transport of coronaviruses in aqueous environments. *npj Clean Water* **2021**, *4*, (1).
- 608 51. Pinon, A.; Vialette, M., Survival of Viruses in Water. *Intervirology* **2018**, *61*, (5), 214-222.
- 609 52. Mejías-Molina, C.; Pico-Tomás, A.; Beltran-Rubinat, A.; Martínez-Puchol, S.; Corominas, L.;
610 Rusiñol, M.; Bofill-Mas, S., Effectiveness of passive sampling for the detection and genetic
611 characterization of human viruses in wastewater. *Environmental Science: Water Research & Technology*
612 **2023**, *9*, (4), 1195-1204.
- 613 53. Strubbia, S.; Schaeffer, J.; Oude Munnink, B. B.; Besnard, A.; Phan, M. V. T.; Nieuwenhuijse, D. F.;
614 de Graaf, M.; Schapendonk, C. M. E.; Wacrenier, C.; Cotten, M.; Koopmans, M. P. G.; Le Guyader, F. S.,
615 Metavirome Sequencing to Evaluate Norovirus Diversity in Sewage and Related Bioaccumulated Oysters.
616 *Front Microbiol* **2019**, *10*, 2394.
- 617 54. Levican, J.; Levican, A.; Ampuero, M.; Gaggero, A., JC polyomavirus circulation in one-year
618 surveillance in wastewater in Santiago, Chile. *Infect Genet Evol* **2019**, *71*, 151-158.
- 619 55. Rafique, A.; Jiang, S. C., Genetic diversity of human polyomavirus JCPyV in Southern California
620 wastewater. *J Water Health* **2008**, *6*, (4), 533-8.
- 621 56. Oghuan, J.; Chavarria, C.; Vanderwal, S. R.; Gitter, A.; Ojaruega, A. A.; Monserrat, C.; Bauer, C. X.;
622 Brown, E. L.; Cregeen, S. J.; Deegan, J.; Hanson, B. M.; Tisza, M.; Ocaranza, H. I.; Balliew, J.; Maresso, A.
623 W.; Rios, J.; Boerwinkle, E.; Mena, K. D.; Wu, F., Wastewater surveillance suggests unreported Mpox
624 cases in a low-prevalence area. *MedRxiv* **2023**.
- 625 57. Wolfe, M. K.; Duong, D.; Hughes, B.; Chan-Herur, V.; White, B. J.; Boehm, A. B., Detection of
626 monkeypox viral DNA in a routine wastewater monitoring program. *MedRxiv* **2022**.
- 627 58. Rehn, A.; Braun, P.; Knüpfer, M.; Wölfel, R.; Antwerpen, M. H.; Walter, M. C., Catching SARS-
628 CoV-2 by sequence hybridization: a comparative analysis. In Cold Spring Harbor Laboratory: 2021.