



Prospective metagenomic sequencing of wastewater across the United States yields robust viral enrichment and concordance with digital PCR measurements

Marlene K. Wolfe,¹ Devin North,¹ Alexander L. Jaffe,² Alessandro Zulli,² Dorothea Duong,³ Bridgette Shelden,³ Miriam Goldman,³ Miles Richardson,³ Peter Thana,³ Vikram Chan-Herur,³ Pouya Kheradpour,³ Amanda L. Bidwell,² Stephen P. Hilton,¹ Sheena Conforti,² Abigail P. Paulos,¹ Alexandria B. Boehm²

AUTHOR AFFILIATIONS See affiliation list on p. 19.

ABSTRACT Metagenomic sequencing is increasingly applied to wastewater to characterize the diversity, dynamics, and relative abundance of human and animal viruses. Among these sequencing approaches are those that enrich viral nucleic acids from the wastewater matrix, aiming to increase the viral read fraction for analysis. However, the feasibility of scaling targeted viral sequencing to diverse sewersheds across large geographic scales is currently unknown. In this study, we apply hybrid capture metagenomic sequencing to nearly 450 weekly wastewater samples collected during the respiratory virus season in the United States and evaluate sequencing performance for generating public health-relevant data. Analysis of data from 15 wastewater treatment plants demonstrates that our approach enabled efficient capture of pathogens of interest, achieving a median viral read fraction over 19%. Importantly, relative abundance estimates of common pathogens correlated with direct quantification of viral targets using reverse transcription digital droplet PCR. Together, our results demonstrate that hybrid capture sequencing of wastewater is a viable tool to monitor both common and rare pathogens across geographically diverse sewersheds.

IMPORTANCE Wastewater testing is commonly used to identify and quantify human pathogens at a community scale. However, the most commonly used approaches rely on targeted, PCR-based methods that are highly specific to a single virus. Metagenomic sequencing provides an opportunity to detect and quantify the relative abundance of a wide range of viruses that are important for human health from wastewater, and hybrid capture approaches work by first enriching samples for these extremely rare targets to increase sensitivity. We demonstrate that hybrid capture metagenomic sequencing successfully enriches wastewater samples from diverse locations across the United States and that data derived from sequences are associated with detections of 11 key viruses using standard RT-ddPCR methods. This work suggests that scaling hybrid capture metagenomics for viruses in wastewater is a feasible way to generate data critical for public health to support infectious disease outbreak response.

KEYWORDS wastewater, virus, metagenomics, hybrid capture, epidemiology

Wastewater-based epidemiology (WBE) is a valuable public health tool for monitoring the occurrence of viral infectious diseases including COVID-19, influenza, mpox, and measles (1–4). Public health agencies, including local, regional, and national programs, consider WBE an important and actionable source of epidemiological information, informing early outbreak detection and guiding public health responses (5, 6). As such, a large number of efforts globally measure viral biomarkers prospectively in

Editor Satoshi Ishii, University of Minnesota Twin Cities, St. Paul, Minnesota, USA

Address correspondence to Marlene K. Wolfe, marlene.wolfe@emory.edu.

Marlene K. Wolfe, Devin North, and Alexander L. Jaffe contributed equally to this article. Author order was determined by seniority on the project team.

D.D., B.S., M.G., M.R., P.T., P.K., and V.C.-H. are employees of Verily Life Sciences.

See the funding table on p. 19.

Received 20 May 2026

Accepted 15 August 2026

Published 28 September 2026

Copyright © 2026 Wolfe et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license](#).

wastewater (7). Prospective monitoring efforts largely use PCR to detect viral genomes in wastewater. Quantitative formats, such as qPCR or digital droplet PCR (ddPCR), offer measurement precision and rapid turnaround times (8). However, PCR methods have several key limitations, including the limited number of different assays that can be run together and the need for highly specific and sensitive oligo primers and probes. As viral genomes can evolve over very short time scales (9), primer and probe sequences may require frequent adjustment to ensure assays remain sensitive and specific to circulating variants (10). Primers and probes can only be designed if a viral genomic sequence is known, limiting PCR assay use for highly novel viruses or divergent variants of known viruses.

Genomic surveillance via sequencing offers an attractive alternative to quantitative PCR approaches for viral WBE (11). Tiled amplicon sequencing to characterize SARS-CoV-2 variants has been widely used for years to understand viral evolution, applied to both clinical (12) and wastewater samples (13). Application to wastewater samples has demonstrated close concordance with clinical results when analyzing SARS-CoV-2 lineages (12) and can provide important information on the evolution of SARS-CoV-2 in areas without clinical testing (13). Tiled amplicon approaches have also been applied to sequencing influenza (14), enterovirus D68, norovirus GII, human adenovirus 41, hepatitis A virus, and measles (15) genomes in wastewater. A major drawback to tiled amplicon approaches is that they focus on a single virus.

Recently, several studies have applied metagenomic approaches to viral pathogen WBE (15, 16). Since viral nucleic acids represent a small fraction of the total nucleic acids in wastewater (17), shotgun metagenomic approaches—which generally employ few, if any, steps to enrich for targets of interest—can fail to detect lower-abundance viruses, resulting in highly variable estimates of sensitivity (18). There are two approaches to overcome this limitation. The first is to sequence deeply, generating billions of sequencing reads per sample, thereby enabling the detection of rare viral targets (16, 19). However, at present, deep, untargeted sequencing may not be feasible for prospective viral WBE efforts due to cost and may not provide the high viral genomic coverage needed to identify the emergence of new variants. The second approach uses probe-based hybrid capture to enrich for viral targets before sequencing (20, 21). This “semi-targeted” approach promises to enrich viral sequencing reads by orders of magnitude over shotgun sequencing approaches while being tolerant to up to 30% sequence divergence from known references (22, 23). Additionally, previous work conducted at a single WWTP in New Haven, CT, USA (24) found that viral sequence read counts obtained using a hybrid capture sequencing correlated positively with ddPCR measurements of respiratory and enteric viral nucleic acids. However, the performance of hybrid capture approaches relative to targeted molecular methods under variable sampling conditions remains unknown.

In this study, we evaluate the feasibility of hybrid capture metagenomic sequencing to detect viral biomarkers in wastewater samples collected weekly from 15 WWTPs across the USA. We examine factors impacting the performance of our sequencing method, including the extent of viral diversity recovered, and compare the relative abundance of sequenced viruses to concentrations measured by reverse transcriptase ddPCR (RT-ddPCR). Spatio-temporal dynamics are presented for select viruses.

MATERIALS AND METHODS

Wastewater samples were collected and processed to obtain viral metagenomic sequences and determine the concentration of viral targets using RT-ddPCR. Measurements were obtained from nucleic acids from wastewater solids generated for an existing academic wastewater monitoring program; solids are used for that program because viral nucleic acids sorb to wastewater solids and are enriched by orders of magnitude on a per mass basis in wastewater solids compared to liquids (25, 26). Therefore, they represent good templates for sensitive and specific ddPCR assays

(8). However, for this study, probe-based metagenomic sequencing was applied to wastewater liquids, in contact with the solids, as described below. Because wastewater solids consist of particles larger than 0.2 μm in diameter (including bacterial and eukaryotic cells and debris, and inorganic particles), the ratio of non-viral to viral sequences is high, and probe-capture metagenomic sequencing methods may not adequately identify viral sequences (23). Pilot work suggests this to be the case in our laboratory, so liquids were used for sequencing.

Sample collection and prospective measurements of viral nucleic acids in wastewater solids

Wastewater samples were collected approximately three times per week for 12 weeks from 15 WWTPs ($n = 443$ samples) across 12 states in the United States (Fig. 1; Table 1). WWTP locations were a subset of sites currently participating in an academic wastewater monitoring program and were chosen to represent a variety of geographic regions and communities, with WWTPs serving populations ranging from 31,000 to 3.5 million across the Midwest, Northeast, South, and West regions of the USA. Sample collection used sterile technique and followed the protocol outlined in Boehm et al. (10). Two sample types were collected depending on location and based on WWTP staff preference: grab samples of sludge from the primary clarifier or 24-h composite primary influent samples (Fig. 1; Table 1). Samples were shipped overnight at 4°C to the laboratory for processing.

Immediately upon receipt, aliquots from samples were centrifuged, and the pellets, representing wastewater solids, were processed for viral nucleic acid quantification as part of the ongoing academic wastewater monitoring program using the standard methods for that program (Fig. 2). Targets included nucleic acids of SARS-CoV-2, influenza A and B, respiratory syncytial virus, human metapneumovirus, enterovirus D68, measles, norovirus GII, rotavirus, hepatitis A, parvovirus, and West Nile virus. Quantification was performed using the methods described in detail elsewhere (8); the exception is that on 2 March 2026, a second measles probe was added to account for sequence divergence observed in some circulating genotypes (see SI). Concentrations are reported as copies per gram of dry-weight solids.

Pre-analytical processing, nucleic acid extraction, and purification for metagenomic sequencing

Pre-analytical processing and nucleic acid extraction and purification for metagenomic sequencing were aimed at minimizing the presence of inhibitors and the ratio of non-viral (bacterial and eukaryotic) to viral targets (23) prior to sequencing. Sludge samples from the primary clarifier were centrifuged at 4,200 $\times g$ for 5 min, and 50 mL of the supernatant was collected. Fifty milliliters of primary influent samples was processed without centrifugation. The 50 mL aliquot was stored between 0 and 8 days at 4°C and processed in weekly batches.

Sludge supernatants and influent samples were vacuum-filtered using 0.45 μm pore-size polyethersulfone (PES) filters (Thermo Scientific #165-0045, Waltham, Massachusetts, USA) to remove bacterial and eukaryotic nucleic acids, and the filtrates were collected in sterile containers. A total of 30–40 mL per sample was filtered, and viral particles were then concentrated from 10 mL of filtrate using Ceres Nanotrap Microbiome A particles (Manassas, Virginia, USA) following manufacturer's directions.

Nucleic acids were extracted from the concentrated viral particle fraction using the MagMAX Viral/Pathogen Nucleic Acid Isolation Kit (Waltham, Massachusetts, USA) using an automated KingFisher instrument (ThermoFisher Scientific, Waltham, MA, USA). PCR inhibitors were removed using the Zymo OneStep PCR Inhibitor Removal Kit (Irvine, California, USA). The final volume of the nucleic acid extract obtained from 10 mL of filtrate was 60 μL .

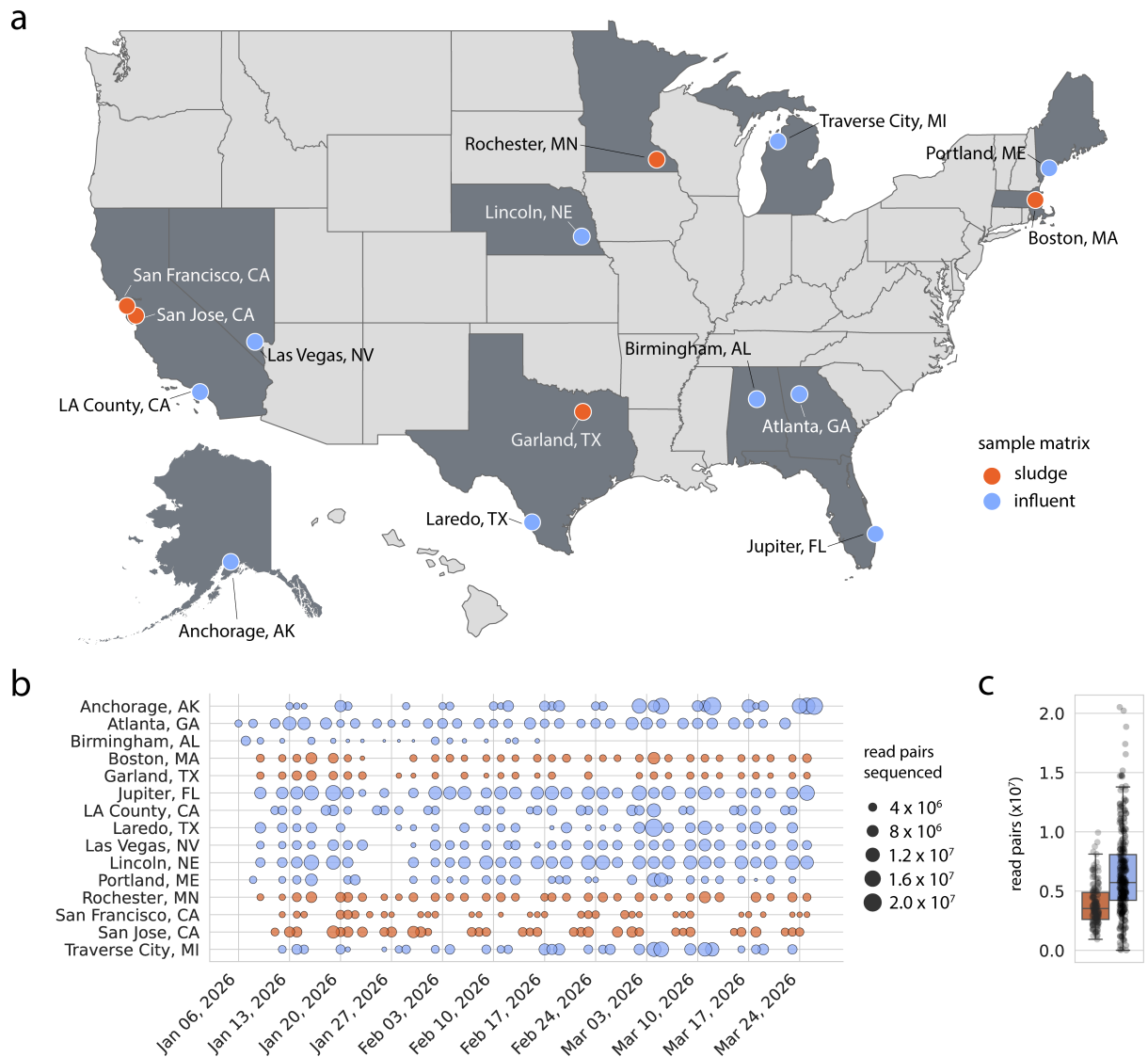


FIG 1 Characteristics of samples analyzed in this study. (a) Map of the 15 wastewater treatment plants (WWTPs) enrolled in this study, along with sample type. States where WWTPs are located are shaded in dark gray. Alaska and Hawaii are not to scale. The map was created using Microsoft Tableau Desktop, using the Web Mercator (EPSG:3857) projection. (b) Temporal frequency of sampling, with the size of points indicating the total number of read pairs sequenced after human read filtering and deduplication. (c) Read pairs sequenced after human read filtering and deduplication as a function of sample matrix.

Library preparation and sequencing

cDNA synthesis was performed with ProtoScript II First Strand cDNA Synthesis Kit (New England Biolabs #E6560L, Ipswich, MA, USA) using the neat nucleic acid as template. cDNA was purified using Kapa HyperPure Beads (Roche, Basel, Switzerland), and the cDNA concentration was measured using Qubit (Thermo Fisher). cDNA was then diluted to 1 ng/ μ L using nuclease-free water for library preparation. The Twist Library Preparation Kit (Twist Biosciences, South San Francisco, CA, #103548) was used for enzymatic fragmentation, adapter ligation, and PCR amplification, with the addition of the Twist Unique Molecular Identifiers (UMI) adapter system. cDNA synthesis efficiency was not assessed and, therefore, could potentially introduce some bias for the detection of viruses with DNA vs RNA genomes.

Prepared libraries were enriched using the Twist Comprehensive Viral Research Panel (Twist Biosciences, #103548), which, as reported by the manufacturer, targets more than 3,000 viral genomes through hybridization at 70°C for 16 h. Bacterial, human, and

TABLE 1 Characteristics of sites included in the study, including location, population served, sample matrix, total number of samples collected per site, and sampling dates (month/day/year format)

Site	Location	Population served	Sample matrix	Total no. of samples	Start date	End date
John M. Asplund Water Pollution Control Facility	Anchorage, Alaska	220,000	Primary influent	28	1/13/26	3/26/26
RM Clayton Water Reclamation Center	Atlanta, Georgia	600,000	Primary influent	33	1/6/26	3/22/26
Village Creek Water Reclamation Facility	Birmingham, Alabama	200,000	Primary influent	17	1/7/26	2/16/26
Deer Island Treatment Plant	Boston, Massachusetts	2,400,000	Sludge	31	1/9/26	3/25/26
City of Garland Rowlett Creek Wastewater Treatment Plant	Garland, Texas	630,000	Sludge	28	1/9/26	3/25/26
Loxahatchee River Environmental Control District	Jupiter, Florida	90,000	Primary influent	32	1/9/26	3/25/26
A.K. Warren Water Resource Recovery Facility	Los Angeles County, California	3,500,000	Primary influent	30	1/11/26	3/23/26
Zacate Creek Wastewater Treatment Plant	Laredo, Texas	140,000	Primary influent	27	1/9/26	3/23/26
Clark County Water Reclamation District Flamingo Water Resource Center	Las Vegas, Nevada	1,250,000	Primary influent	31	1/9/26	3/25/26
Theresa Street Water Resource Recovery Facility	Lincoln, Nebraska	240,000	Primary influent	30	1/9/26	3/25/26
Portland Water District East End Wastewater Treatment Facility	Portland, Maine	65,000	Primary influent	29	1/8/26	3/23/26
City Of Rochester MN Water Reclamation Plant	Rochester, Minnesota	120,000	Sludge	32	1/9/26	3/25/26
Southeast San Francisco Wastewater Treatment Plant	San Francisco, California	750,000	Sludge	33	1/12/26	3/25/26
San Jose-Santa Clara Regional Wastewater Facility	San Jose, California	1,500,000	Sludge	33	1/11/26	3/24/26
Traverse City Regional Waste Water Treatment Plant	Traverse City, Michigan	30,623	Primary influent	29	1/12/26	3/23/26

other mammalian ribosomal RNA (rRNA) was depleted from the samples using CRISPR-Cas9 technology via the CRISPRclean Stranded Total RNA Prep with rRNA Depletion kit (Jumpcode Genomics, San Diego, CA #KIT1014). Prepared libraries were sequenced on an Illumina (Foster City, CA) NovaSeq X Plus (10B flow cells, 150-bp paired-end reads), alongside negative sequencing controls. Negative sequencing controls used DNA/RNA-free water as template for the cDNA synthesis and were moved through the workflow as samples for each plate (11 total were processed). Samples were pooled in batches of 8, with a target sequencing depth of 100 million reads per sample. Full library preparation details are provided on protocols.io (ID 225762) (27).

Computational analyses

After demultiplexing, reads were deduplicated using HUMID (<https://github.com/jfjaros/HUMID>), incorporating information from the unique molecular identifiers (UMIs). We next performed host read removal by iteratively applying Kraken2 (v2.1.3) (28), bowtie2 (v2.5.2) (29) with reference GRCh38.p14, and HRRT (30) (v2.2.1) with the human_filter.db.20250916v2 database. Additionally, raw sequencing reads were trimmed of sequencing adapters and UMIs using fastp (v0.23.4) (31). Taxonomic classification and read quantification were computed using EsVirtu (32) (v1.1.0) with database version 3.2.2. This approach employs competitive alignment with minimap2 against a curated database of high-quality virus genome assemblies, enabling simultaneous detection of many viruses. Intermediate alignment files were retained (*--keep True*), and the initial, unfiltered BAM file was used to compute the number of putative viral read pairs for each sample (Table S1). Specifically, read pairs were designated as viral if one or both mates aligned to an EsVirtu reference sequence during the initial mapping. Putative rRNA reads were classified using sortmerna (33) (v4.3.6) with the smr_v4.3_sensitive_db.fasta database and an alignment e-value threshold of 1×10^{-5} . Read pairs that were preliminarily classified as both viral and rRNA (typically <0.1%) were designated as ambiguous and removed from final viral and rRNA read counts (Table S1).

To examine off-target sequences generated by our protocol, we filtered out reads initially classified as viral or as rRNA using the tools above and assembled the remaining reads using two variants of the SPAdes assembler ([metaspades.py](https://github.com/abranjean/metaspades.py) and [rnaspades.py](https://github.com/abranjean/rnaspades.py))

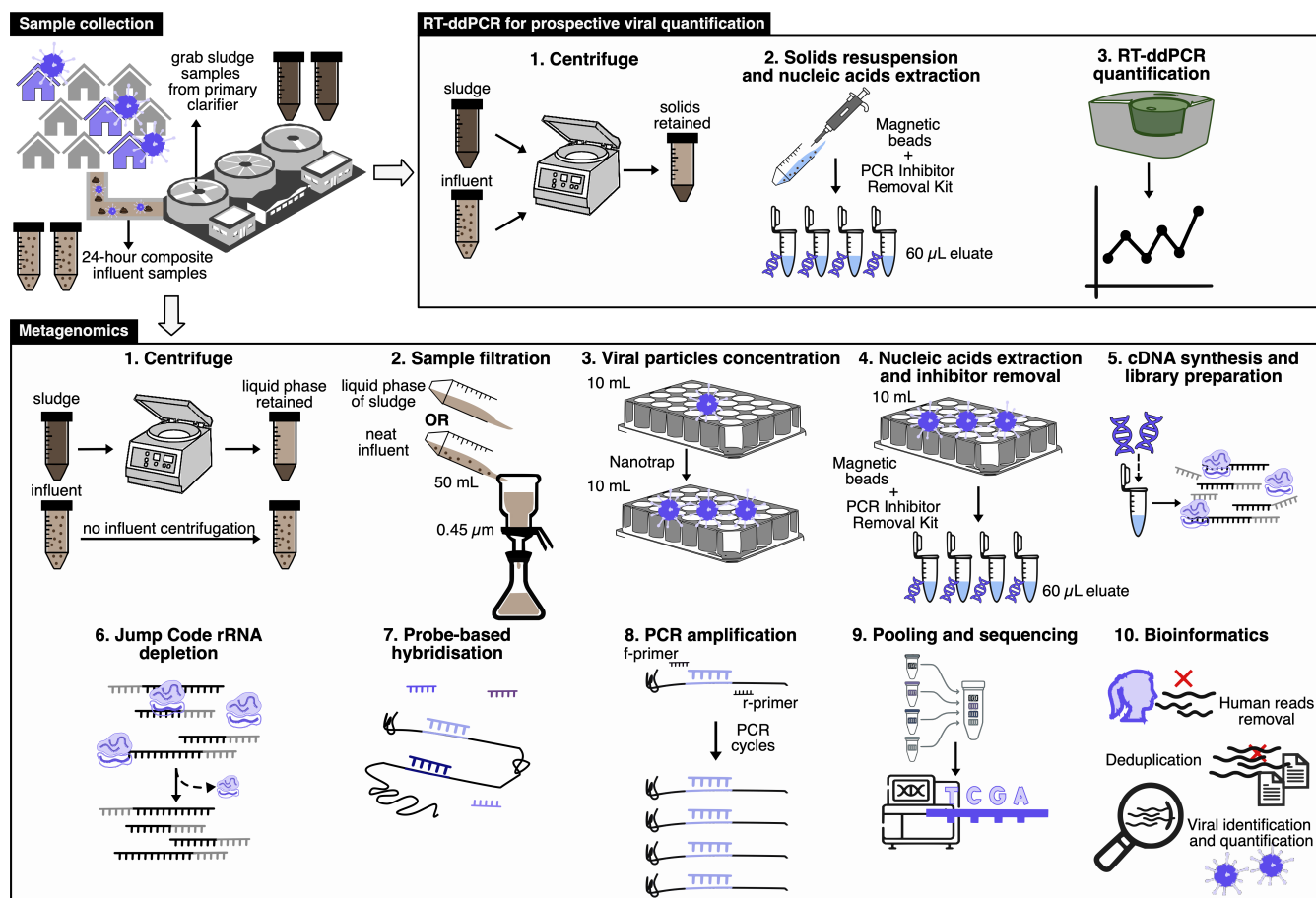


FIG 2 Workflow of wastewater samples processing and comparison between RT-ddPCR and metagenomics sequencing approaches. Abbreviations: RT-ddPCR, reverse transcription digital droplet PCR; cDNA, complementary DNA; rRNA, ribosomal RNA.

(34). Resulting contigs were filtered to those at least 1,000 bp in length. Focusing on three sites—Las Vegas, Traverse City, and Atlanta—we combined all filtered contigs on a per-site basis and clustered them using *mmseqs2* (35) (*easy-linclust --min-seq-id 0.95 --cluster-mode 2 --cov-mode 1*), reasoning that there would be significant sequence redundancy from high-frequency sampling of the same sewersheds over time. For each site, raw metagenomic reads for each sample were competitively mapped back to their corresponding contig set using *bowtie2* (29) and *samtools* (36). Alignment files were processed using *CoverM* (37) with both loose (*--min-read-percent-identity 0*) and stringent (*--min-read-percent-identity 0.95*) mapping parameters, generating per-contig read counts and coverage breadth values for downstream analyses. Contigs were considered “detected” in a sample if their breadth of coverage at 95% read identity met or exceeded 50%. Preliminary classifications were made for contigs recruiting 50,000 or more reads in any sample using *BLAST* against *core_nt* and *refseq_rna* (*blastn*). Figures were generated using the *seaborn* (38) (v0.11.2) package in Python (v3.6.1) and the *ggplot2* (39) package(s) in R.

Direct validation

For a subset of targets classified as associated with high-consequence diseases (HCD), manual confirmation of classification was performed. HCDs are associated with serious illness with high case fatality rates. The list of high-consequence pathogens for this study (Table S2) was based on global HCD classifications or on pathogens that share high genetic similarity with these pathogens (40). For a flagged target, reads were manually

inspected using BAM files from the EsVirtu output. NCBI BLAST was used to manually classify the reads against the entirety of NCBI's core_nt database using the blastn algorithm (41, 42). Exclusionary BLAST was also performed to ensure there were no other potential classifications. Lastly, sequences were BLASTed against the NCBI patent database to identify potential synthetic matches. If the EsVirtu-based classification was confirmed, further analysis of the sequences was performed by comparing them to known synthetic vectors and determining whether multiple classified reads were distributed across the flagged target genome. For sequences classified as poliovirus, a pipeline was built to BLAST reads against a curated database of complete poliovirus genomes (polio-blast-validation v.0) (43). Alignment of the reads was also performed to determine which genomic region they were present in. Reads were then manually inspected as described above, with only reads in the protein capsid region considered informative.

Statistical analyses

Sequence run quality was assessed using the duplication rate, rRNA sequence relative abundance, and the richness and diversity of viral species. Duplication rate was defined as the percentage of total sequencing reads that were redundant after UMI/sequence-based clustering by HUMID (see above). rRNA relative abundance was defined as the fraction of total reads mapped to ribosomal RNA sequences. The number of unique viral species (viral richness) and the Shannon diversity index were calculated using the vegan R package (44). We compared alpha diversity over time for each site using a linear regression in the lme4 package in R (45). A Bray-Curtis dissimilarity matrix was used to compare viral species composition, defined as relative abundance (number of detected reads of a specific target divided by the total number of filtered reads in the sample), among all samples. Then, non-metric multidimensional scaling (NMDS) ordinations were used to visualize results, with sample placement in the NMDS ordination colored according to WWTP and sample matrix. A PERMANOVA was used to test the null hypothesis of no significant difference in viral community composition ($P > 0.05$) between groupings (WWTP and sample matrix, tested separately). This beta diversity analysis was also performed using the vegan R package (44).

Results from hybrid capture sequencing and RT-ddPCR results were compared for a short list of viral targets: SARS-CoV-2, influenza A, influenza B, RSV (inclusive of subtypes A and B), human metapneumovirus (HMPV), rotavirus A, norovirus GII, adenovirus F, measles, hepatitis A, and West Nile virus. RT-ddPCR results were compared to results for the same viral species based on taxonomic classification and read quantification from EsVirtu; the list of species used for each viral target is listed in Table S3. The exception is Norovirus GII, where only GII subspecies were retained for comparison. We evaluated both agreement in detection across samples and the association between the abundance estimates for each sample. Agreement between hybrid capture sequencing and RT-ddPCR results was evaluated using Percent Positive Agreement (PPA), Percent Negative Agreement (PNA), and Overall Percent Agreement (OPA). PPA is the percentage of samples that tested positive for a given target by both sequencing and RT-ddPCR, and PNA is the percentage of samples that tested negative by both methods. OPA is the overall percentage of samples for which RT-ddPCR and sequencing results were concordant, either both negative or both positive. PPA, PNA, and OPA were computed across all sites and for each site individually for every viral target measured using both methods. This beta diversity analysis was also performed using the vegan R package (44).

Measurements below the limit of detection were replaced with values approximately half the estimated limit of detection (0.001 for RPKMF and 500 for gc/g), and all measurements were log₁₀-transformed prior to analysis with the lme4 package in R. The limit of detection can vary across targets, so the estimated limit of detection for RPKMF and gc/g was based on the lowest observed positive detection for any virus by each method. Random intercepts were included for city, virus, and city/date to account for these differences and repeated measures, and the intraclass correlation coefficient

(ICC) was calculated for each random effect. The performance package (46) in R was used to compute conditional and marginal R^2 . When a viral target was detected by both methods, Kendall's tau was used to assess the association between sequencing reads (measured in RPKMF) and RT-ddPCR results at each site using base R. For Kendall's tau, P -values were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate correction, but no adjustment was made for repeated measures.

RESULTS

Assessing sequencing quality of hybrid capture metagenomes

After human read removal and deduplication, a median of 4.9×10^6 unique read pairs was sequenced across the 443 samples (Fig. 1b), with a median duplication rate of 0.97 (IQR 0.97–0.98) (Tables S1 and S4). Deduplicated, de-humanized reads were then subjected to viral classification, revealing a median of 8.4×10^5 viral read pairs (19.4%) across the 443 samples, with substantial variation across sites (Fig. 3a; Table S1). Notably, samples collected from five sites—Jupiter, Lincoln, Traverse City, Las Vegas, and Anchorage—attained median viral read percentages over 25% ($N = 32, 30, 29, 31$, and 28 for the sites, respectively) (Fig. 3b).

We investigated the relationship between the total number of sequenced read pairs and the number of viral read pairs recovered, observing a positive, linear relationship between $\log_{10}$ -transformed values (Fig. S1). This positive, linear trend was particularly apparent for influent-derived samples ($R^2 = 0.73$). Similar trends were observed between the total number of read pairs sequenced and the number of unique, high-confidence viral detections per sample (Methods) after log transformation ($R^2 = 0.68$). rRNA reads were relatively abundant in sequenced read sets (median $3.6 \times 10^5/7.7\%$) despite efforts to reduce off-target nucleic acids during processing (Fig. 3c and d; Table S1).

Eleven no-template controls (NTCs) were processed to evaluate the potential for reagent and/or other sources of contamination during library preparation. Very few viral reads were sequenced in these controls (fewer than 300 viral read pairs per sample across all controls), and rRNA made up a higher proportion of reads in NTCs compared to wastewater samples (Table S5; Fig. 3c and d). However, in the cases where viral reads were observed (6/11 NTCs), they were predicted to derive from only four viruses—human mastadenovirus F, mamastrovirus 1, human papillomavirus 28, and Acheta domestica mini ambidensovirus—with read counts generally far below those observed in wastewater samples (Fig. S2). Given this, we inferred that widespread reagent contamination is unlikely to have occurred in our work, though we cannot entirely rule out the possibility of some sporadic well-to-well contamination, a common phenomenon in low-biomass metagenomic studies (47).

Sequencing runs were considered high-quality if they did not exceed a 99% duplication rate and produced $\geq 10^6$ unique reads. These thresholds were chosen based on our assessment of sample characteristics generally yielded a useful number of viral reads ($\sim 100,000$) and high quality viral detections (~ 30) (Fig. S1). Overall, 94.6% of samples (419/443) were considered high quality (420/443 with $\leq 99\%$ duplication rate, 437/443 with $\geq 10^6$ unique reads). Performance varied by site, with the percentage of high-quality samples at an individual site ranging from 29.4% (Birmingham) to 100% at 9 of the 15 sites (Table S4). One sample from Birmingham yielded no viral detections (1/26/2026, month/day/year format). Sequencing performance, measured as the number of viral reads and the duplication rate, was compared between samples derived from influent wastewater and those derived from sludge. The number of viral read pairs obtained using influent was larger than the number obtained using sludge (difference in median viral read pairs = 4.9×10^6 , Kruskal-Wallis $H = 85.8$, $P = 2.0 \times 10^{-20}$, $n = 443$). There was no difference in the duplication rate between the two sample types (Fig. S3). Birmingham sample sequencing was terminated after 17 samples on 16 February (Table S1), resulting in fewer samples in the data than other sites. The decision was made to conclude testing at Birmingham to conserve resources because samples from the site consistently failed to meet performance thresholds and showed low levels of

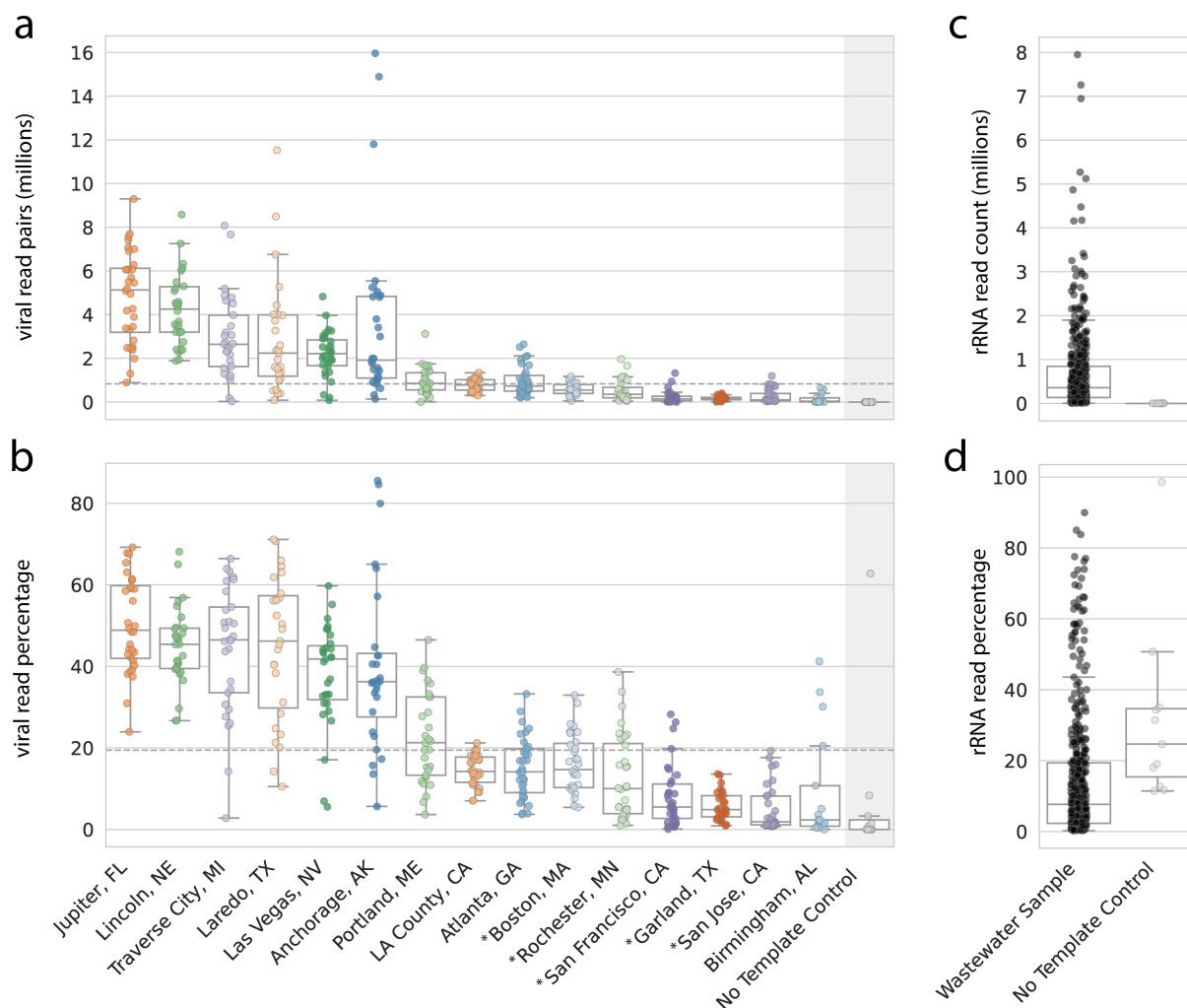


FIG 3 Components of hybrid capture metagenomes. (a) Number of viral read pairs classified per site and (b) percentage of total, deduplicated read pairs classified as viral. Dashed lines indicate the median values across all samples, excluding no-template controls. Asterisks indicate sites where sludge samples were collected. (c) Number and (d) percentage of total, deduplicated read pairs classified as rRNA by sample category. Abbreviations: NTC, no-template control.

the detection of common viruses such as norovirus despite samples passing our QA/QC process and the absence of clear laboratory errors. The data from the 17 samples that were processed are included in our analysis.

Sample composition of hybrid capture metagenomes

From the ~760 million viral read pairs obtained from 443 wastewater samples, 1,334 unique viral species and 2,227 viral subspecies were identified across 50 viral families. Across all samples, the most abundant viral families were typically *Sedoreoviridae* (rotaviruses) and *Astroviridae* (astroviruses), which together accounted for over 75% of reads in most samples. *Caliciviridae* (noroviruses) and *Adenoviridae* (adenoviruses) were the next most abundant families, accounting, respectively, for 2.0%–19.7% and 1.3%–55.1% of reads on average (Fig. S4a). In most locations, *Adenoviridae* (adenoviruses) accounted for ~2%–8% of reads across all samples; however, in certain locations (Rochester, San Francisco, and San Jose), *Adenoviridae* was much more prevalent (~30%–50%, with some samples having >90%) (Fig. S4a). Notably, all locations with increased relative abundances of *Adenoviridae* provided sludge samples. Other common human viral pathogen families, such as *Polyomaviridae* (polyomaviruses), *Coronaviridae* (coronaviruses), *Papillomaviridae* (HPV), and *Picobirnaviridae* (enteroviruses), were

identified in over 97% (434/443) of samples, while *Paramyxoviridae* (parainfluenza) and *Orthomyxoviridae* (influenza) were present in 76% (337/443) and 83% (369/443), respectively (Fig. S4b).

Across 443 samples, a median of 230 viral species (minimum: 6, maximum: 357) was found in each sample, with the most abundant species being human adenovirus F (*Human mastadenovirus F*), human astrovirus (*Mamastrovirus hominis*), and rotavirus A (*Rotavirus alphagastroenteritidis*) found in 442/443 samples. For this analysis, we refer to viruses using their common names; the corresponding species names can be found in Table S3. Common respiratory pathogens were found in a majority of samples, concordant with the winter onset periods of many of these diseases; SARS-CoV-2, influenza A, rhinovirus A and C, and respiratory syncytial virus (RSV) were found in >71% of samples (316/443) (Fig. S5). Human metapneumovirus (HMPV) was detected in 243/443 (55%) samples. During the study period, many other pathogens of clinical relevance were found, including Epstein-Barr virus, cytomegalovirus, measles, and chickenpox, in 69%, 42%, 6.3%, and 5.0% of samples, respectively. Multiple strains of human papillomavirus (HPV), both high- and low-risk, as well as hepatitis A, B, C, and E, were detected (Fig. S5). Heatmaps showing the relative abundances of 75 viral targets (highlighted in Table S3) at each sampling location can be found in the supplementary information.

Sample diversity was analyzed both temporally and spatially. Average species richness, the number of species detected in each sample, ranged from a minimum of 57 in Birmingham to a maximum of 302 in Lincoln, with a group median of 222 species. Alpha diversity, as measured via the Shannon index, ranged from 0.25 to 2.93 across all samples, with a median of 1.79 (Fig. S6a). Individual samples from multiple locations spanned a wide range of diversity values, with some falling well above or below the site median, but overall diversity distributions were broadly similar across sites when aggregated across the study period, indicating that even low-performing locations had comparable diversity yields. The slope and *P*-values of linear regressions of alpha diversity for each site were used to determine significant variation over time. Throughout the sampling period, all locations except Jupiter, Portland, and Rochester showed statistically significant decreases in alpha diversity with respect to time from January to March ($P < 0.05$ for San Francisco, $P < 0.001$ for all others) despite no significant change in species richness in any location over that period (Fig. S7). To assess variation in viral diversity among sites, Bray-Curtis dissimilarity was used to compare relative abundances of all species across samples. The NMDS 2-D ordination had a stress of 0.11, suggesting a fair fit. The null hypothesis of no variation among samples due to grouping was rejected and indicated samples of the same type (influent vs sludge) ($R^2 = 0.14$, $P < 0.001$) (Fig. S6b) and from the same WWTP ($R^2 = 0.44$, $P < 0.001$) tended to be more similar to each other than to other samples (Fig. S6c).

To explore the off-target components of the sequenced viromes, we removed reads initially classified as viral or rRNA from metagenomic samples from three sites—Las Vegas, Atlanta, and Traverse City—and performed *de novo* assembly of the remainder to create larger contiguous DNA/RNA fragments. Assembled fragments recruited a median of ~67.4% of filtered reads (Fig. S8a), and many were consistently detected across the time series at three sites. This was particularly true in Las Vegas, where ~12,000 unique fragments were detected in 75% or more of samples (Fig. S8b). Among the fragments recruiting the most reads across samples ($\geq 50,000$ reads), we detected genetic signatures of protistan ribosomal RNA, potential phage sequences, and Picornaviridae viruses that were apparently not identified by our viral classification pipeline, possibly due to poor database representation.

Relationship between quantitative sequencing and ddPCR measurements

The percentage of samples positive for a given virus was higher with RT-ddPCR than with sequencing for most targets, except for measles (Fig. 4a). When considering whether results from the two methods agreed on a per-sample basis, overall percent agreement (OPA) was typically highest for viral targets that are detected very frequently (e.g., 99.8%

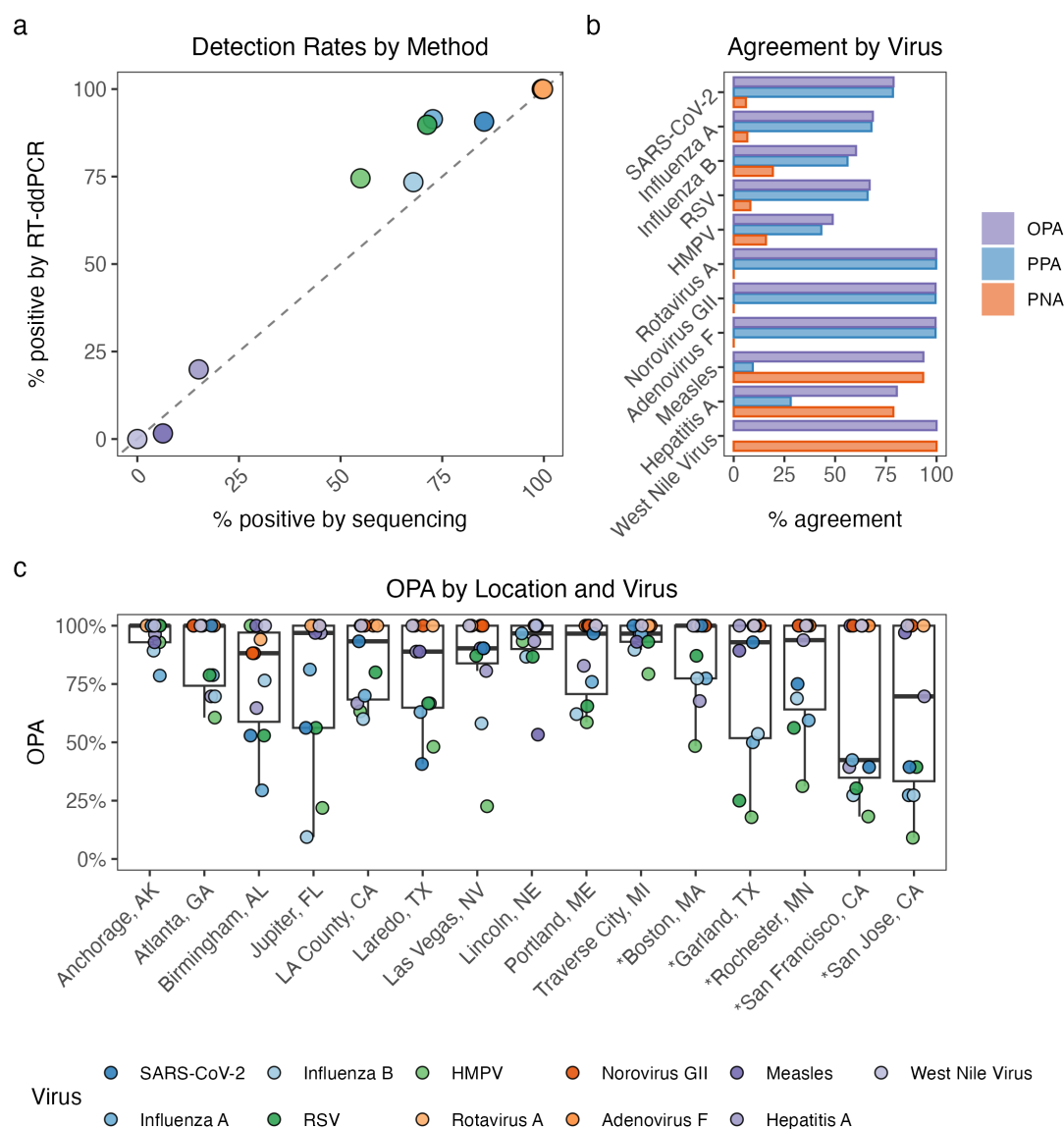


FIG 4 Detection rates and agreement between sequencing methods and RT-ddPCR. (a) Percent of samples positive by ddPCR compared to percent of samples positive by sequencing for a subset of viruses. The dashed line represents a 1:1 relationship. Norovirus, Rotavirus A, and Adenovirus F are overlapping points in the top right corner of the plot. (b) Percentage agreement (OPA, PPA, and PNA) by virus. (c) OPA by sampling location (x-axis) and by viral target (color). Individual points represent OPA for each viral target at a given sampling location, and boxplots represent the distribution of OPA across the subset of viral targets included at each site. Asterisks indicate sites where sludge samples were collected. Abbreviations: RT-ddPCR, reverse transcription digital droplet PCR; OPA, overall percent agreement; PPA, positive percent agreement; PNA, percent negative agreement.

for rotavirus) or infrequently (e.g., 93.5% for measles) and lowest for respiratory targets which were detected moderately frequently (e.g., 48.8% for HMPV). Positive Percent Agreement (PPA) was typically high for pathogens that were detected frequently (e.g., enteric viruses), and Percent Negative Agreement (PNA) was typically high for pathogens that were detected infrequently. Percent positive agreement (PPA) ranged from 9.4% for measles and 28.1% for hepatitis A to 99.5% for both adenovirus group F and norovirus GII and 99.8% for rotavirus (Fig. 4b). Percent negative agreement (PNA) ranged from 0% for norovirus GII, adenovirus group F, and rotavirus to 100% for West Nile Virus, which was never detected by either method.

Percent agreement varied by location and by viral target (Fig. 4c). OPA at individual locations varied the most for HMPV, from 9.1% in San Jose to 93.3% in Lincoln (although

OPA was 100% for HMPV in Birmingham, the virus was not detected in any sample), followed by Influenza B, with an OPA of 9.4% in Jupiter and 89.7% in Traverse City. There were no significant differences in the median PPA, PNA, or OPA values for viral targets between locations providing influent samples and those providing sludge (Wilcoxon rank-sum test, $P = 0.14$, 0.33 , and 0.27 , respectively; Table S6).

A linear mixed model revealed that target concentrations measured by RT-ddPCR were a significant positive predictor of RPKMF across all measurements (slope = 0.52 , 95% CI 0.47 – 0.57 , $t = 19.8$, $P < 2.0 \times 10^{-16}$). RT-ddPCR concentration explained 12.4% of variance (marginal R^2), and when accounting for variation across cities and viruses and repeated sample measures, the model explained 83.5% of the total variance (conditional R^2). The majority of total variance in the model was attributable to viral target (variance = 3.15 , ICC = 71.8%) with some variation by sampling location (variance = 0.41 , ICC = 9.4%). A time series of measurements at selected locations and viral targets are shown in Fig. 5, all sampling locations and viral targets are shown in Fig. S9.

When Kendall's tau was used to investigate associations within each sampling location for all viral targets considered together, there was a moderate and significant association between RPKMF and target concentrations measured by RT-ddPCR with a median Kendall's tau = 0.51 across all sites, ranging from tau = 0.42 in San Jose (adjusted $P = 1.6 \times 10^{-22}$, $n = 275$) to tau = 0.61 in Lincoln (adjusted $P = 2.5 \times 10^{-46}$, $n = 256$) (Fig. 6a; Table S6). There was a small but significant difference between tau for samples collected at locations providing influent compared to sludge (Wilcoxon rank-sum test, $P = 0.04$, median tau for influent = 0.54 , median tau for sludge = 0.50). When results were grouped by viral target across all locations, the association between the two methods was weaker, with a median of tau = 0.07 across all sites (Fig. 6b). There was no significant association across sites for RSV, measles, norovirus, or hepatitis A. Among the significant associations, tau ranged from -0.14 for HMPV (adjusted $P = 1.3 \times 10^{-4}$) to tau = 0.27 for SARS-CoV-2 (adjusted $P = 2.9 \times 10^{-16}$).

Measles sensitivity and genotyping

Measles was detected seven times throughout the sampling period via RT-ddPCR: four times in Garland, TX; one time in Lincoln, NE; one time each in Laredo, TX, and San Jose, CA (Fig. 7; Table S7). Of these seven detections, three coincided with sequencing detections. The RT-ddPCR assay used is inclusive of both wild strains of measles genotypes D8 and B3 but does not distinguish between the two (4). Measles was, therefore, chosen as an example to further explore the classification of sequencing data by genotype, using the same pipeline described above. Measles was detected via sequencing in 29 out of 443 samples across the sampling period, in 7 out of 15 locations. All alignments to measles obtained >96.6% average read identity, representing the percentage of read bases matching the reference genome, but read counts were typically low, with 18/29 samples containing <10 measles reads (median = 8, range = 1–1,070, median percent coverage = 2.7%, range = 0.8%–96.8%) (Table S7). Of these samples, 86% (24/29) produced reads that predominantly mapped to genotype D8. In three samples, measles reads were attributed to genotype A, corresponding to the measles vaccine. Two other genotypes were identified in wastewater samples: D4, which was circulating globally during the study period and was identified in a sample from a location with multiple D8 detection in other samples (1 sample; 2 reads, 99.3% identity, 0.90% coverage breadth), and C1, a genotype that is not known to be currently circulating (1 sample; 4 reads, 99.27% identity, 1.72% coverage breadth) (Fig. 7; Table S7). There was only one sample in which reads were attributable to multiple genotypes, both D8 and A.

In one sample from Lincoln, 1,070 reads were attributed to genotype D8, covering 96.8% of the reference genome, demonstrating the potential to sequence complete pathogen genomes from wastewater metagenomes (Table S7). Three detections were attributed to genotype A (two in Traverse City, one in Lincoln), which is attributed to the measles vaccine. Overall, hybrid capture sequencing was more sensitive than RT-ddPCR

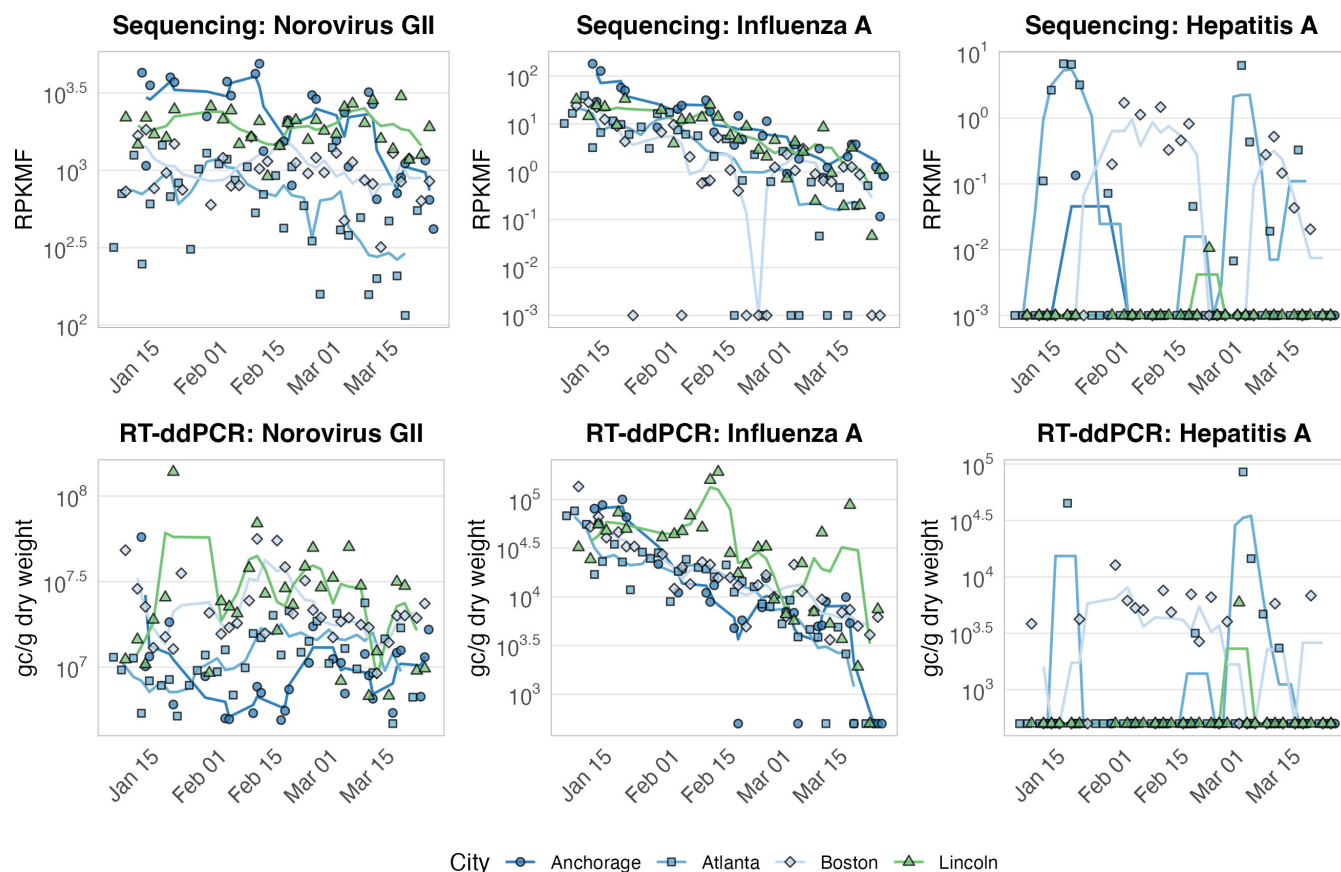


FIG 5 Time series of relative abundance of select viral targets (norovirus GII, influenza A, and hepatitis A) via sequencing (RPKM, top row) and abundance in gc/g by RT-ddPCR (bottom row) for four sites, one from each region of the United States. Points represent values for individual samples, and lines represent a 3-sample rolling average. A full time series for all targets is available in the SI. Note that y-axis scales are specific to each plot. Abbreviations: RT-ddPCR, reverse transcription digital droplet PCR; RPKM, reads per kilobase of genome per million filtered reads.

for measles detection, and enabled genotype differentiation, which is critical for public health interpretation of circulating strains.

Manual and semi-automated inspection of reads assigned to pathogens of concern

Reads assigned to pathogens of concern, including H5N1 Influenza A virus, Mpox virus, Marburg virus, poliovirus, swine vesicular disease, and avian paramyxovirus 1, using the bioinformatics pipeline (Table S8) were further investigated using additional direct validation steps as described in Materials and Methods. Using the polio direct validation pipeline, we determined that none of the read pairs classified as poliovirus represented a true detection of poliovirus; they were closer matches to non-polio enterovirus C or to conserved regions of the enterovirus C genome. Avian paramyxovirus 1 read pairs were determined not to span the fusion protein cleavage site that defines them as velogenic (48). All H5N1 read pairs were confirmed as described above. All mpox virus read pairs were confirmed to align to mpox and other *Poxviridae*. Swine vesicular disease read pairs were determined to be better matches to other enterovirus B species rather than swine vesicular disease. Lastly, the Marburg virus read pairs did not align to any known circulating Marburg virus but were sequences present only in a synthetic Marburg virus genome used in research of antibody-mediated virus neutralization. Specifically, the reads were aligned against synthetic orthomarburgvirus (GenBank MK271062.1, 100% nucleotide match, 100% query cover), reference orthomarburgvirus (RefSeq NC_024781.1, 0% nucleotide match, 0% query cover), and a recent wild-type

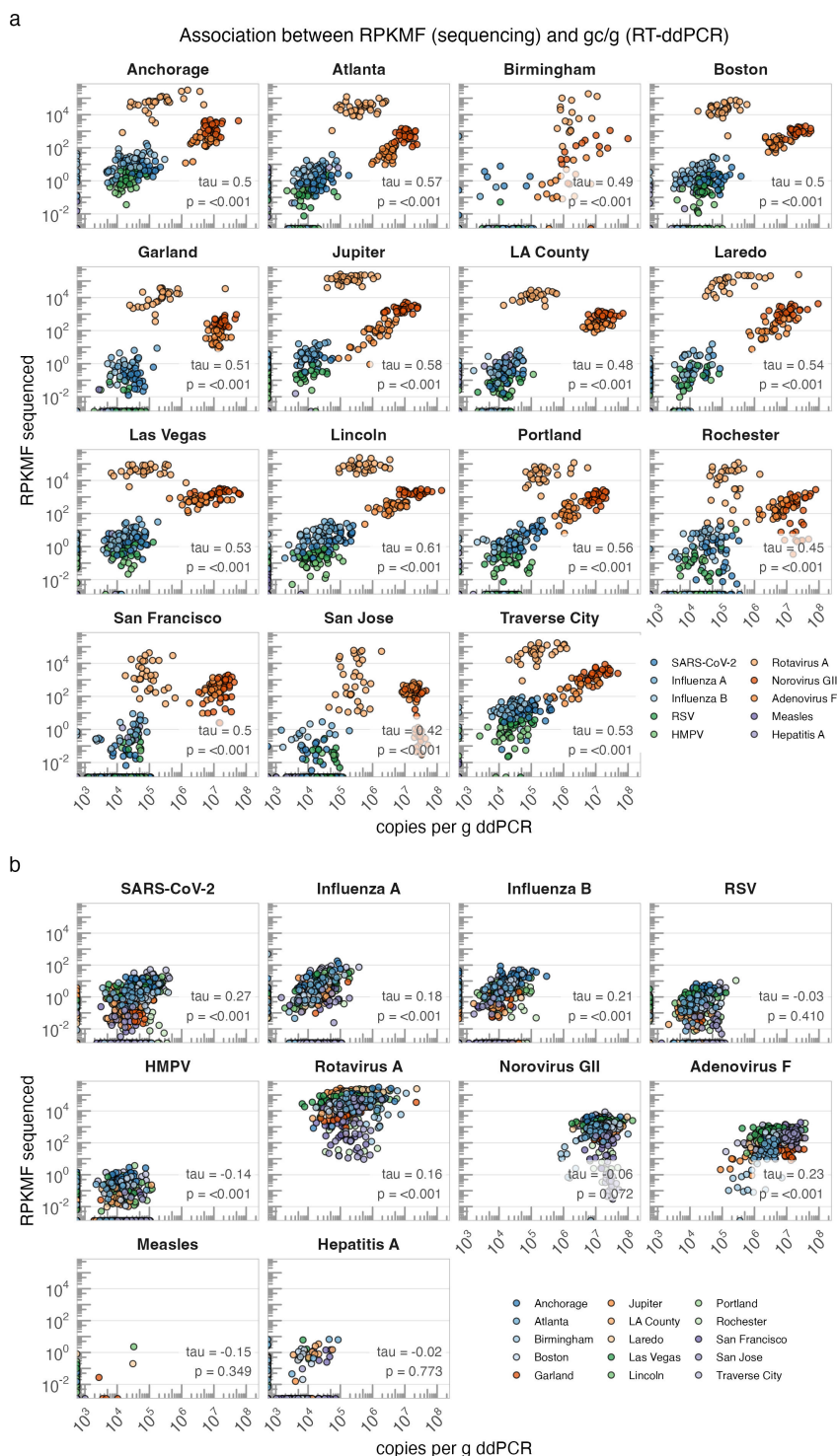


FIG 6 Relationship between relative abundance of viral targets via sequencing (RPKMF) and abundance in gc/g by RT-ddPCR for (a) each location across all viral targets with color representing different targets and (b) each viral target across all sampling locations with color representing locations. For each relationship plotted, Kendall's tau was used to assess the association between sequencing (RPKMF) and RT-ddPCR (gc/g) results at each site. *P*-values shown were adjusted for multiple comparisons using the Benjamini-Hochberg false discovery rate correction. Abbreviations: RT-ddPCR, reverse transcription digital droplet PCR; RPKMF, reads per kilobase of genome per million filtered reads.



FIG 7 Heatmap showing detections of measles by both RT-ddPCR (top row) and sequencing (bottom row) by WWTP. Color represents the genotype of measles associated with the reads from each sample. Note that different reads from the sample from Lincoln on 2/6/2026 were attributed to both genotypes D8 and A. Abbreviations: RT-ddPCR, reverse transcription digital droplet PCR; WWTP, wastewater treatment plant.

orthomarburgvirus (GenBank PV700537.1, 0% nucleotide match, 0% query cover). Subsequent detections of Marburg virus in these samples all displayed the same pattern, aligning only to the synthetic orthomarburgvirus vector. The metagenomic reads are only aligned to the synthetic orthomarburgvirus, being present in a region of the genome entirely absent from both the reference genome and the wild-type sequences (Fig. S10).

DISCUSSION

In this study, we demonstrate that hybrid capture metagenomic sequencing effectively enriches wastewater samples for sequencing of human viruses, enables sensitive, high-quality detection of rare viral targets, and produces results that are aligned with quantitative RT-ddPCR measurements of key targets across diverse U.S. sewersheds. Previous studies have demonstrated success in sequencing human viral targets in wastewater, particularly through the recovery of specific pathogen genomes such as SARS-CoV-2 using amplicon-based sequencing approaches, and through broad, unbiased approaches that have recovered portions of pathogen genomes without viral enrichment (2, 16, 32, 49). In this study, we extend this work to evaluate the performance of hybrid capture sequencing for wastewater in 15 locations across the United States with a probe panel—a semi-agnostic approach enriching for a large suite of human viruses. The goal of this approach is to allow for the broad recovery of viral genomes relevant to public health practice, while retaining more breadth of genomic information than shotgun sequencing approaches. Our findings for respiratory viruses are consistent with limited prior studies using shotgun and hybrid capture metagenomics, and we extend this comparison to a broader diversity of enteric and vaccine-preventable viruses.

Importantly, we were able to detect nearly full reference genomes for many targets, including measles.

Hybrid capture metagenomic sequencing resulted in effective enrichment of a broad set of viruses in wastewater samples. While the percent of reads that were viral varied across WWTPs, viral reads were enriched by orders of magnitude above what has been reported by Kantor et al. (range 0.1%–2.56%) and Justen et al. (median: 1.6%, range: 0%–14.3%) using shogun metagenomics (16, 23). One third of WWTPs (5/15) had a median percent viral reads higher than 25%, demonstrating consistently effective enrichment at these WWTPs. The Twist Comprehensive Viral Research Panel used in this study targets 3,153 viruses, enabling broad capture of viral diversity in wastewater samples. We detected 1,334 different viral species across 443 samples, indicating diverse viral capture.

To assess whether hybrid capture sequencing can serve as a semi-quantitative alternative to PCR methods, we compared sequencing-derived reads per kilobase of genome per million filtered reads (RPKMF) with RT-ddPCR-measured genome copies per gram (gc/g) of dry-weight solids across 11 targets from an ongoing academic wastewater monitoring project. Because the sequencing results generated for this study were compared to results from testing for an ongoing academic wastewater monitoring program, liquid was used for sequencing, and solids were used for RT-ddPCR. Despite this difference in the matrix tested, RT-ddPCR gc/g concentrations were a significant predictor of sequencing-derived RPKMF across all 15 sites, with much of the variance in the linear mixed model attributable to differences across viral targets and WWTP. Associations between sequencing-derived RPKMF and RT-ddPCR gc/g across viral pathogens within a site were consistent regardless of whether the original sample was influent or sludge, with a median Kendall's tau of 0.51 across sites. Associations were weaker for individual viral pathogens across sites and, in some cases, were not significant. While the association between measurements from sequencing and RT-ddPCR was only moderate, we note that this is likely in part because the material tested for the two methods was different. Agreement was strongest for consistently abundant enteric viruses. Although these same viruses (Rotavirus A, Norovirus GII, and Adenovirus F) had a very low percent negative agreement between the methods (0%), this was not concerning because this represented a mismatch in results among the very few samples in which these viruses were not detected. Overall, the results are consistent with the limited, previous literature on respiratory viruses, showing similar agreement between signals from shotgun metagenomics and other probe-based hybridization approaches (16, 23, 24).

While RT-ddPCR was more sensitive than sequencing for most viral targets, positivity rates were similar across methods even for rarer targets. In one case (measles), sequencing produced more detections than RT-ddPCR. This could be due to a variety of factors. For example, hybrid capture probes have a high tolerance to sequence divergence, while RT-ddPCR assays need to be highly specific. Furthermore, using sequencing, a sample can be scored positive for a virus if any portion of its genome is present, whereas for RT-ddPCR, a sample is only scored as positive if the selected portion of its genome that is targeted by the assay is present and successfully amplified. The latter may be particularly important when concentrations of the target are low, or when there is high variability or rate of change in the viral genome (4, 50).

An important advantage of hybrid capture sequencing over PCR approaches is the ability to identify subtypes or characteristics of viral genomes. Targeted PCR assays amplify a single, predefined genomic region and are less robust to single nucleotide variants and other changes in a genome. This makes these assays vulnerable to false negatives when circulating variants accumulate mismatches at primer or probe binding sites, a challenge that requires continuous assay redesign and up-to-date sequencing data, particularly for pathogens with frequent changes such as influenza and measles (8). PCR assays typically target highly conserved regions of the viral genome, and multiple assays are needed in order to detect different sub-species or genotypes depending

on which targets are important for public health goals. Hybrid capture probes theoretically tolerate up to 30% sequence divergence and interrogate multiple regions of each viral genome (51), providing built-in resilience to variant evolution and reducing the operational burden of continuous assay maintenance. This promiscuity allows for the generation of data with minimal, if any, changes in the laboratory methods that can be used to identify specific subspecies and genotypes/strains.

Sequencing yielded substantial genome coverage for several priority pathogens, including SARS-CoV-2, influenza, and RSV (52.38%, 17.65%, 13% of detections attaining $\geq 50\%$ coverage of the genome, respectively), which can enable downstream analyses such as subtyping and phylogenetic placement (52) (Fig. S11). Previous studies have shown that this enables tracking of H5N1 influenza subtypes (53). Additionally, reads distributed across the viral genome can be assembled into larger contigs or even near-complete genomes. We recently employed this *de novo* assembly approach to recover divergent fragments of the influenza A genome and link them to avian-associated lineages (52). Success of subtyping for viruses depends on variability in the viral genome, depth of sequencing coverage, and the inclusion of circulating subtypes in databases. While approaches will need to be tailored to specific viral targets, this sequence-level resolution positions hybrid capture metagenomics to greatly broaden the questions addressable by WBE.

Recovery of duplicate reads—likely from single cDNA fragments that are sequenced multiple times—in a sequencing effort is expected to be due to over-sequencing. This can occur when input cDNA has low sequence diversity, suggesting that viral enrichment did not occur as intended. In the present study, we identified samples with high-quality viral sequences as those with a duplication rate $< 99\%$. Duplication rate ranged from a median 96% to 99% across sites; sequencing was halted partway through the study at Birmingham, which had $> 99\%$ duplication rate in most samples and the lowest number of unique reads among all sites (median = 2.53×10^6). There was no significant difference in duplication rate by sample matrix; however, the number of viral reads per site did appear to differ by matrix, suggesting that influent samples may yield higher-quality data in terms of quantity of viral reads. Despite filtration and CRISPR-based rRNA depletion, prokaryotic rRNA still accounted for a median of 7.7% of reads, indicating that off-target nucleic acids remain a meaningful portion of the nucleic acids sequenced. Sequencing depth was also a key determinant of viral recovery, with both total viral reads and unique viral species scaling log-linearly with read pairs sequenced. These observations suggest that for hybrid capture WBE, sample matrix selection, upstream depletion of host and bacterial nucleic acids, and per-sample sequencing budget jointly govern the sensitivity of the assay. Future research should focus on clarifying the reasons why analysis of wastewater from sites like Birmingham may consistently fail to produce useful results with a generally successful method.

Several limitations should be considered when interpreting these results. Probe-based hybridization was chosen for this work to enable sensitive and efficient detection of human viruses in wastewater; however, enriching the sample manipulates the material such that presence and quantity of different viruses may be over or under-estimated. When enriched, relative abundance of viral targets may be altered by the efficiency of cDNA synthesis or viral capture. Viruses that are present in the sample but not targeted may not be sequenced. Another consideration for this study is that we used all mapped reads in viral classification, and did not implement a coverage cutoff or other metric to further filter detections. It is possible that implementing stricter thresholds may alter the overall estimated abundance of viruses in these samples and their relationship with other measurements such as those obtained using RT-ddPCR. Additionally, we acknowledge that this less stringent approach to coverage thresholding may result in false positive detections, particularly when reference genomes attain low breadth of coverage (percent of genome covered by reads). The identification of a rare measles genotype (C1) in a sample with only four reads and low coverage (1.72%) is an example of a subtype assignment that may be important to investigate further. However, manual

inspection of reads is costly in terms of effort and stricter computational thresholds may reduce sensitivity. In some cases, low breadth of coverage may result from mis-mapping of reads to genomic regions shared between related viruses, rather than true presence in the sample. The presence of synthetic viral vectors may also confound result interpretation. Discerning true, low-abundance detections from spurious ones will require further, careful work that balances the risk of false positives with that of false negatives, the latter of which may have serious implications for public health response. Here, we present all results to demonstrate the potential of targeted viral sequencing to provide nuanced information about both common, higher coverage pathogens as well as rarer ones, rather than a definitive accounting of viral presence/absence over our time series. Finally, viral targets referred to in this study are exclusive of prokaryotic viruses (bacteriophages), which are likely to account for larger biomass than human viruses in these samples.

It is important to note that the matrix-type comparisons of sequencing data between sites in this study compare two different sample sources rather than partitioning behavior of viruses between liquids and solids. Metagenomic sequencing was performed on the liquid fraction of both sample types. Sludge-derived samples were centrifuged prior to processing to remove excess solids, whereas influent samples were processed without removal of solids. Future studies that include direct comparisons of sludge and liquid-derived samples from the same sites would also provide evidence for the best sample types for hybrid capture sequencing. The discrepancy in the sample material for the comparison between RT-ddPCR measurements and sequencing is another limitation of this study. Future work could pursue a comparison of the two methods using the same nucleic acid extracts and/or pursuing an experiment recovering spiked in viral material from the solids compared to liquids in a sample. More research on normalizing results (e.g., to another biomarker such as a widely abundant viral target) may further help with comparison of quantitative results between PCR and sequencing data.

In summary, this study illustrates that hybrid capture sequencing can be scaled to generate information on presence and quantity of human viruses in wastewater from WWTPs across the United States, each with unique characteristics and population coverage, to infer information on infectious diseases circulating in the contributing populations. A remaining challenge is to understand how to best prioritize public health actionable data within the large amounts of information obtained with this approach, and how to move beyond tracking presence and abundance of targets over time. The goal of the present study was to establish the scientific basis for this sequencing approach so that future translational work can explore how the data can be used to protect public health. We anticipate that the data can be used to identify emerging viral infections and outbreaks, for situational awareness, and various response activities as is done with PCR-based data now. To turn this process into a useful tool for public health, it is a priority to enable quick turnaround of both wet lab and analytical pipelines, confirming assignments where necessary, and flagging results that are associated with important health outcomes. The scientific community should work closely with public health officials and clinicians to understand how to prioritize and communicate outcomes from these data and explore how subtyping and characterization of viral genomes may best serve the needs of the community.

ACKNOWLEDGMENTS

We thank the wastewater treatment plant partners at A.K. Warren Water Resource Recovery Facility, City of Garland Rowlett Creek Wastewater Treatment Plant, City of Rochester MN Water Reclamation Plant, Clark County Water Reclamation District Flamingo Water Resource Center, Deer Island Treatment Plant, John M. Asplund Water Pollution Control Facility, Loxahatchee River Environmental Control District, Portland Water District East End Wastewater Treatment Facility, RM Clayton Water Reclamation Center, San Jose-Santa Clara Regional Wastewater Facility, Southeast San Francisco Wastewater Treatment Plant, Theresa Street Water Resource Recovery Facility, Traverse City Regional Waste Water Treatment Plant, Village Creek Water Reclamation Facility,

and Zacate Creek Wastewater Treatment Plant for assistance with sample collection. We thank the local and state public health department partners in Alabama, Alaska, California, Georgia, Florida, Maine, Massachusetts, Michigan, Minnesota, Nebraska, Nevada, and Texas for their support of this work.

This work was funded by gifts to M.K.W. and A.B.B. from the Sergey Brin Family Foundation.

AUTHOR AFFILIATIONS

¹Gangarosa Department of Environmental Health, Emory University, Atlanta, Georgia, USA

²Department of Civil and Environmental Engineering, Stanford University, Stanford, California, USA

³Verily Health, South San Francisco, California, USA

AUTHOR ORCIDs

Marlene K. Wolfe  <http://orcid.org/0000-0002-6476-0450>

Devin North  <http://orcid.org/0000-0002-0671-3381>

Alexander L. Jaffe  <http://orcid.org/0000-0002-6903-1687>

Alexandria B. Boehm  <http://orcid.org/0000-0002-8162-5090>

FUNDING

Funder	Grant(s)	Author(s)
Sergey Brin Family Foundation		Marlene K. Wolfe
		Alexandria B. Boehm

AUTHOR CONTRIBUTIONS

Marlene K. Wolfe, Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review and editing | Devin North, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing | Alexander L. Jaffe, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing | Alessandro Zulli, Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review and editing | Dorothea Duong, Data curation, Investigation, Methodology, Project administration, Supervision, Writing – review and editing | Bridgette Shelden, Data curation, Investigation, Methodology, Project administration, Supervision, Writing – review and editing | Miriam Goldman, Data curation, Formal analysis, Investigation, Methodology, Writing – review and editing | Miles Richardson, Data curation, Formal analysis, Investigation, Methodology, Writing – review and editing | Peter Thana, Data curation, Methodology, Project administration, Supervision, Writing – review and editing | Vikram Chan-Herur, Data curation, Investigation, Methodology, Writing – review and editing | Pouya Kheradpour, Data curation, Formal analysis, Methodology, Project administration, Supervision, Writing – review and editing | Amanda L. Bidwell, Data curation, Project administration, Supervision, Visualization, Writing – original draft, Writing – review and editing | Stephen P. Hilton, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review and editing | Sheena Conforti, Methodology, Visualization, Writing – original draft, Writing – review and editing | Abigail P. Paulos, Methodology, Writing – original draft, Writing – review and editing | Alexandria B. Boehm, Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

Deduplicated sequencing reads are available through the NCBI under BioProject [PRJNA1438722](https://doi.org/10.25740/cx707zg0146). Accessions are also listed in Table S1. Digital PCR data are available at <https://doi.org/10.25740/cx707zg0146>.

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

Supplemental material (AEM01020-26-S0001.docx). Supplemental methods; Fig. S1 to S11.

Supplemental heatmaps (AEM01020-26-S0002.pdf). Showing common viruses for each WWTP.

Supplemental tables (AEM01020-26-S0003.xlsx). Tables S1 to S8.

REFERENCES

- Wolfe MK, Paulos AH, Zulli A, Duong D, Shelden B, White BJ, Boehm AB. 2024. Wastewater detection of emerging arbovirus infections: case study of dengue in the United States. *Environ Sci Technol Lett* 11:9–15. <https://doi.org/10.1021/acs.estlett.3c00769>
- Peccia J, Zulli A, Brackney DE, Grubaugh ND, Kaplan EH, Casanovas-Massana A, Ko AI, Malik AA, Wang D, Wang M, Warren JL, Weinberger DM, Arnold W, Omer SB. 2020. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nat Biotechnol* 38:1164–1167. <https://doi.org/10.1038/s41587-020-0684-z>
- Boehm AB, Wolfe MK, Bidwell AL, White BJ, Shelden B, Duong D. 2025. Detection of monkeypox virus clade Ib DNA in wastewater solids at wastewater treatment plants, United States. *Emerg Infect Dis* 31:1984–1987. <https://doi.org/10.3201/eid3110.250270>
- Paulos AP, Zulli A, Bidwell A, Hilton SP, Shelden B, Duong D, Boehm AB, Wolfe MK. 2026. Measles RNA detection in wastewater solids: longitudinal monitoring at a national scale. *Environ Sci: Water Res Technol* 12:2074–2081. <https://doi.org/10.1039/D6EW00020G>
2026. Wastewater surveillance. Available from: <https://wise.ethz.ch/?page=overview>
- CDC. 2026. CDC's wastewater monitoring program. Available from: <https://www.cdc.gov/wastewater/index.html>
- Naughton CC, Holm RH, Lin NJ, James BP, Smith T. 2023. Online dashboards for SARS-CoV-2 wastewater data need standard best practices: an environmental health communication agenda. *J Water Health* 21:615–624. <https://doi.org/10.2166/wh.2023.312>
- Boehm AB, Wolfe MK, Bidwell AL, Zulli A, White BJ, Shelden B, Duong D. 2026. Pathogen nucleic acids data in wastewater solids from 147 treatment plants in the United States: 2024–2025. *Data Brief* 65:112503. <https://doi.org/10.1016/j.dib.2026.112503>
- Xue KS, Stevens-Ayers T, Campbell AP, Englund JA, Pergam SA, Boeckh M, Bloom JD. 2017. Parallel evolution of influenza across multiple spatiotemporal scales. *eLife* 6:e26875. <https://doi.org/10.7554/eLife.26875>
- Boehm AB, Wolfe MK, Bidwell AL, Zulli A, Chan-Herur V, White BJ, Shelden B, Duong D. 2024. Human pathogen nucleic acids in wastewater solids from 191 wastewater treatment plants in the United States. *Sci Data* 11:1141. <https://doi.org/10.1038/s41597-024-03969-8>
- Jiang M, Wang ALW, Be NA, Mulakken N, Nelson KL, Kantor RS. 2024. Evaluation of the impact of concentration and extraction methods on the targeted sequencing of human viruses from wastewater. *Environ Sci Technol* 58:8239–8250. <https://doi.org/10.1021/acs.est.4c00580>
- Zulli A, White B, Duong D, Shelden B, Chan-Herur V, Wolfe M, Boehm AB. 2026. Nationwide wastewater sequencing surveillance of SARS-CoV-2 lineages: validation against clinical data across 28 U.S. states. *PeerJ* 14:e20941. <https://doi.org/10.7717/peerj.20941>
- Shakya M, Ma KC, Hughes LJ, Smith C, Atherton LJ, Boehm AB, Cook PW, Cornforth DM, Gardner M, Goraichuk IV, Harcourt JL, Hicks H, Johnson M, Kelleher A, Reese HE, Sabin SJ, Smith TC, Tamin A, Wolfe MK, Silk BJ, Paden CR, Thornburg N, MacNeil A. 2024. Early detection and surveillance of the SARS-CoV-2 variant BA.3.2 — worldwide, November 2024–February 2026. *MMWR Morb Mortal Wkly Rep* 75:130–137. <https://doi.org/10.15585/mmwr.mm7510a1>
- John A, Kang S, Fuhrmann L, Topolsky I, Kent C, Quick J, Stadler T, Julian TR, Beerenwinkel N. 2025. Characterizing influenza A virus lineages and clinically relevant mutations through high-coverage wastewater sequencing. *Water Res* 287:124453. <https://doi.org/10.1016/j.watres.2025.124453>
- Child HT, Airey G, Maloney DM, Parker A, Wild J, McGinley S, Evens N, Porter J, Templeton K, Paterson S, van Aerle R, Wade MJ, Jeffries AR, Bassano I. 2023. Comparison of metagenomic and targeted methods for sequencing human pathogenic viruses from wastewater. *mBio* 14:e0146823. <https://doi.org/10.1128/mbio.01468-23>
- Justen LJ, Rushford C, Hershey OS, Floyd-O'Sullivan R, Grimm SL, Bradshaw WJ, Bhasin H, Rice DP, Stansifer K, Faraguna JD, et al. 2026. Deep untargeted wastewater metagenomic sequencing from sewersheds across the United States. *medRxiv*. <https://doi.org/10.64898/2026.03.05.26345726>
- Newton RJ, McLellan SL, Dila DK, Vineis JH, Morrison HG, Eren AM, Sogin ML. 2015. Sewage reflects the microbiomes of human populations. *mBio* 6:e02574. <https://doi.org/10.1128/mBio.02574-14>
- Grimm SL, Kaufman JT, Rice DP, Whittaker C, Bradshaw WJ, McLaren MR. 2025. Inferring the sensitivity of wastewater metagenomic sequencing for early detection of viruses: a statistical modelling study. *Lancet Microbe* 6:101187. <https://doi.org/10.1016/j.lanmic.2025.101187>
- Rushford C, Gregory D, Copen E, Naydenov J, Tovar-Mendez A, Li P-E, Kantor RS, Shakya M, Ruth N, Chain PSG, Hsieh H, Hunter T, Kome J, Frank M, Lyddon T, Kaufman J, O'Connor DH, Johnson MC. 2025. Untargeted longitudinal ultra deep metagenomic sequencing of wastewater provides a comprehensive readout of expected and unexpected viral pathogens. *medRxiv*. <https://doi.org/10.1101/2025.10.27.25338874>
- Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27:182–189. <https://doi.org/10.1038/nbt.1523>
2020. *Comprehensive Viral Research Panel*. Available from: <https://twistbioscience.com/products/ngs/fixed-panels/comprehensive-viral-research-panel>
- Alanko JN, Slizovskiy IB, Lokshantov D, Gagie T, Noyes NR, Boucher C. 2022. Syotti: scalable bait design for DNA enrichment. *Bioinformatics* 38:i177–i184. <https://doi.org/10.1093/bioinformatics/btac226>
- Kantor RS, Jiang M. 2024. Considerations and opportunities for probe capture enrichment sequencing of emerging viruses from wastewater. *Environ Sci Technol* 58:8161–8168. <https://doi.org/10.1021/acs.est.4c02638>
- Zulli A, Bowie KR, Ruedaflares M, Sahly J, Angel D, Rohatgi K, Malik WS, Hao RM, Shepherd J, Peccia J. 2025. Probe-based enrichment sequencing applied to wastewater surveillance accurately tracks multiple viral

- respiratory outbreaks. *ACS EST Water* 5:4575–4583. <https://doi.org/10.1021/acsestwater.5c00262>
25. Roldan-Hernandez L, Van Oost C, Boehm AB. 2024. Solid-liquid partitioning of dengue, west nile, zika, hepatitis A, influenza A, and SARS-CoV-2 viruses in wastewater from across the USA. *Environ Sci: Water Res Technol* 11:88–99. <https://doi.org/10.1039/D4EW00225C>
 26. Roldan-Hernandez L, Boehm AB. 2023. Adsorption of respiratory syncytial virus, rhinovirus, SARS-CoV-2, and F+ bacteriophage MS2 RNA onto wastewater solids from raw wastewater. *Environ Sci Technol* 57:13346–13355. <https://doi.org/10.1021/acs.est.3c03376>
 27. Zulli A, Duong D, Hughes B, Chan-Herur V, Catoe D, Jaffe A, Paulos A, Bidwell A, Wolfe M, Boehm A. 2026. WWSCAN probe-based hybridization of wastewater for metagenomic analysis of viruses. Available from: <https://www.protocols.io/view/wwscan-probe-based-hybridization-of-wastewater-for-e6nvw41w7lmk/v1>
 28. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20:257. <https://doi.org/10.1186/s13059-019-1891-0>
 29. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
 30. Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. 2021. STAT: a fast, scalable, MinHash-based k-mer tool to assess sequence read archive next-generation sequence submissions. *Genome Biol* 22:270. <https://doi.org/10.1186/s13059-021-02490-0>
 31. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
 32. Tisza M, Javornik Cregeen S, Avadhanula V, Zhang P, Ayvaz T, Feliz K, Hoffman KL, Clark JR, Terwilliger A, Ross MC, et al. 2023. Wastewater sequencing reveals community and variant dynamics of the collective human virome. *Nat Commun* 14:6878. <https://doi.org/10.1038/s41467-023-42064-1>
 33. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. 2012 *Bioinformatics*. Oxford Academic. Available from: <https://academic.oup.com/bioinformatics/article/28/24/3211/246053?log-in=false>
 34. Using SPAdes *de novo* assembler - Prjibelski. 2020. Current protocols in bioinformatics. Wiley Online Library. Available from: <https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.102>
 35. Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 35:1026–1028. <https://doi.org/10.1038/nbt.3988>
 36. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10:giab008. <https://doi.org/10.1093/gigascience/giab008>
 37. Aroney STN, Newell RJP, Nissen JN, Camargo AP, Tyson GW, Woodcroft BJ. 2025. CoverM: read alignment statistics for metagenomics. *Bioinformatics* 41:btaf147. <https://doi.org/10.1093/bioinformatics/btaf147>
 38. Waskom ML. 2021. Seaborn: statistical data visualization. *J Open Source Softw* 6:3021. <https://doi.org/10.21105/joss.03021>
 39. Wickham H. 2011. ggplot2. *WIREs Comput Stat* 3:180–185. <https://doi.org/10.1002/wics.147>
 40. Select agents and toxins list. 2025. Federal Select Agent Program. Available from: <https://www.selectagents.gov/sat/list.htm>
 41. Staff N. 2024. Get faster, more focused search results with NCBI's new BLAST core nucleotide database (core_nt). NCBI Insights. Available from: <https://ncbiinsights.ncbi.nlm.nih.gov/2024/07/18/new-blast-core-nucleotide-database/>
 42. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. <https://doi.org/10.1186/1471-2105-10-421>
 43. AZulli946. 2026. AZulli946/polio-blast-validation
 44. Oksanen J, Simpson G, Blanchet F, et al. 2026. vegan: an R package for community ecologists. <https://doi.org/10.32614/CRAN.package.vegan>
 45. Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J Stat Softw* 67. <https://doi.org/10.18637/jss.v067.i01>
 46. Lüdtke D, Ben-Shachar MS, Patil I, Waggoner P, Makowski D. 2021. performance: an R package for assessment, comparison and testing of statistical models. *J Open Source Softw* 6:3139. <https://doi.org/10.21105/joss.03139>
 47. Lou YC, Hoff J, Olm MR, West-Roberts J, Diamond S, Firek BA, Morowitz MJ, Banfield JF. 2023. Using strain-resolved analysis to identify contamination in metagenomics data. *Microbiome* 11:36. <https://doi.org/10.1186/s40168-023-01477-2>
 48. Hines NL, Miller CL. 2012. Avian paramyxovirus serotype-1: a review of disease distribution, clinical symptoms, and laboratory diagnostics. *Vet Med Int* 2012:708216. <https://doi.org/10.1155/2012/708216>
 49. Poretsky RS, Dhiman VK, Hendricks DL, Lin C-Y, Sanchez Gonzalez D, Greenwald S, Owens SM, Williams CH, Leslie MT, Bemis K, Frias M, Kaufman JT, O'Connor DH, Johnson MC. 2026. Detection of a single measles infection using untargeted ultra-deep metagenomic sequencing of wastewater in Cook County, Illinois. *NEJM Evid* 5:EVIDpha2600079. <https://doi.org/10.1056/EVIDpha2600079>
 50. CDC. 2025. CDC's influenza SARS-CoV-2 multiplex assay. Influenza (Flu). Available from: <https://www.cdc.gov/flu/php/laboratories/influenza-sars-cov-2-multiplex-assay.html>
 51. 2021. *Novel Virus Detection Using the Twist Comprehensive Viral Research Panel*. Available from: <https://twistbioscience.com/resources/application-note/novel-virus-detection-using-twist-comprehensive-viral-research-panel>
 52. Jaffe AL, Zulli A, Duong D, Sheldon B, Goldman M, Richardson M, Wolfe MK, Boehm AB. 2026. Divergent avian strains drive an off-season influenza A peak in municipal wastewater 14:e02027-26. <https://doi.org/10.1128/spectrum.02027-26>
 53. Tisza MJ, Hanson BM, Clark JR, Wang L, Payne K, Ross MC, Mena KD, Gitter A, Javornik Cregeen SJ, Cormier J, Avadhanula V, Terwilliger A, Balliew J, Wu F, Rios J, Deegan J, Piedra PA, Petrosino JF, Boerwinkle E, Maresso AW. 2024. Sequencing-based detection of avian influenza A(H5N1) virus in wastewater in ten cities. *N Engl J Med* 391:1157–1159. <https://doi.org/10.1056/NEJMc2405937>