Untargeted longitudinal ultra deep metagenomic sequencing of wastewater provides a comprehensive readout of expected and unexpected viral pathogens

Clayton Rushford[a], Devon Gregory[b], Emma Copen[b], Jonathan Naydenov[b], Alejandro Tovar-Mendez[b], Po-E Li[c], Rose S. Kantor[d], Migun Shakya[c], Nelson Ruth[c], Patrick S. G. Chain[c], Hsinyeh Hsieh[e], Torin Hunter[b], Josh Kome[b], Matt Frank[b], Terri Lyddon[b], Jeff Kaufman[f], David H. O'Connor[g], Marc C. Johnson[b#]

Author Affiliations

[a]University of Missouri, Department of Veterinary Pathobiology

[b]University of Missouri, Department of Molecular Microbiology and Immunology

[c]B-GEN, Bioscience Division, Los Alamos National Laboratory

[d]Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory

[e]University of Missouri, School of Natural Resources

[f]Nucleic Acid Observatory, SecureBio

[g]University of Wisconsin-Madison, Department of Pathology and Laboratory Medicine

[#]Corresponding Author: marcjohnson@health.missouri.edu

Abstract

Wastewater surveillance has become a powerful tool to monitor circulating viruses at a community level. Currently, most wastewater surveillance efforts use target-based approaches such as quantitative PCR techniques or hybrid capture. This study explores the feasibility of using unbiased ultra-deep metagenomic sequencing as a comprehensive approach to wastewater surveillance. To test this, composite influent wastewater samples were collected weekly from January 2024 through June 2025. Unbiased sequencing was performed on all samples with an average depth of 1.1 billion reads per sample. Human enteric viruses such as rotaviruses, astroviruses, and noroviruses were detected at high levels in virtually every sample. SARS-CoV-2 was also detected in most samples and the counts per sample positively correlated with digital PCR (dPCR) measurements. Less abundant respiratory pathogens such as influenza A and B, rhinoviruses, parainfluenzaviruses, and human coronaviruses 229E, OC43, NL63, and HKU1 were also regularly detected. However, those pathogens displayed distinct and reproducible winter and spring seasonality. Several unexpected viruses were also detected, such as several detections of highly pathogenic avian influenza H5N1 (HPAI H5N1) genotype B3.13, a month-long surge of hepatitis A virus, and a large season-specific surge in influenza C virus. The most abundant known virus detected was the *Tobamovirus* tomato brown rugose fruit virus, which was present year-round at high abundance. However, other tobamoviruses such as tomato mosaic virus were detected primarily in the late growing season. This eighteen-month study highlights that unbiased ultra deep sequencing enables detection of expected and unexpected viral pathogens without targeted enrichment.

Importance

This study demonstrates that untargeted ultra deep metagenomic sequencing can provide a comprehensive and scalable tool for wastewater surveillance of viral pathogens. By generating approximately 1 billion reads per sample across 78 consecutive weeks, we captured expected pathogens such as SARS-CoV-2, noroviruses, and influenza viruses.

Additionally, we captured unexpected viral signals such as influenza C and highly pathogenic avian influenza H5N1. The wide range of viral taxa captured in this study also displays epidemiologically relevant seasonality. We also observed a correlation between metagenomic SARS-CoV-2 read counts and dPCR values to validate this method against other wastewater surveillance methods currently in use. Our findings highlight how ultra deep metagenomics can enhance pandemic preparedness, enable early detection of novel or re-emerging viruses, and broaden the scope of One Health monitoring by capturing human, animal, and plant viral signatures from a composite wastewater sample.

Keywords

Wastewater surveillance, metagenomic, deep sequencing, SARS-CoV-2, influenza H5N1, influenza C, tomato brown rugose fruit virus, *NVD*, *GOTTCHA2*, OneHealth, public health

Introduction

Wastewater surveillance is a powerful tool for public health that leverages minimally biased sampling on a population level to facilitate pandemic preparedness by providing early warnings for circulating pathogens in sampled communities based on sequence data [1, 2]. This form of community-level sampling reduces biases inherent to clinical surveillance, such as underreporting due to asymptomatic infections or limited access to healthcare. Several labs have demonstrated this concept through reverse transcription PCR (RT-PCR), real time quantitative PCR (RT-qPCR), and digital PCR (dPCR) designed to target human-infecting pathogens such as SARS-CoV-2, influenza, and *Enterovirus* spp. [3-5]. Rapid multiplex PCR panels are also available for clinically relevant viruses and have been used for tracking multiple viruses in wastewater [6]. However, PCR-based approaches require frequent modifications due to evolution around primer binding sites and are limited to known targets [7]. Alternatively, probe-capture approaches facilitate detection of a wider range of pathogens, have been utilized in wastewater systems, and have potential to detect pathogens that differ by up to 40% from reference genomes [8, 9]. Additionally, commercially available probe panels have been developed

72 and implemented for wastewater surveillance [10-14]. While PCR and probe-capture
73 approaches facilitate detection of viruses with *a priori* knowledge of genetic makeup, they
74 are unable to detect highly divergent, novel viral sequences that may be circulating in a
75 population.

76 Unbiased sequencing and metagenomics can detect a wider range of known and
77 unknown pathogens compared to targeted and hybrid capture methods if they receive
78 sufficient coverage. Additionally, pathogen concentration methods, such as polyethylene
79 glycol (PEG) precipitation and vacuum-based direct capture, have been shown to be the
80 most reproducible while improving depth and breadth of viral signals [15]. However, a deep
81 dataset is a prerequisite for unbiased metagenomic surveillance due to a low target-to-
82 nontarget ratio of sequences [16]. Fortunately, the rapidly decreasing cost of sequencing
83 has facilitated data generation at scales that would, until recently, have been infeasible
84 [17].

85 The goal of this study was to test the feasibility of viral surveillance using unbiased
86 metagenomic sequencing. A variety of previous studies have utilized metagenomic
87 approaches for screening antimicrobial resistance genes, microbiomes, and viromes in
88 wastewater [2, 18-26]. Most of these take place over a short period, yet a few longitudinal
89 studies have surveyed wastewater for longer than a year to determine how detectable
90 sequences change over time [11, 18, 27]. While previous studies have demonstrated the
91 efficacy of metagenomics for viral surveillance, our study is unique due to our combination
92 of longitudinal sampling (78 consecutive weeks), use of PEG precipitation for viral
93 concentration, and sequencing at a depth greater than previous work. To provide a high-
94 confidence multi-kingdom breakdown of sample composition, we assigned sequence
95 reads through *GOTTCHA2* (Genomic Origin Through Taxonomic CHAllenge) classification
96 using an expanded reference database. For the detection of human-infecting viral
97 sequences, we implement *NVD* (New Virus Discovery), a workflow designed to assemble
98 putative viral sequences from human-infecting virus families and validate assemblies.

99 **Methods**

100 *Sample collection*

101 Influent wastewater from Columbia, MO was collected weekly as a 24-hour
102 composite sample using an ISCO autosampler. A 35 mL aliquot was transferred to sterile
103 50 mL conical and stored at 4°C prior to processing. At the time of collection, water quality
104 parameters including temperature, pH, total suspended solids, average influent flow over
105 sampling period, and chemical oxygen demand were recorded and provided by the
106 Missouri's Department of Health and Senior Services.

107 *Viral concentration and RNA extraction for Metagenomics*

108 Samples were centrifuged for 2000 xg at 4°C for 5 minutes and the supernatant was
109 filtered through 0.22 µm filters (Millipore, SCGP00525) to remove solids and larger
110 microbes. A subset of samples collected between January 2nd and March 19th, 2024 were
111 then treated with 10,000 ng of RNAse A (ThermoFisher, R1253), 2 units of TURBO DNAse
112 (Invitrogen, AM1907), and a 4.75 mL of 5 mM $CaCl_2$ and 25 mM $MgCl_2$ buffer, in an attempt
113 to remove bacterial ribosomal RNA. Separately, another sample subset was treated with
114 nucleases (as above) and 30 mM EDTA to dissociate rRNA from ribosomes. For nuclease
115 treated samples, 400 units of RNaseOUT™ Recombinant Ribonuclease Inhibitor
116 (Invitrogen, 10777019) was added prior to extraction. A breakdown of treatments per
117 sample date is summarized in Supplementary Table 1.

118 Untreated, nuclease treated, and nuclease and EDTA treated sample filtrates were
119 mixed with 12.5 mL of 50% (w/vol) PEG (Research Products International, P48080-1000.0)
120 and 1.2M NaCl for 1 h at 4°C. Nucleic acids were extracted from pellets using QIAamp Viral
121 RNA Mini Kit (Qiagen, 52906) using the manufacturer's instructions excluding the addition
122 of carrier RNA. Samples were not DNase treated after extraction.

123 *RT-qPCR of pepper mild mottle virus*

124 Viral RNA in samples was assessed prior to library preparation through comparative
125 RT-qPCR targeting pepper mild mottle virus (PMMoV) to determine if viruses were

126 successfully concentrated. Pepper mild mottle virus was selected as an endogenous
127 control due to its ubiquitous distribution in wastewater [28]. Forward (5'
128 GGCGTAGATCCATTGGTGC 3') and reverse (5' CGAACCTTCCTCCTTTGATG 3') primers
129 were designed for RT-qPCR to target the PMMoV genome were mixed and resuspended at
130 100 μM concentrations. A Taqman probe (VIC-5' GCTGTGGTTTCAAATGAGAGTGG 3'-QSY)
131 was prepared at 100 μM concentration. The RT-qPCR reaction was carried out using
132 TaqPath 1-Step RT-qPCR Master Mix, CG (4x) (Applied Biosystems, A15299), forward and
133 reverse primers (500 nM each), taqman probe (125 nM), and 5 μL of RNA template for a
134 total reaction volume of 20 μL. Reactions were carried out on an Applied Biosystems 7500
135 Fast System, with thermocycler settings of 48°C for 15 min, 95°C for 10 min, followed by 40
136 cycles of 95°C for 15 s and 60°C for 1 min.

137 *Library preparation and deep sequencing*

138 Total RNA concentrations were quantified through Qubit RNA High Sensitivity
139 Assay, although some samples were below the detection limit for this assay. Libraries were
140 prepped from RNA for deep sequencing using manufacturers recommendations with slight
141 modifications. RNA inputs for library preparation were normalized based on Qubit results
142 to 10-15 ng in 9.5 μL. If input concentrations were <0.2 ng/μl then 9.5 μL was used for
143 input. The mRNA capture step was removed with library preps beginning with RNA
144 fragmentation and cDNA synthesis. Normalized RNA inputs were evenly split into two
145 aliquots. One aliquot was fragmented so longer RNAs were broken into a target range
146 compatible with sequencing, while the other half was left unfragmented to retain
147 sequences already within the target range. After fragmentation, aliquots were pooled
148 together for cDNA synthesis. Lastly, final indexed libraries underwent size selection during
149 the final cleanup by mixing 30 μL of AMPure XP (Beckman Coulter, A63881) with 50 μL of
150 post-indexing PCR amplicons to bind to long library sequences. Magnetic beads were
151 incubated for 5 minutes, and the supernatant was moved to a new tube where 20 μL of
152 AMPure XP beads were added to bind to shorter library sequences. These shorter
153 fragments were washed, eluted in 30 μL of resuspension buffer, and used for sample
154 pooling. Final libraries were sequenced using a NovaSeq 6000 until April 16th 2024, and

155 subsequent libraries were sequenced using a NovaSeq X with PE150 reads
156 (Supplementary Table 1).

157 *Assessing sequencing depth and rRNA abundance in samples*

158 Analysis by qPCR of PMMoV showed no noticeable deviation in Ct values for sample
159 dates that underwent multiple treatments. However upon analysis of rRNA abundance
160 (described below), treatments did not show any reliable reduction of rRNA in samples.
161 Therefore, treatment use was discontinued, and fastq files from samples with multiple
162 treatments per date were combined for analysis. Fastq files were trimmed for adapters,
163 homopolymers, low entropy sequences, and short length sequences using fastp (version
164 0.23.4) using the following options, "-w 16 --dont_eval_duplication -l 50 -r -c --trim_poly_g -
165 -poly_g_min_len=5 --trim_poly_x --poly_x_min_len=5 -y --
166 adapter_sequence=CTGTCTCTTATACACATCT --
167 adapter_sequence_r2=CTGTCTCTTATACACATCT". Samples were split into rRNA and non-
168 rRNA fractions to assess the approximate depth of potential viral sequences. Ribosomal
169 RNA separation was performed using bbduk.sh from BBmap (version 38.18) and a custom
170 rRNA database curated by Brian Bushnell [29]. The depth of rRNA fractions and non-rRNA
171 fractions were determined using stats.sh from BBmap. Both BBTools commands were
172 performed using default settings.

173 *Taxonomic assignment with GOTTCHA2.*

174 Overall sample composition was determined with high confidence through a
175 species-level *GOTTCHA2* (version 2.1.9.0) analysis using fastp trimmed sequences and
176 arguments "-nc -rm no" [30]. *GOTTCHA2* was selected for reporting for definitive sample
177 composition due to its curated multi-kingdom reference database which uses unique
178 genomic signatures at the queried taxonomic level. Taxonomic assignments were grouped
179 by species, read counts were summed for each species, and read counts were normalized
180 to reads per million (RPM). Host domains for each detected viral phylum were manually
181 annotated (Supplemental Table 2).

182 *Taxonomic assignment with NVD*.

183 We developed a Snakemake-based pipeline to detect and classify viral sequences
184 from high-throughput sequencing data. The pipeline was containerized using Apptainer
185 (formerly Singularity) to ensure reproducibility across computing environments. Raw
186 sequencing reads were first interleaved using BBMap (v38.18). To reduce computational
187 requirements and improve specificity, reads were pre-filtered using NCBI's Sequence
188 Taxonomy Analysis Tool (STAT; aligns_to v3.1.1) to retain only those with k-mer matches to
189 known human-infecting virus families, including *Adenoviridae*, *Anelloviridae*, *Arenaviridae*,
190 *Astroviridae*, *Bornaviridae*, *Peribunyaviridae*, *Caliciviridae*, *Coronaviridae*, *Filoviridae*,
191 *Flaviviridae*, *Hepadnaviridae*, *Hepeviridae*, *Orthoherpesviridae*, *Orthomyxoviridae*,
192 *Papillomaviridae*, *Paramyxoviridae*, *Parvoviridae*, *Picobirnaviridae*, *Picornaviridae*,
193 *Pneumoviridae*, *Polyomaviridae*, *Poxviridae*, *Sedoreoviridae*, *Retroviridae*, *Rhabdoviridae*,
194 *Togaviridae*, and *Kolmioviridae*. *Arteriviridae* were also included due to their ability to infect
195 other primate species.

196 The filtered reads were assembled using SPAdes (v4.0.0) in sewage mode. The
197 resulting contigs underwent quality control using BBMask to remove low-complexity
198 regions using an entropy threshold of 0.9. Following masking, contigs were filtered to retain
199 only sequences with at least 200 consecutive non-N bases after trimming terminal N's, as
200 shorter sequences can interfere with subsequent classification steps. The quality-
201 controlled contigs then underwent a hierarchical classification process beginning with
202 analysis using a NCBI STAT's tree index database provided by Kenneth Katz (downloaded
203 from NCBI on 2024-08-30; tree_filter.dbss, tree_filter.dbss.annotation, tree_index.dbs).
204 From this first-pass classification, a focused taxa list was generated and used for a
205 second, more detailed classification using STAT's dbss database. This two-stage STAT
206 analysis retained matches with greater than 0.001% confidence and formed the basis for
207 subsequent BLAST analysis.

208 The BLAST-based classification employed a two-stage approach for maximum
209 sensitivity. Initially, contigs were analyzed using MEGABLAST against NCBI's nucleotide

database with default parameters to identify close matches to known sequences. Contigs that remained unclassified after MEGABLAST underwent a more sensitive BLASTN search using default parameters. The results from both BLAST stages were filtered to remove non-viral matches and bacteriophages, then merged to create a comprehensive classification of the viral contigs. To validate the assembled viral sequences and assess their coverage, the initially filtered reads were mapped back to the identified viral contigs using minimap2 (v2.1). Coverage calculations were performed using samtools (v1.20), retaining only primary alignments for accurate depth estimation. A SQL script was generated to query BLAST validated assemblies for human specific pathogen names and group them based on tropism for use in heatmap generation. Similarly to *GOTTCHA2* read counts, *NVD* read counts were normalized to RPM for each species detected.

Digital PCR targeting SARS-CoV-2

Raw samples were prepped in accordance with Missouri's SARS-CoV-2 wastewater surveillance program. Twenty-four-hour composite samples as described above were acquired at Missouri wastewater treatment facilities (WWTF) and stored at 4°C until processing. After mixing the samples briefly by gently inverting the tubes, a total of 9.75 mL of wastewater per sample was loaded into 24-deep-well plates. A total of 20 µL of the previously described NL4-3 derived HIV with the CMV-driven puromycin-resistance gene and lacking the *env*, *nef*, *vif*, and *vpr* genes were added to each sample as a processing control [31]. Then, a total of 100 µL of Nanotrap Enhancement Reagent 2 (Ceres Nanosciences, 10112-10) and 150 µL of Nanotrap Microbiome A Particles (Ceres Nanosciences, 44202) were added to each sample to capture, concentrate, and enhance downstream detection of viruses. A total of 5.6 µL of carrier RNA (Qiagen, 52906) was mixed with 700 µL of Buffer AVL (Qiagen, 52906) on a separate 24-deep-well plate to lyse viruses and help with RNA recovery from the wastewater samples. All plates were loaded into the KingFisher Apex (ThermoFisher Scientific) and processed. After processing in the KingFisher Apex, 700 µL of the lysate was loaded into the QIAcube Connect (Qiagen), and

237 RNA was extracted and eluted into a final volume of 60 µL using the QIAamp Viral RNA Mini
238 Kit (Qiagen, 52906), according to the manufacturer's instructions.

239 Forward (5' GACCCCAAAATCAGCGAAAT 3' and 5' TTACAAACATTGGCCGCAAA 3')
240 and reverse (5' TCTGGTTACTGCCAGTTGAATCTG 3' and 5' GCGCGACATTCCGAAGAA 3')
241 primers and TaqMan probes (FAM-5' ACCCCGCATTACGTTTGGTGGACC 3'-MGB-NFQ and
242 ABY-5' ACAATTTGCCCCCAGCGCTTCAG 3'-QSY) were designed for RT-dPCR to target the
243 N1 and N2 regions of N gene of SARS-CoV-2, respectively. The primers were resuspended
244 at 100 µM concentrations as TaqMan probe. The RT-dPCR reaction was carried out using
245 QIAcuity One Step Advanced Probe kits (Qiagen, 250131), forward and reverse primers
246 (500nM each), TaqMan probe (125 nM) and 10 µL of RNA template for a total reaction
247 volume of 40 µL in a QIAcuity Four digital PCR system, with thermocycler settings of 50°C
248 for 30 min, 95°C for 2 min, followed by 45 cycles of 95°C for 10 s and 55°C for 30 s, after
249 priming.

250 *Statistical analysis*.

251 Relative abundance of rRNA was assessed for correlation against collection
252 temperature, water pH, total suspended solids, average influent flow over sampling
253 period, and chemical oxygen demand using Spearman's Correlation Coefficient.

254 For *GOTTCHA2* assigned sequences, Shannon–Wiener diversity (H') at the species
255 level was calculated separately for prokaryotic-host and eukaryotic-host viruses in each
256 sample using the diversity function in the Vegan package (version 2.6.8) of R (version
257 4.4.2). Paired differences in H' between prokaryote-infecting and eukaryote-infecting
258 viruses were tested for normality by Shapiro–Wilk test, then compared using a two-tailed
259 paired Student's t-test. A Wilcoxon signed-rank test was also performed as a
260 nonparametric alternative. A 95 % confidence interval for the mean H' difference was
261 estimated by nonparametric bootstrap (10 000 replicates) via the boot package (v1.3-31).
262 Statistical significance was defined as $p < 0.05$.

Metagenomic read counts for SARS-CoV-2 were compared with dPCR reports through Missouri's SARS-CoV-2 wastewater surveillance program to determine if the metagenomic approach showed consistent epidemiological trends and was consistent with our current SARS-CoV-2 surveillance protocol. For each sampling week, dPCR results were flow normalized by multiplying mean copies per liter by average influent flow during the sampling period to calculate total SARS-CoV-2 viral load. The metagenomic SARS-CoV-2 (NC_045512.2) and ToBRFV (KT383474.1) sequences were extracted from trimmed fastq files using minimap2 (version 2.14-r883) using the argument "-ax sr". Secondary alignments and matches with MAPQ scores <30 were removed using samtools (version 1.20). The association between flow-normalized dPCR or with SARS-CoV-2 counts normalized to RPM or with ToBRFV were quantified in R using Pearson's correlation (two-sided). For visualization only, the metagenomic counts were linearly rescaled to overlay the dPCR results on a secondary y-axis.

For *NVD* assigned sequences and *GOTTCHA2* sequences assigned to the same viral families, beta diversity was longitudinally assessed on the species level through Bray-Curtis dissimilarity of both outputs using the vegdist function from the vegan package in R. Permutational analysis of variance (PERMANOVA) was performed to evaluate the effects of season on the full data set. Since samples were collected longitudinally from a single site, results were interpreted as indicative of overall compositional differences rather than as strictly independent statistical tests, recognizing the potential for temporal autocorrelation among samples. Seasons were assigned based on solstice and equinox dates. PERMANOVA was also used to evaluate season, year, and their interaction on a subset of data focusing on winter and spring. Non-Metric Multidimensional Scaling (NMDS) was performed with $k = 2$ dimensions and ordination was repeated with multiple random starts (up to 200 iterations) to ensure stable solutions. The first two NMDS axes were plotted to visualize community dissimilarity.

Results

Sequencing depth and sample quality

291 Raw 24-hour composite Influent wastewater samples from Columbia, MO were
292 collected weekly from January 2024 through June 2025. Samples were filtered, PEG
293 concentrated, and extracted nucleic acids were sequenced using unbiased ultra-deep
294 sequencing. While this study targeted RNA viruses due to the method used for library
295 preparation, purified nucleic acids contained both RNA and DNA, and DNA viruses were
296 included in the output. Over the study period, 85.8 billion read pairs were obtained with an
297 average depth of 1.1 billion read pairs per sample (**Figure 1**). On average, ribosomal RNA
298 (rRNA) accounted for 66.2 ± 20.7% of reads per sample (range 14.2% - 95.5%). No strong
299 correlations were observed ($|\rho| \geq 0.5$) between rRNA abundance and temperature, water
300 pH, total suspended solids, average influent flow rate over the sampling period, or
301 chemical oxygen demand associated with samples. However, high rRNA abundances are
302 expected in nonselective metagenomic sample preps [21, 22]. The average RT-qPCR Ct
303 value for the wastewater-ubiquitous plant virus PMMoV was 24.2 ± 1.1, providing evidence
304 of viral concentration, even when overall RNA concentrations were low or below the
305 detection limit.

306 *Definitive taxonomic profile*

307 We used *GOTTCHA2* to identify reads that matched unique genomic regions
308 specific to each taxonomic level to provide a baseline for species that are unambiguously
309 present in each sample. The average percentage of sequences that could be assigned at
310 the species level using *GOTTCHA2* was 3.79% ± 1.78%. This represents approximately 40
311 million reads per sample, even though this excludes sequences that map to more than one
312 species. In nearly all samples, eukaryotic viruses were more abundant than prokaryotic
313 viruses (**Figure 2A**). The most abundant viral family, *Virgaviridae*, consists mostly of tomato
314 brown rugose fruit virus (ToBRFV), with PMMoV also consistently present throughout the
315 year but at lower levels (**Figures 2B and 2D**). Notably, other plant viruses such as tomato
316 mosaic virus (TMV) and cucumber green mottle mosaic virus (CGMMV) were detected
317 predominantly during US growing seasons (**Figure 2D**). A similar breakdown of abundant
318 prokaryotic virus families is shown in **Supplemental Figure 1**.

319 Although prokaryotic viruses were less abundant in most samples, the prokaryotic
320 virus fraction was much more diverse compared to the eukaryotic virus fraction
321 (**Supplemental Figure 2**). Mean Shannon diversity was 3.15 ± 0.42 for prokaryotic viruses
322 and 1.10 ± 0.24 for eukaryotic viruses with a paired mean difference of 2.05 (95% CI 1.94-
323 2.16) that was statistically significant (paired t-test $p < 2.2 \times 10^{-16}$, Wilcoxon $p = 2.51 \times 10^{-14}$).
324 If the distribution of species were even, then these values would roughly equate to
325 prokaryotic virus fractions having 8-fold more detectable species than the eukaryotic virus
326 fraction. However, this difference is primarily due to the strong ToBRFV signal skewing the
327 diversity of the eukaryotic virus fraction, causing subsequent underreporting of eukaryotic
328 viral diversity.

329 *Comparison between dPCR and metagenomic read counts*

330 Many metagenomic studies normalize read counts by an endogenous control, such
331 as PMMoV, so counts are more reflective of true viral signal trends and not skewed by
332 changes in wastewater flow rate or variability in sample preparation. Since ToBRFV was the
333 most abundant and consistent viral signal in our data, we compared SARS-CoV-2
334 metagenomic counts normalized to RPM and SARS-CoV-2 counts normalized against
335 ToBRFV to dPCR findings from Missouri's wastewater surveillance program. Both SARS-
336 CoV-2 counts in RPM and ToBRFV normalized counts were positively correlated with flow
337 normalized dPCR measurements ($R^2 = 0.416$ and $R^2 = 0.471$ respectively), showing
338 normalization slightly improved correlations (**Figure 3**).

339 *Human virus profiles*

340 To better investigate the viruses that infect humans, we developed and applied a
341 custom pipeline. Many viruses and viral groups were detected across the study period,
342 such as SARS-CoV-2, enteroviruses, rhinoviruses, and parechovirus A (**Figure 4**) while
343 other viruses and viral serotypes were only present at distinct seasons or timeframes
344 (**Figure 5**). For example, during the winter and early spring months of both years, the
345 classic respiratory seasonal viruses such as influenza A virus, and human coronaviruses
346 NL63 and OC43 were readily detected, while during the summer months these viruses

347 were virtually absent (**Figure 5**). Some respiratory viruses appeared to surge during specific
348 seasons. For instance, in both 2024 and 2025 influenza A virus was predominantly
349 detected from January to March, while parainfluenza virus 3 detection was primarily from
350 March to June, which is consistent with what is observed clinically [32, 33]. Enterovirus
351 D68, on the other hand, was detected primarily in the fall, consistent with clinical
352 observations [34]. Some of the viruses detected had a more complex distribution when the
353 specific serotypes were considered. For instance, rhinoviruses were detected year-round,
354 with higher detection levels in the winter and spring (**Figure 4**). However, when individual
355 serotypes were analyzed, there were distinct waves (**Figure 5**). While most rhinovirus
356 serotypes were abundant in the late spring months (e.g. A58, A18, and A68 in the spring of
357 2024), some serotypes surged in the fall and early winter, such as rhinovirus C42 in late
358 2024. Notably, while rhinoviruses were most pronounced in the spring, the season-specific
359 serotypes were not the same from year to year. For instance, rhinoviruses A61 and A30
360 were both prevalent in the spring of 2025, but were completely absent in the spring of
361 2024. Rhinovirus infections are rarely typed, so it is difficult to say if these observations are
362 consistent with what was seen clinically.

363 The parechoviruses also displayed a year-to-year serotype change. In 2024,
364 parechovirus 5 was the main serotype detected in 2024, while parechovirus 3 was the
365 primary serotype in 2025. This is consistent with a report that demonstrated that
366 parechovirus 5 emerged as the dominant serotype in nearby Kansas City in 2024 [35].

367 Some of the human-infecting viruses detected were not expected. One example
368 was detection of the respiratory pathogen influenza C virus, an influenza type that
369 generally causes milder disease and is typically not included in respiratory panels (**Figures**
370 **4 and 5**). Influenza C virus was the most frequently detected species of influenza in the
371 winter of 2023/24, but did not return the following winter. Another unexpected finding from
372 wastewater was the detection of influenza A virus subtype H5N1 from early March-May of
373 2024. The assembled sequence matched the influenza H5N1 strain B3.13 found in the
374 mammary tissue of dairy cattle and reported by the USDA in late March 2024, followed by
375 dissemination to 16 states by the end of the year [36]. Additionally, our H5N1 detections

376 align to the same timeframe as detections throughout Texas wastewater [37]. While
377 Missouri did not have any reported H5N1 dairy herds, the signal was believed to be derived
378 from a local dairy that received milk for processing from a state with H5N1-infected herds.
379 This finding highlights the fact that this metagenomic approach can detect pathogen
380 surges that are likely to be missed by standard patient surveillance.

381 Viruses associated with gastrointestinal disease, such as norovirus genogroup II,
382 norovirus genogroup I, rotavirus A, astroviruses, Sapporo virus, human mastadenovirus F,
383 and human picobirnavirus were detected throughout the study (**Figure 6**). Some viruses,
384 such as rotavirus A and astrovirus MLB, displayed slight increases during colder months. A
385 month-long surge of hepatitis A virus was detected the summer of 2024, but the signal did
386 not return for the remainder of this study. As expected for wastewater, read counts for
387 enteric viruses were higher than detected respiratory viruses (**Figures 4-6**). Despite the
388 lower counts associated with respiratory viruses, all reported human-infecting viruses
389 were captured with sufficient coverage to assemble enough reads to confidently assign
390 taxonomy.

391 When analyzed in terms of the relative abundance of viruses at the species level
392 using only viral families listed in *NVD* analysis, season explained 18.8% of the variance of
393 sample groupings for *NVD*, and 27.3% of the variance for *GOTTCHA2* ($p < 0.001$ for both
394 methods; **Figure 7**). Additionally, both year and the interaction between season and year
395 were significant ($p < 0.01$) for spring and winter suggesting that seasonal trends are mostly
396 consistent between years with slight shifts in community composition from year to year.
397 The seasonality associated with this data reflects known epidemiological patterns and
398 reinforces previous studies that show unbiased metagenomic sequencing captures
399 expected biological signals in wastewater samples.

400 **Discussion**

401 This study demonstrates the feasibility of an untargeted sequencing approach to
402 detect human infecting viral sequences in deep wastewater data. By generating
403 approximately one billion reads per week, we were able to detect a wide array of viruses,

404 across a range of hosts, with sufficient sensitivity to track temporal trends and detect low-
405 abundant or clinically overlooked pathogens. A major challenge in generating this type of
406 data is the high proportion of rRNA sequences in the final output. While these ribosomal
407 signals dilute target signals, they also provide important contextual information about the
408 microbial communities, potential hosts, and other species signatures in wastewater.
409 Additionally, removing rRNA during sample preparation introduces risks of removing target
410 sequences. Given the low rate of classification of non-rRNA sequences using standard
411 databases, there is a demand to generate a database of wastewater associated
412 sequences for classifying future samples. While low classification rates in this study are
413 consistent with previous work, Guajardo-Leiva, S. et al have recently shown that family
414 level resolution is currently possible [24, 26, 38]

415 The detection of influenza C virus, and H5N1 influenza demonstrate that minimally
416 biased sequencing approaches can capture novel and clinically overlooked pathogens in a
417 time-sensitive manner. Additionally, the correlation between metagenomic read counts,
418 dPCR, and clinical cases suggests that this method provides similar trends to current
419 methods already being used for surveillance. The consistent detection of clinically relevant
420 viruses such as SARS-CoV-2, parechovirus A, rhinovirus A, norovirus genogroup II,
421 norovirus genogroup I, rotavirus A, astroviruses, Sapporo virus, human mastadenovirus F,
422 and human picobirnavirus across all seasons strengthens the value of metagenomics for
423 tracking endemic viral transmission in wastewater. Seasonal increases in virus abundance
424 were observed for several of these pathogens, such as influenza virus, enterovirus A, and
425 rotavirus A during colder months, which is consistent with established clinical trends. The
426 similar human virus trends observed by two different taxonomic assignment methods
427 between seasons reinforces the observations made in this study that are reflective of a
428 wastewater dataset that follows traditionally expected temporal trends [2]

429 Beyond human health, the wide breath of other detectable species emphasizes the
430 expanded utility of metagenomic surveillance. We selected *GOTTCHA2* for definitive
431 baseline characterization because this method is well suited for highly heterogeneous
432 samples such as wastewater that contain a wide range of organisms such as viruses,

433 bacteria, archaea, fungi, plants, and animals. The low classification rates using this
434 method are a direct result of the reference database only containing unique kmers, and
435 our classified percentages underrepresent the total amount of classifiable sequences.
436 Other classifiers, such as Kraken2 using the PlusPFP database, have potential to classify
437 93-99% of rRNA sequences or 14-55% of non-rRNA sequences (data not shown). However,
438 because this method contains non-unique kmers in the reference database, classified
439 reads are not always accurate. This ambiguity led us to our decision to use *GOTTCHA2* for
440 baseline characterization in this study due to its low false positive rate.

441 The detection of highly abundant plant viruses such as ToBRFV and TMV highlights
442 this expanded utility in detecting unexpected sequences. While these viruses have
443 geographic distributions that overlap with Missouri, it is likely that these viruses are
444 introduced to wastewater through consumption of crops within *Solanaceae*, such as
445 tomato and peppers, or suggest disease outbreaks in nearby plant communities, and
446 further highlights the utility of an expanded breath of detectable outbreaks when using
447 metagenomic approaches to detect and identify viruses. While this study provides
448 evidence that ToBRFV may be prevalent enough to be used as an endogenous control,
449 similar to PMMoV, the presence of ToBRFV is not consistent across all studies [26]. In
450 addition to ToBRFV, the detection of plant viruses in this study is consistent with previous
451 work showing high abundance of plant viruses in wastewater [23]. These results, in
452 combination with human viral detection, demonstrate that deep metagenomic sequencing
453 can support timely, minimally biased, and ecologically broad surveillance that supports
454 the initiatives of CDC's OneHealth.

455 The dataset generated here has already proven valuable for the development of
456 foundation large language models that reduce computational burden for downstream
457 users, particularly those lacking high-performance computing resources, to detect low
458 abundant human viruses and putative chimeric viruses [39]. Additionally, similar
459 heatmaps, such as **Figures 4-6**, are publicly available so that trends can be followed in real
460 time during the continuation of this study. The dataset produced here provides a starting
461 point for other research efforts to begin development of new bioinformatic workflows to

further improve the detection of impactful nucleic acid signatures in deep data. As sequencing costs continue to decline and analytical workflows improve, our findings support broader implementation of this approach as a scalable solution for public health preparedness and environmental monitoring.

Acknowledgments

Authors would like to thank the University of Missouri Genomics Core for technical support in weekly sequencing efforts and optimization of library preparation protocols. Additionally, authors would like to thank the Research Support Solutions in the Division of IT at the University of Missouri and the Center for High Throughput Computing at the University of Wisconsin-Madison for access to high performance computing infrastructure (Columbia MO DOI: https://doi.org/10.32469/10355/97710). We would also like to thank Tami Hansen and the wastewater treatment plant in Columbia, MO for weekly sample collection.

Author Contributions

Clayton Rushford, Data curation, Formal analysis, Investigation, Methodology, Project Administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing | Devon Gregory, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review and editing | Emma Copen, Data Curation, Investigation, Methodology, Resources, Writing – review and editing | Jonathan Naydenov, Methodology | Alejandro Tovar-Mendez, Methodology, Writing – review and editing | Po-E Li, Data curation, Formal analysis, Software | Rose S. Kantor, Data curation, Methodology, Validation, Writing – review and editing | Migun Shakya, Data curation, Methodology, Validation, Writing – review and editing | Nelson Ruth, Software | Patrick S. G. Chain, Software | Hsinyeh Hsieh, Investigation, Methodology, Resources, Writing – Review & Editing | Torin Hunter, Investigation, Methodology, Writing – Review & Editing | Josh Kome, Software | Matt Frank, Investigation | Terri Lyddon, Resources | Jeff Kaufman, Conceptualization, Funding Acquisition, Methodology, Project administration, Writing – Review & Editing | David H.

O'Connor, Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Methodology, Resources, Software, Validation, Visualization, Writing – Original Draft, Writing – Review & Editing | Marc C. Johnson, Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft, Writing – Review & Editing.

Data Availability

Sequence data has been uploaded to the NCBI Sequence Read Archive under accession number PRJNA1247874. The workflow for identifying human-infecting virus sequences with high confidence is available at https://github.com/dhoconno/nvd/. The workflow for *GOTTCHA2* analysis, and subsequent visualization of *NVD* and *GOTTCHA2* results are available at https://github.com/ClaytonRushford/Columbia_MO_viral_surveillance_2024-2025/.

Supplemental material

Table S1: Breakdown of treatments in each sample and sequencing platform used.

Table S2: Host annotation for *GOTTCHA2* detected phyla.

Conflict of Interest

Authors declare no conflict of interest.

Funding

Funder	Grant	Author
Open Philanthropy		Jeff Kaufman
Inkfish		Marc Johnson David H. O'Connor

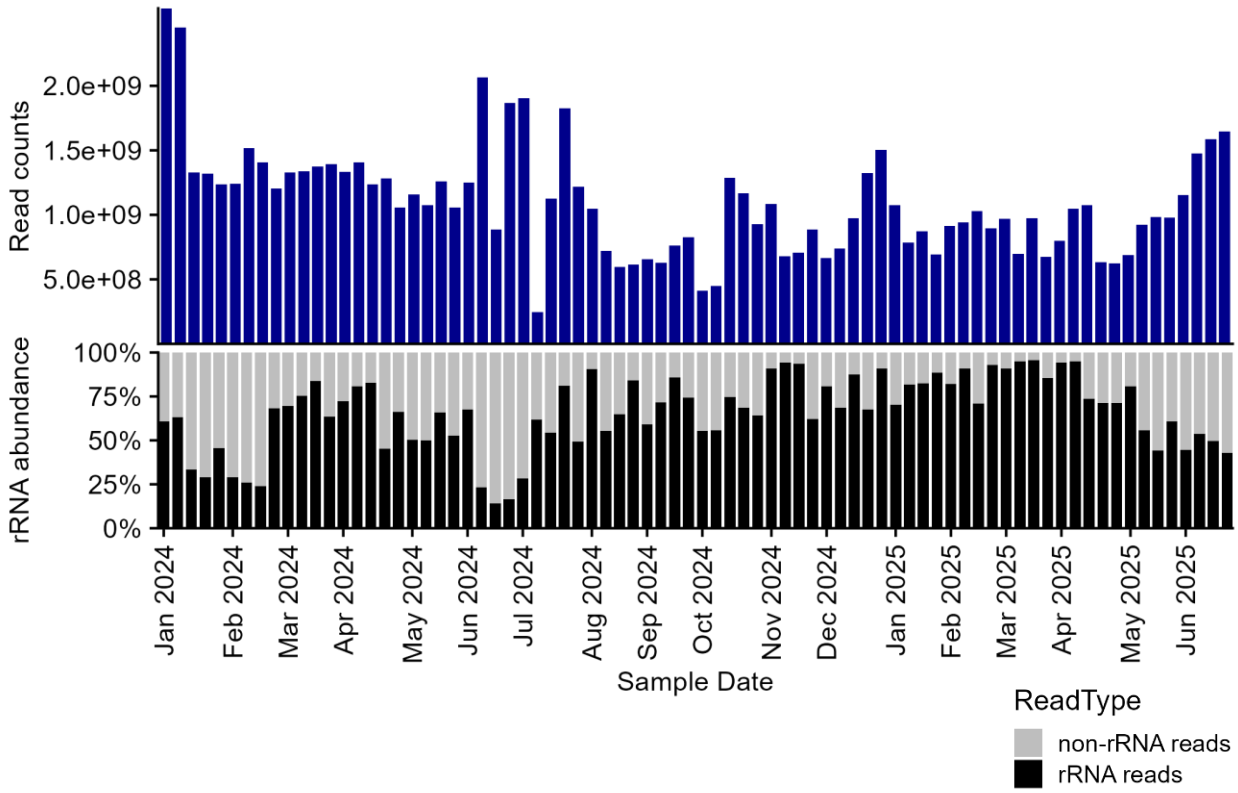
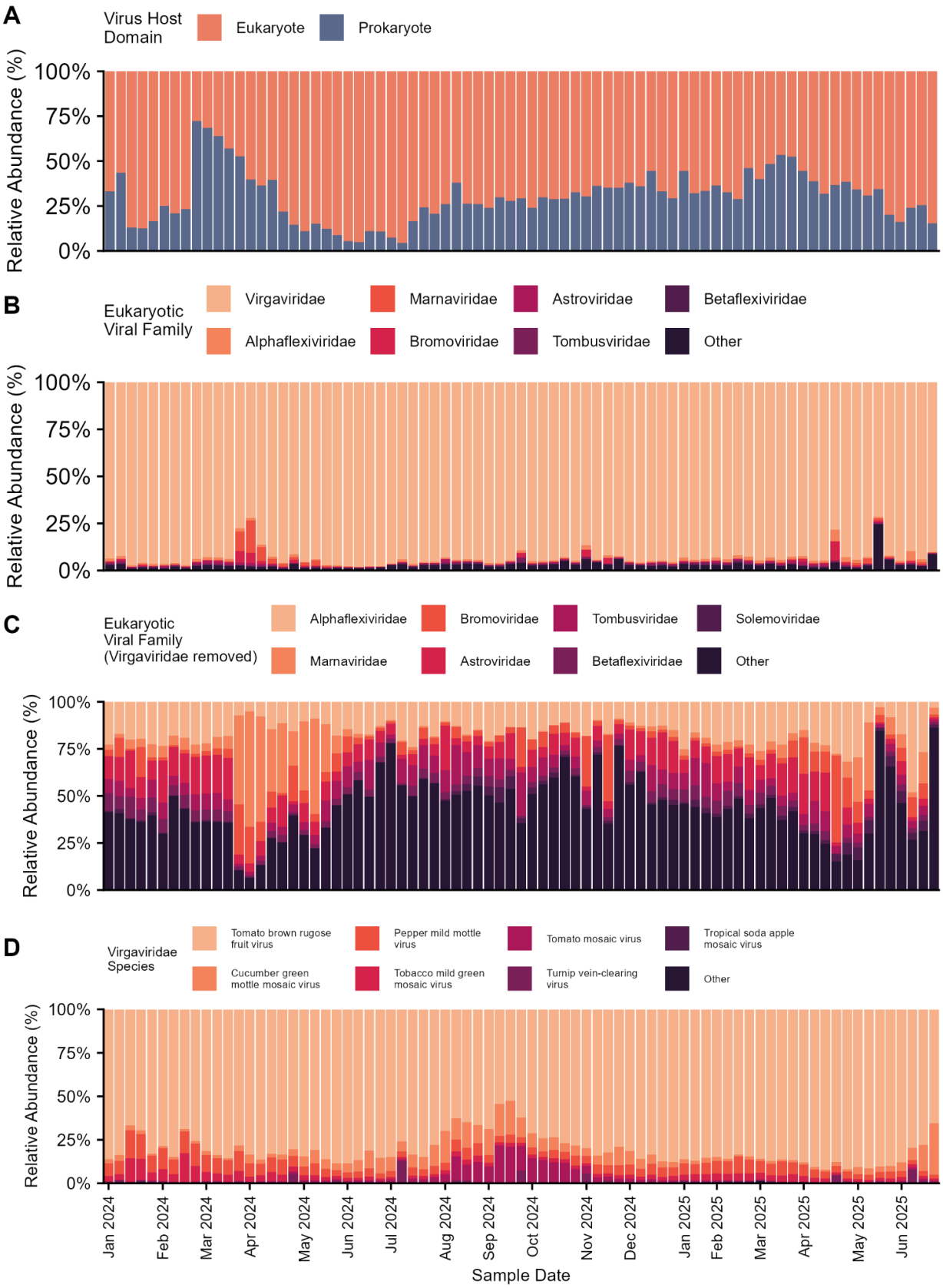
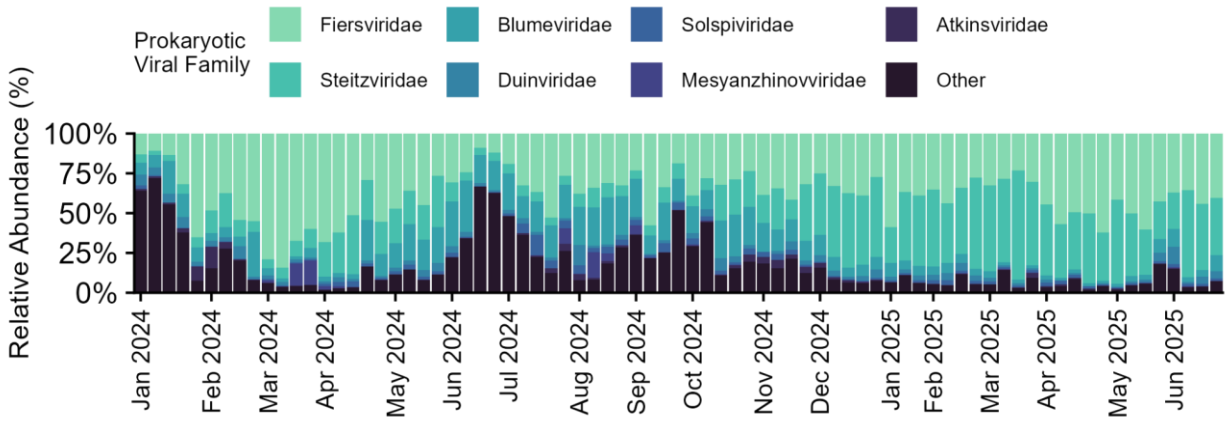


Figure 1) Sequencing depth and rRNA abundance of Columbia MO wastewater samples from January 2024 through June 2025. Blue bars represent total read counts for each sample date. The x-axis indicates the first sample date of the corresponding month labeled.


A
Virus Host Domain
Eukaryote
Prokaryote
Relative Abundance (%)
100%
75%
50%
25%
0%
B
Eukaryotic Viral Family
Virgaviridae
Marnaviridae
Astroviridae
Betaflexiviridae
Alphaflexiviridae
Bromoviridae
Tombusviridae
Other
Relative Abundance (%)
100%
75%
50%
25%
0%
C
Eukaryotic Viral Family (Virgaviridae removed)
Alphaflexiviridae
Bromoviridae
Tombusviridae
Solemoviridae
Marnaviridae
Astroviridae
Betaflexiviridae
Other
Relative Abundance (%)
100%
75%
50%
25%
0%
D
Virgaviridae Species
Tomato brown rugose fruit virus
Pepper mild mottle virus
Tomato mosaic virus
Tropical soda apple mosaic virus
Cucumber green mottle mosaic virus
Tobacco mild green mosaic virus
Turnip vein-clearing virus
Other
Relative Abundance (%)
100%
75%
50%
25%
0%
Jan 2024
Feb 2024
Mar 2024
Apr 2024
May 2024
Jun 2024
Jul 2024
Aug 2024
Sep 2024
Oct 2024
Nov 2024
Dec 2024
Jan 2025
Feb 2025
Mar 2025
Apr 2025
May 2025
Jun 2025
Sample Date

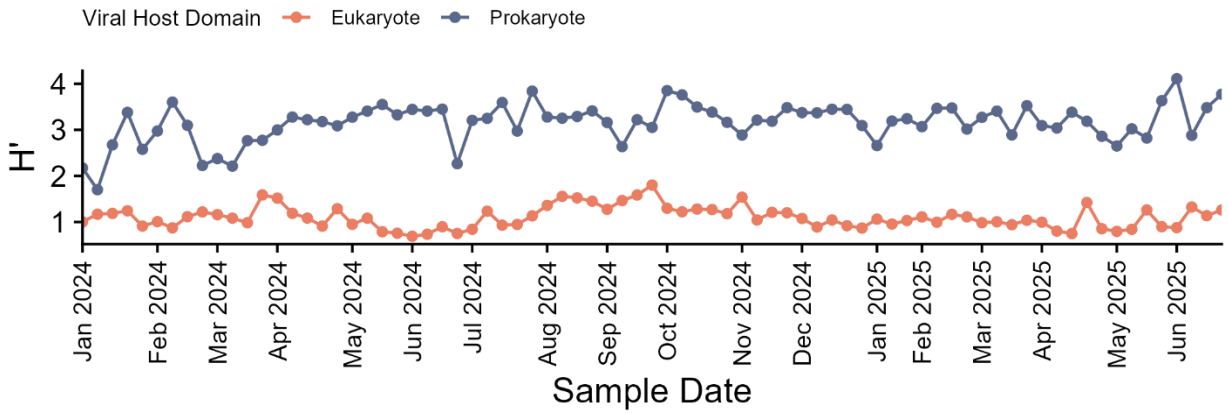
522

523 **Figure 2**) Relative abundance of *GOTTCHA2* classified viral reads out of total classified
524 viral reads, colored by host domain based on manual annotation (A). Relative abundance
525 of the seven most abundant eukaryotic infecting virus families (B). Relative abundance of
526 the seven most abundant families within eukaryotic viruses after removal of the family
527 *Virgaviridae* (C). Breakdown of the most abundant species within the family *Virgaviridae*
528 (D).



529

530 **Supplemental Figure 1**) Relative abundance of *GOTTCHA2* classified viral families with
531 prokaryotic hosts.

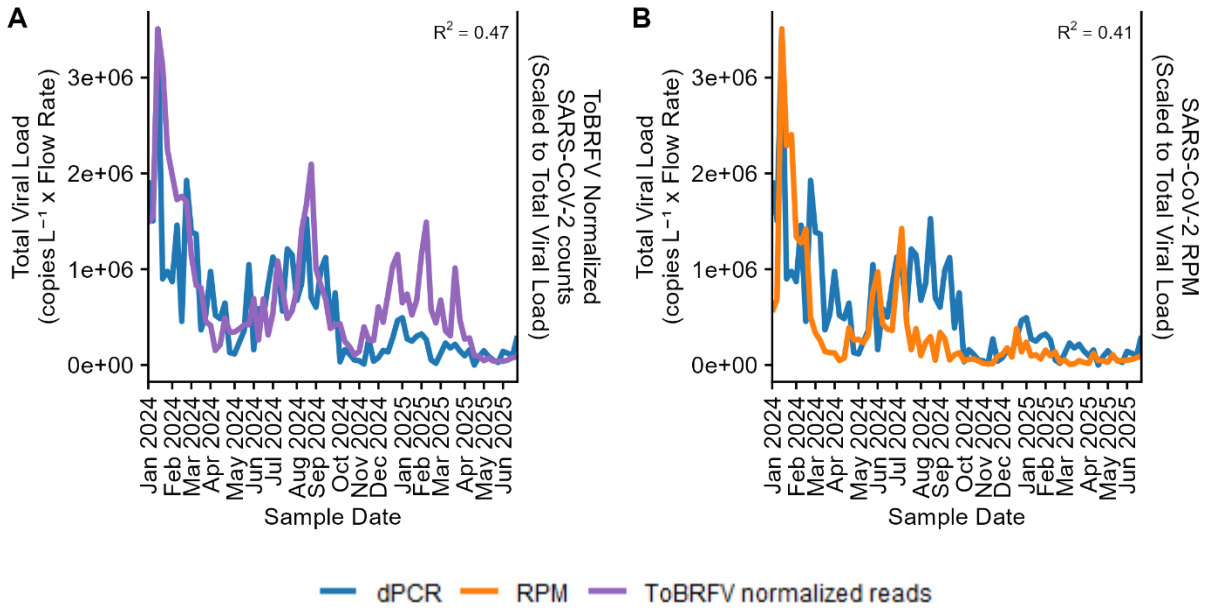


532

533 **Supplemental Figure 2**) Shannon-Wiener diversity index values for *GOTTCHA2* assigned
534 prokaryotic infecting (Blue) and eukaryotic infecting (Orange) virus groups. H' was analyzed
535 on the species level.

536

537



538

539

540 **Figure 3)** Comparison of SARS-CoV-2 read counts derived from metagenomics and dPCR
541 results from Missouri's Wastewater Surveillance Program from January 2024 through June
542 2025. SARS-CoV-2 read counts normalized to ToBRFV (A) or RPM (B) were scaled prior to
543 plotting by dividing the max value of clinical cases or total viral load by the max value of
544 SARS-CoV-2 read counts.

545

546

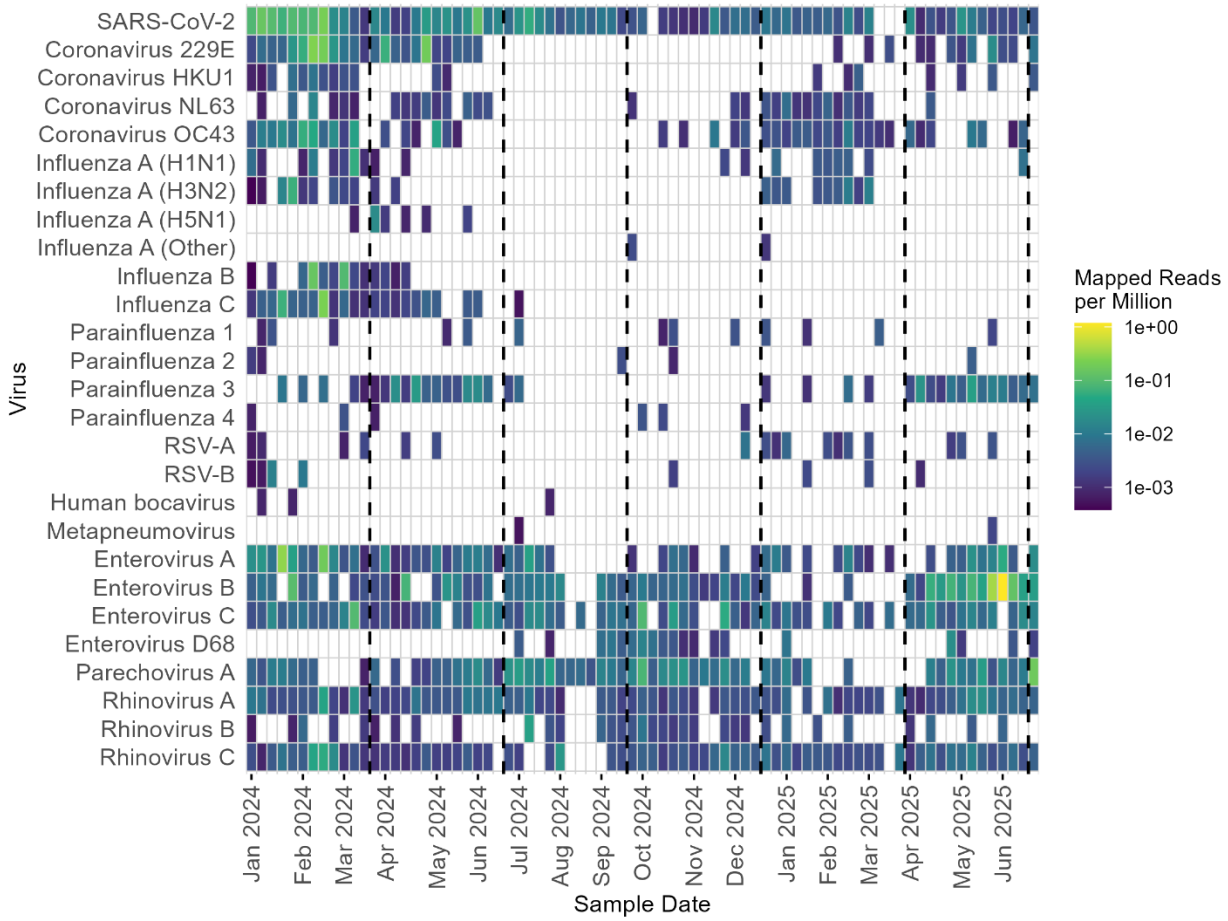


Figure 4) Relative abundance of human respiratory viruses in Columbia MO wastewater from January 2024 to June 2025. Reads counts were collected by *NVD* and normalized to mapped reads per million (RPM) prior to plotting as opposed to mapped reads per kilobase per million (RPKM) due to groupings of viruses with different genome lengths in some rows. A $\log_{10}$ color scale was used for visualization to enhance dynamic range and improve interpretability of variation across samples. Vertical dashed lines signify solstice and equinox dates to highlight season timeframes.

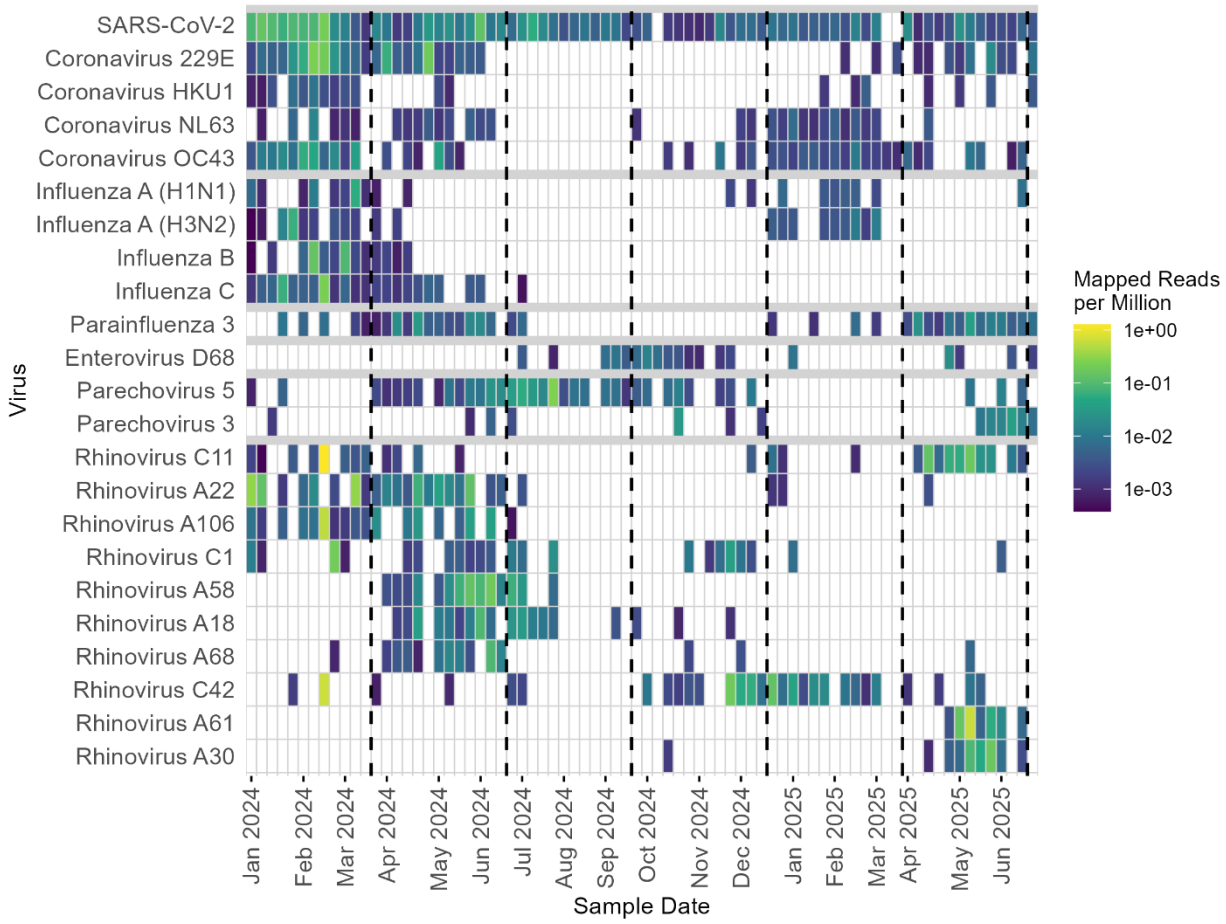


Figure 5) Distribution of individual *NVD* assigned viruses associated with human respiratory disease that exhibit a seasonal trend. Reads counts were normalized to mapped reads per million prior to plotting to keep normalization consistent across heatmaps. A $\log_{10}$ color scale was used for visualization to enhance dynamic range and improve interpretability of variation across samples. Vertical dashed lines signify solstice and equinox dates to highlight season timeframes.

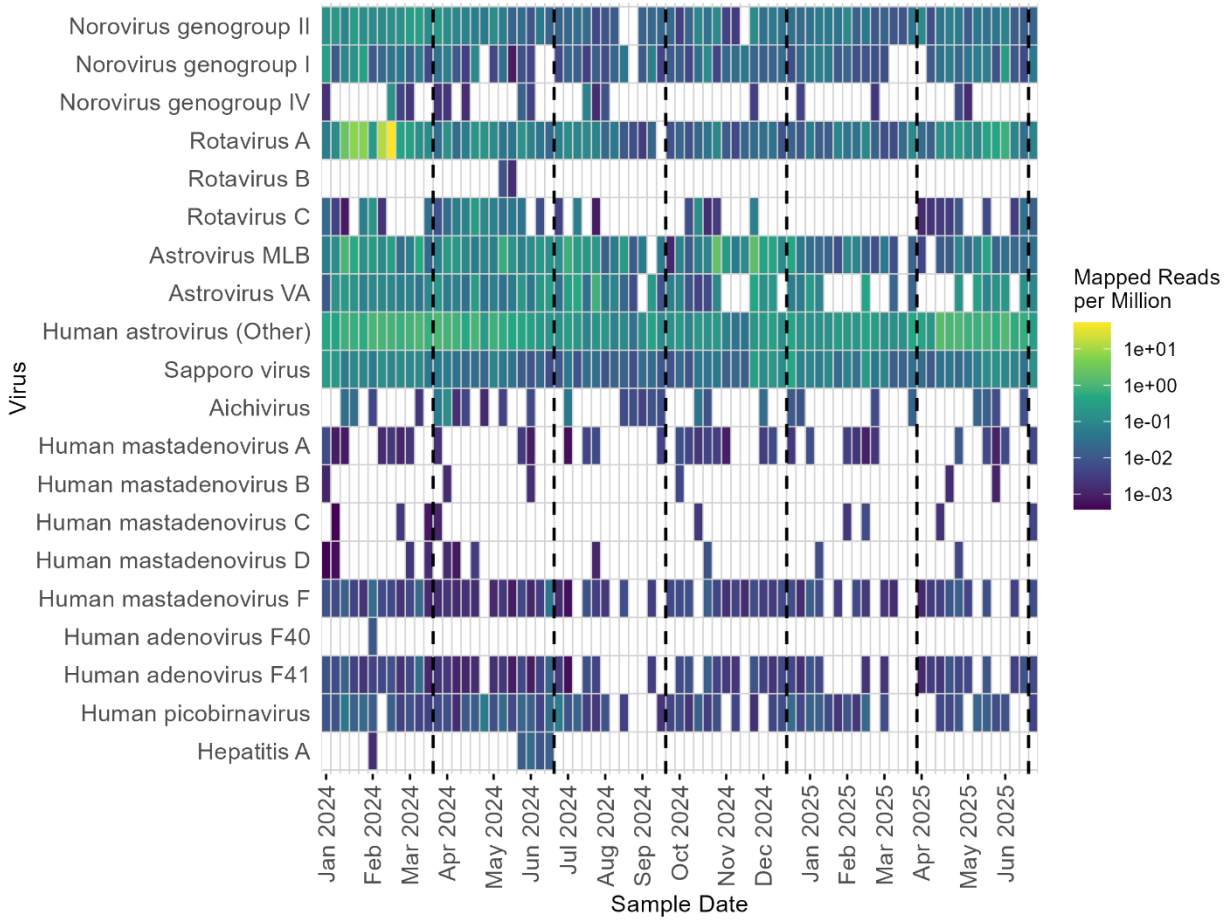


Figure 6) Distribution of *NVD* assigned virus reads associated with human gastrointestinal disease in Columbia MO wastewater from January 2024 to June 2025. Reads counts were normalized to mapped reads per million prior to plotting as opposed to mapped reads per kilobase per million due to high level groupings of viruses with different genome lengths. A $\log_{10}$ color scale was used for visualization to enhance dynamic range and improve interpretability of variation across samples. Vertical dashed lines signify solstice and equinox dates to highlight season timeframes.

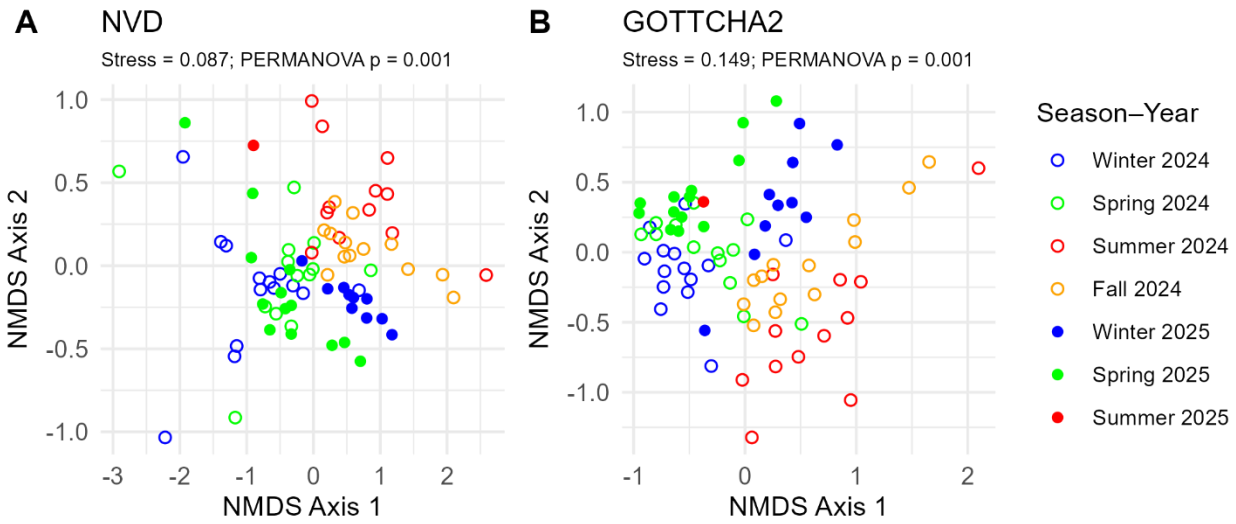


Figure 7) Sample clustering using non-metric multidimensional scaling of Bray-Curtis dissimilarity using human infecting virus counts derived from *NVD* and *GOTTCHA2*. Seasonal groupings were significant based on PERMANOVA results ($p < 0.001$ for both analysis).

References

1. Grassly, N.C., A.G. Shaw, and M. Owusu, *Global wastewater surveillance for pathogens with pandemic potential: opportunities and challenges*. The Lancet Microbe, 2024.
2. Smith, M.F., et al., *Seasonality of respiratory, enteric, and urinary viruses revealed by wastewater genomic surveillance*. Msphere, 2024. **9**(5): p. e00105-24.
3. Kilaru, P., et al., *Wastewater surveillance for infectious disease: a systematic review*. American journal of epidemiology, 2023. **192**(2): p. 305-322.
4. Majumdar, M., et al., *Detection of Enterovirus D68 in wastewater samples from the United Kingdom during outbreaks reported globally between 2015 and 2018*. BioRxiv, 2019: p. 738948.
5. Jahn, K., et al., *Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC*. Nature Microbiology, 2022. **7**(8): p. 1151-1160.
6. Kim, N.-Y., et al., *Wastewater Knows Pathogen Spread: Analysis of Residential Wastewater for Infectious Microorganisms including SARS-CoV-2*. Infection & Chemotherapy, 2023. **55**(2): p. 214.
7. Kantor, R.S. and M. Jiang, *Considerations and Opportunities for Probe Capture Enrichment Sequencing of Emerging Viruses from Wastewater*. Environmental Science & Technology, 2024. **58**(19): p. 8161-8168.
8. Santos, A.F., et al., *Wastewater Metavirome Diversity: Exploring Replicate Inconsistencies and Bioinformatic Tool Disparities*. International Journal of Environmental Research and Public Health, 2025. **22**(5): p. 707.
9. Pogka, V., et al., *Targeted virome sequencing enhances unbiased detection and genome assembly of known and emerging viruses—the example of SARS-CoV-2*. Viruses, 2022. **14**(6): p. 1272.
10. McCall, C., et al., *Targeted metagenomic sequencing for detection of vertebrate viruses in wastewater for public health surveillance*. ACS ES&T Water, 2023. **3**(9): p. 2955-2965.
11. Tisza, M., et al., *Wastewater sequencing reveals community and variant dynamics of the collective human virome*. Nature Communications, 2023. **14**(1): p. 6878.
12. Zulli, A., et al., *Probe-Based Enrichment Sequencing Applied to Wastewater Surveillance Accurately Tracks Multiple Viral Respiratory Outbreaks*. ACS ES&T Water, 2025.
13. Javornik Cregeen, S., et al., *Sequencing-Based Detection of Measles in Wastewater: Texas, January 2025*. American Journal of Public Health, 2025(0): p. e1-e5.
14. Clark, J.R., et al., *Wastewater pandemic preparedness: Toward an end-to-end pathogen monitoring program*. Frontiers in Public Health, 2023. **11**: p. 1137881.
15. Thornton, M., et al., *Comparative wastewater virome analysis with different enrichment methods*. Water Research, 2025: p. 123985.
16. Liu, S., et al., *Analysis of metagenomic data*. Nature Reviews Methods Primers, 2025. **5**(1): p. 5.
17. Wetterstrand, K. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program*.

18. Rajput, V., et al., *Metagenomics based longitudinal monitoring of antibiotic resistome and microbiome in the inlets of wastewater treatment plants in an Indian megacity*. Science of The Total Environment, 2025. **986**: p. 179691.
19. Becsei, Á., et al., *Time-series sewage metagenomics distinguishes seasonal, human-derived and environmental microbial communities potentially allowing source-attributed surveillance*. Nature communications, 2024. **15**(1): p. 7551.
20. Child, H.T., et al., *Comparison of metagenomic and targeted methods for sequencing human pathogenic viruses from wastewater*. MBio, 2023. **14**(6): p. e01468-23.
21. Li, Y., et al., *A broad wastewater screening and clinical data surveillance for virus-related diseases in the metropolitan Detroit area in Michigan*. Human Genomics, 2024. **18**(1): p. 14.
22. Cantalupo, P.G., et al., *Raw sewage harbors diverse viral populations*. MBio, 2011. **2**(5): p. 10.1128/mbio. 00180-11.
23. Nieuwenhuijse, D.F., et al., *Setting a baseline for global urban virome surveillance in sewage*. Scientific reports, 2020. **10**(1): p. 13748.
24. Guajardo-Leiva, S., J. Chnaiderman, A. Gaggero, and B. Díez, *Metagenomic insights into the sewage RNA virosphere of a large city*. Viruses, 2020. **12**(9): p. 1050.
25. Gulino, K., et al., *Initial mapping of the New York City wastewater virome*. Msystems, 2020. **5**(3): p. 10.1128/msystems. 00876-19.
26. Guajardo-Leiva, S., et al., *From sewage to genomes: Expanding our understanding of the urban and semi-urban wastewater RNA virome*. Environmental Research, 2025. **276**: p. 121509.
27. Wyler, E., et al., *Pathogen dynamics and discovery of novel viruses and enzymes by deep nucleic acid sequencing of wastewater*. Environment International, 2024. **190**: p. 108875.
28. Kitajima, M., H.P. Sassi, and J.R. Torrey, *Pepper mild mottle virus as a water quality indicator*. NPJ Clean Water, 2018. **1**(1): p. 19.
29. Bushnell, B. *BBmap*. 2025 2025-06-05 [cited 2025 6-20]; Available from: https://sourceforge.net/projects/bbmap/.
30. Freitas, T.A.K., P.-E. Li, M.B. Scholz, and P.S. Chain, *Accurate read-based metagenome characterization using a hierarchical suite of unique signatures*. Nucleic acids research, 2015. **43**(10): p. e69-e69.
31. Robinson, C.A., et al., *Defining biological and biophysical properties of SARS-CoV-2 genetic material in wastewater*. Science of The Total Environment, 2022. **807**: p. 150786.
32. Russell, E., A. Yang, S. Tardrew, and M.G. Ison, *Parainfluenza virus in hospitalized adults: a 7-year retrospective study*. Clinical Infectious Diseases, 2019. **68**(2): p. 298-305.
33. Control, C.f.D. and Prevention, *Influenza activity in the United States during the 2023–2024 season and composition of the 2024–2025 influenza vaccine*. 2024.
34. Grizer, C.S., K. Messacar, and J.J. Mattapallil, *Enterovirus-D68–a reemerging non-polio enterovirus that causes severe respiratory and neurological disease in children*. Frontiers in Virology, 2024. **4**: p. 1328457.

686 35. Dhar, D., et al., *Emergence of Parechovirus-A5 central nervous system infections in*
687 *children from Kansas City, Missouri, USA*. Journal of Clinical Virology, 2025: p.
688 105835.
689 36. Campbell, A., K. Brizuela, and S.S. Lakdawala, *mGem: Transmission and exposure*
690 *risks of dairy cow H5N1 influenza virus*. MBio, 2025. **16**(3): p. e02944-24.
691 37. Tisza, M.J., et al., *Virome Sequencing Detects H5N1 Avian Influenza in Wastewater*
692 *in Ten Cities*. The New England journal of medicine, 2024. **391**(12): p. 1157.
693 38. Tamaki, H., et al., *Metagenomic analysis of DNA viruses in a wastewater treatment*
694 *plant in tropical climate*. Environmental microbiology, 2012. **14**(2): p. 441-452.
695 39. Liu, O., et al., *METAGENE-1: Metagenomic Foundation Model for Pandemic*
696 *Monitoring*. arXiv preprint arXiv:2501.02045, 2025.

697