Deep untargeted wastewater metagenomic sequencing from sewersheds across the United States

Lennart J. Justen*[1, 2, 3], Clayton Rushford*[4], Olivia S. Hershey[5], Róisín Floyd-O'Sullivan[5], Simon L. Grimm[5], William J. Bradshaw[5], Harmon Bhasin[5], Daniel P. Rice[5], Katherine Stansifer[5], Jo D. Faraguna[5], Michael R. McLaren[5], Alessandro Zulli[5], Alejandro Tovar-Mendez[6], Emma E. Copen[6], Kristen K. Shelton[7], Ayaaz Amirali[8], Sherin Kannoly[9], Sofia Pesantez[9], Aiden Stanciu[9], Iñigo Caballero Quiroga[9], Leopolda Silvera[10], Nicole Greenwood[11], Barbra Bongiovi[12], Austin Walkins[12], Ryan Love[13], Scott Lening[13], Kaylyn Patterson[14], Theresa Johnston[14], Sandra Hernandez[15], Aymara Benitez[15], Billie Jo McCarley[15], Samantha Engelage[16], Suguna Pillay[16], Cindy Calender[17], Brent Herring[18], Carey Robinson[19], Monett Wastewater Treatment Plant, Columbia Missouri Wastewater Treatment Plant, Daniel Cunningham-Bryant[20], Gordon Adams[3,21], Jillian Paull[3,22], Jamie Devlin[21], Vamsi Thiriveedhi[23], Sarah E. Turbett[21,23,24], Jacob E. Lemieux[3,21,24], Rose S. Kantor[25], David H. O'Connor[26], John J. Dennehy[9,27], Rachel Poretsky[28], Jason A. Rothman[29], Helena M. Solo-Gabriele[8], Jason R. Vogel[7], Pardis C. Sabeti[3,22,30,31], Jeff Kaufman[†5], Marc Johnson[†6]

*These authors contributed equally to this work.
[†]Corresponding authors: jeff@securebio.org, marcjohnson@health.missouri.edu

Affiliations
[1]Massachusetts Institute of Technology, Media Lab, Cambridge, MA, USA
[2]The Charles Stark Draper Laboratory, Inc., Draper Scholar, Cambridge, MA, USA
[3]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[4]University of Missouri, Department of Veterinary Pathobiology, Columbia, MO, USA
[5]SecureBio, Cambridge, MA, USA
[6]University of Missouri, Department of Molecular Microbiology and Immunology, Columbia, MO, USA
[7]University of Oklahoma, School of Civil Engineering and Environmental Science, Norman, OK, USA
[8]University of Miami, Department of Chemical, Environmental, and Materials Engineering, Miami, FL, USA
[9]Queens College of The City University of New York, Biology Department, Flushing, NY, USA
[10]NYC Health + Hospitals, New York, NY, USA
[11]Riverside Water Quality Control Plant, Riverside, CA, USA
[12]City of Boise, Boise, ID, USA
[13]Inland Empire Utilities Agency, Chino, CA, USA
[14]Metropolitan Water Reclamation District of Greater Chicago, Chicago, IL, USA
[15]Miami-Dade Water and Sewer Department, Miami, FL, USA
[16]City of Palo Alto, Palo Alto, CA, USA
[17]City of Kansas City Missouri, KC Health, Communicable Disease and Prevention Division, Kansas City, MO
[18]City of Kansas City Missouri, KC Water, Water and Wastewater Operations, Kansas City, MO
[19]City of Kansas City Missouri, KC Water, Wastewater Treatment Division, Kansas City, MO
[20]The Charles Stark Draper Laboratory, Inc., Cambridge, MA, USA

[21]Massachusetts General Hospital, Division of Infectious Diseases, Boston, MA, USA
[22]Howard Hughes Medical Institute, Chevy Chase, MD, USA
[23]Massachusetts General Hospital, Department of Pathology, Boston, MA, USA
[24]Harvard Medical School, Boston, MA, USA
[25]Lawrence Livermore National Laboratory, Physical and Life Sciences Directorate, Livermore, CA, USA
[26]University of Wisconsin-Madison, Department of Pathology and Laboratory Medicine, Madison, WI, USA
[27]The Graduate Center of The City University of New York, New York, NY, USA
[28]University of Illinois Chicago, Department of Biological Sciences, Chicago, IL, USA
[29]University of California, Riverside, Department of Microbiology and Plant Pathology, Riverside, CA, USA
[30]Harvard University, Harvard T.H. Chan School of Public Health, Department of Immunology and Infectious Diseases, Boston, MA, USA
[31]Harvard University, Faculty of Arts and Sciences, Department of Organismic and Evolutionary Biology, Cambridge, MA, USA

Abstract

Wastewater monitoring enables non-invasive, population-scale tracking of community infections independent of healthcare-seeking behavior and clinical diagnosis. Metagenomic sequencing extends this capability by enabling broad, pathogen-agnostic detection, genomic characterization, and identification of novel or unexpected threats. Here, we present data from CASPER (the Coalition for Agnostic Sequencing of Pathogens from Environmental Reservoirs), a U.S.-based wastewater metagenomic sequencing network designed for deep, untargeted pathogen monitoring at national scale. This release includes 1,206 samples collected between December 2023 and December 2025 from 27 sites across nine states, covering 13 million people. Deep sequencing (~1 billion read pairs per sample) generated 1.2 trillion read pairs (347 terabases), enabling detection of even rare taxa, with CASPER representing 66% of all untargeted wastewater sequencing data currently available on the NCBI Sequence Read Archive. Virus abundance trends correlate with nationwide wastewater PCR and clinical data for SARS-CoV-2, influenza A, and respiratory syncytial virus, while the pathogen-agnostic approach captures emerging threats, including avian influenza H5N1 during initial dairy cattle outbreaks, West Nile virus, and measles, among hundreds of viral taxa. As the largest publicly available untargeted wastewater sequencing dataset to date, CASPER provides a shared and growing resource for pathogen surveillance and microbial ecology.

Main

Wastewater monitoring has emerged as a powerful tool for providing early warning of pathogen outbreaks and tracking community-level infections independent of health care-seeking behavior (Kirby et al. 2022; Diamond et al. 2022; Hellmér et al. 2014; Peccia et al. 2020; Karthikeyan et al. 2022; Rebecca Falender et al. 2026). A single wastewater sample can capture pathogens shed by millions of individuals, including those with asymptomatic or pre-symptomatic infections, providing a comparatively unbiased snapshot of pathogen circulation at scale (Wolfe et al. 2021; Corrin et al. 2024; Committee on Community Wastewater-based Infectious Disease Surveillance et al. 2024).

In practice, current wastewater monitoring programs rely on targeted detection approaches such as quantitative PCR (qPCR), digital PCR (dPCR), or amplicon sequencing to detect predefined pathogen targets (Adams, Bias, et al. 2024; Boehm et al. 2024; Duvallet et al. 2022; Morfino et al. 2023). These approaches are sensitive, cost-effective, and valuable for public health, but face limitations: they cannot detect novel pathogens outside their targets, may lose sensitivity as mutations occur in primer binding sites (Lambisia et al. 2022; Schuele et al. 2024; Public Health England 2020; Kidd et al. 2021; Wu et al. 2023), and require sometimes lengthy assay design and validation for each new target, often during critical periods of an outbreak (Adams, Kirby, et al. 2024; Wolfe et al. 2022). Hybrid-capture sequencing is a promising semi-targeted alternative detection platform, using probe panels to enrich for sequences similar to known pathogens and tolerating greater sequence divergence (20–30%) than amplicon-based methods (Tisza et al. 2023; Kantor and Jiang 2024; Zulli et al. 2025).

Unlike other approaches, wastewater metagenomic sequencing (WW-MGS) resolves the entire microbial community present, enabling untargeted, pathogen-agnostic detection of known, novel, and unexpected pathogens simultaneously (Wongprommoon et al. 2024). WW-MGS has already demonstrated its utility for large-scale microbial ecology surveys (Fiamenghi et al. 2025; Edgar et al. 2022; Neri et al. 2022; Worp et al. 2025) and has shown the ability to detect a broad range of human and animal pathogens (Rushford et al. 2025; Spurbeck et al. 2023; Rothman et al. 2021), including unexpected signals with public health relevance such as highly pathogenic avian influenza H5N1 (Rushford et al. 2025), West Nile virus (Missouri DHSS 2025), and measles virus (Hawaii DOH 2025).

Despite this potential, routine untargeted WW-MGS has seen limited adoption as a biosecurity or public health tool. Vertebrate-infecting viruses constitute a small fraction of total nucleic acids in wastewater, requiring deep and costly sequencing for sensitive detection (Grimm, Kaufman, et al. 2025); computational infrastructure must scale to hundreds of terabases; and interpretation of detections is challenged by the diverse sources of nucleic acids in wastewater (Grassly et al. 2025). However, continued declines in sequencing costs and the expansion of wastewater monitoring infrastructure have made routine deep, untargeted WW-MGS increasingly feasible.

We established CASPER (Coalition for Agnostic Sequencing of Pathogens from Environmental Reservoirs) to enable deep, untargeted wastewater metagenomic sequencing at national scale. CASPER is a collaboration of U.S.-based partners, including academic institutions, national laboratories, municipal wastewater utilities, and hospitals, advancing pathogen monitoring through untargeted sequencing. Launched through a partnership between SecureBio (SB) and the University of Missouri (MU), the network has since expanded across the country and continues to grow, with data released periodically on the NCBI Sequence Read Archive (SRA). Sequencing data support real-time dashboards ("Wastewater Surveillance Dashboards," n.d.), and detections of immediate concern are reported to relevant public health and biosecurity authorities. CASPER targets RNA, capturing an important diversity of plant-, bacteria-, and vertebrate-infecting viruses, including likely sources of future human-infecting pandemic pathogens (Adalja et al. 2018), though some DNA and DNA viruses are also captured. To enable sensitive detection of rare taxa, we sequence to a depth of approximately one billion read pairs per sample—far deeper than most previous efforts.

Here, we report the public release of CASPER data from 27 sites approved for sharing. The release comprises 1.2 trillion read pairs (347 terabases (Tb)) from samples collected between December 2023 and December 2025, spanning nine U.S. states and covering over 13 million people. A previous, smaller CASPER release comprising 13 Tb from a Los Angeles pilot study is also publicly available (Grimm, Rothman, et al. 2025; BioProject PRJNA1198001). Together, CASPER constitutes the largest public repository of WW-MGS data to date, accounting for 66% of all untargeted wastewater sequencing data on the NCBI Sequence Read Archive (SRA). Beyond pathogen surveillance, this data provides a shared resource for research in microbial ecology, viral discovery, antimicrobial resistance, and methodological development. All data is publicly available (BioProject PRJNA1247874). We invite the research and public health communities to use and build upon this growing resource.

Results

Sampling collection, processing, and sequencing

CASPER sampling sites in this data release span nine U.S. states and encompass diverse geographic regions and major population centers (Figure 1a). This release includes 27 distinct wastewater catchments (Table 1). Using sewershed population estimates from the National Wastewater Surveillance System (NWSS) (CDC 2025c) and other sources, we estimate that these sites collectively serve approximately 13 million people (Table 1). This figure counts each physical catchment once and excludes hospital and airport sites, which lack reliable population estimates, as well as one site that opted not to disclose.

Table 1. Summary of CASPER sampling sites included in this data release. The table lists site location, estimated population served, sample type, collection start date, and samples in this data release. Population estimates are approximate and based on sewershed coverage data from the National Wastewater Surveillance System (NWSS) where available. "n/a" indicates non-wastewater treatment plant (WWTP) sites for which contributing population estimates are difficult to determine, and "–" indicates sites that opted to keep population estimates anonymous. MU, University of Missouri; SB, SecureBio.

Sampling site	Location type	Sample type	Population served	Sequencing parter	Collection start date	Samples
Northeast						
Boston Deer Island Treatment Plant (DITP) North, MA	WWTP	24h composite	1.1M	MU	2024-03-20	65
Boston DITP North, MA	WWTP	24h composite	1.1M	SB	2025-02-12	44
Boston DITP South, MA	WWTP	24h composite	1.4M	MU	2024-03-20	65
Boston DITP South, MA	WWTP	24h composite	1.4M	SB	2025-02-12	44
New York (Hospital A), NY	Hospital (service hatch)	Grab	n/a	MU	2025-08-28	10
New York (Hospital B), NY	Hospital (manhole)	Grab	n/a	MU	2024-01-16	23
New York (Hospital C), NY	Hospital (manhole)	Grab	n/a	MU	2024-11-13	8
New York (Hospital D), NY	Hospital (manhole)	Grab	n/a	MU	2025-09-02	16
Midwest						
Chicago (CHI-A), IL	WWTP	24h composite	1.3M	MU	2024-09-08	69
Chicago (CHI-B), IL	WWTP	24h composite	1.1M	MU	2024-09-08	63
Chicago (CHI-C), IL	Airport (manhole)	Moore swab	n/a	MU	2024-10-14	52
Chicago (CHI-D1), IL	WWTP	24h composite	1.1M	MU	2024-09-15	63
Chicago (CHI-D2), IL	WWTP	24h composite	1.1M	MU	2024-09-15	63
Chicago (CHI-D3), IL	WWTP	24h composite	2.2M	MU	2024-09-08	6
Ottumwa Water Pollution Control Facility (WPCF), IA	WWTP	24h composite	26k	MU	2025-02-19	46
Columbia WWTP, MO	WWTP	24h composite	120k	MU	2023-12-26	106
Kansas City Blue River WWTP, MO	WWTP	24h composite	490k	MU	2025-08-12	21
Kansas City Westside WWTP, MO	WWTP	24h composite	61k	MU	2025-10-01	12
Milan Wastewater Treatment Facility (WWTF), MO	WWTP	24h composite	1.8k	MU	2025-02-24	31
Monett WWTP, MO	WWTP	24h composite	9.1k	MU	2025-06-10	29
South						
Miami-Dade Central District Wastewater Treatment Plant (CDWWTP), FL	WWTP	24h composite	900k	SB	2025-03-25	40
Central Oklahoma (OK-A), OK	WWTP	24h composite	110k	MU	2025-08-18	20
Central Oklahoma (OK-B), OK	WWTP	24h composite	440k	MU	2025-08-18	20
West						
Ontario, CA	WWTP	24h composite	890k	MU	2025-08-25	28
Palo Alto Regional Water Quality Control Plant (RWQCP), CA	WWTP	24h composite	240k	MU	2025-08-25	18
Riverside Water Quality Control Plant (WQCP), CA	WWTP	24h composite	350k	MU	2024-12-09	96
Sacramento, CA	WWTP	24h composite	1.6M	MU	2025-08-28	18
Southern California, CA	WWTP	24h composite	–	MU	2025-04-16	73
West Boise and Lander Street Water Renewal Facility (Boise RWF), ID	WWTP	24h composite	270k	MU	2024-11-04	57

After collection, local partners ship samples on ice packs to one of two sample processing and sequencing laboratories: Professor Marc Johnson at MU or SB in Boston (Figure 1b). Each laboratory applies a distinct sample processing workflow optimized for viral enrichment and RNA sequencing (details in the Methods). In both workflows, teams perform viral concentration, nucleic acid extraction, cDNA synthesis, and Illumina library preparation. MU processed the majority of samples in this release (1,078 samples from 26 sites), while SB processed samples from Boston and Miami (128 samples from 3 sites) (Table 1; Figure 2a). Both laboratories independently process the same physical samples from Boston's two Deer Island Treatment Plant (DITP) sites, enabling direct comparison of protocols.

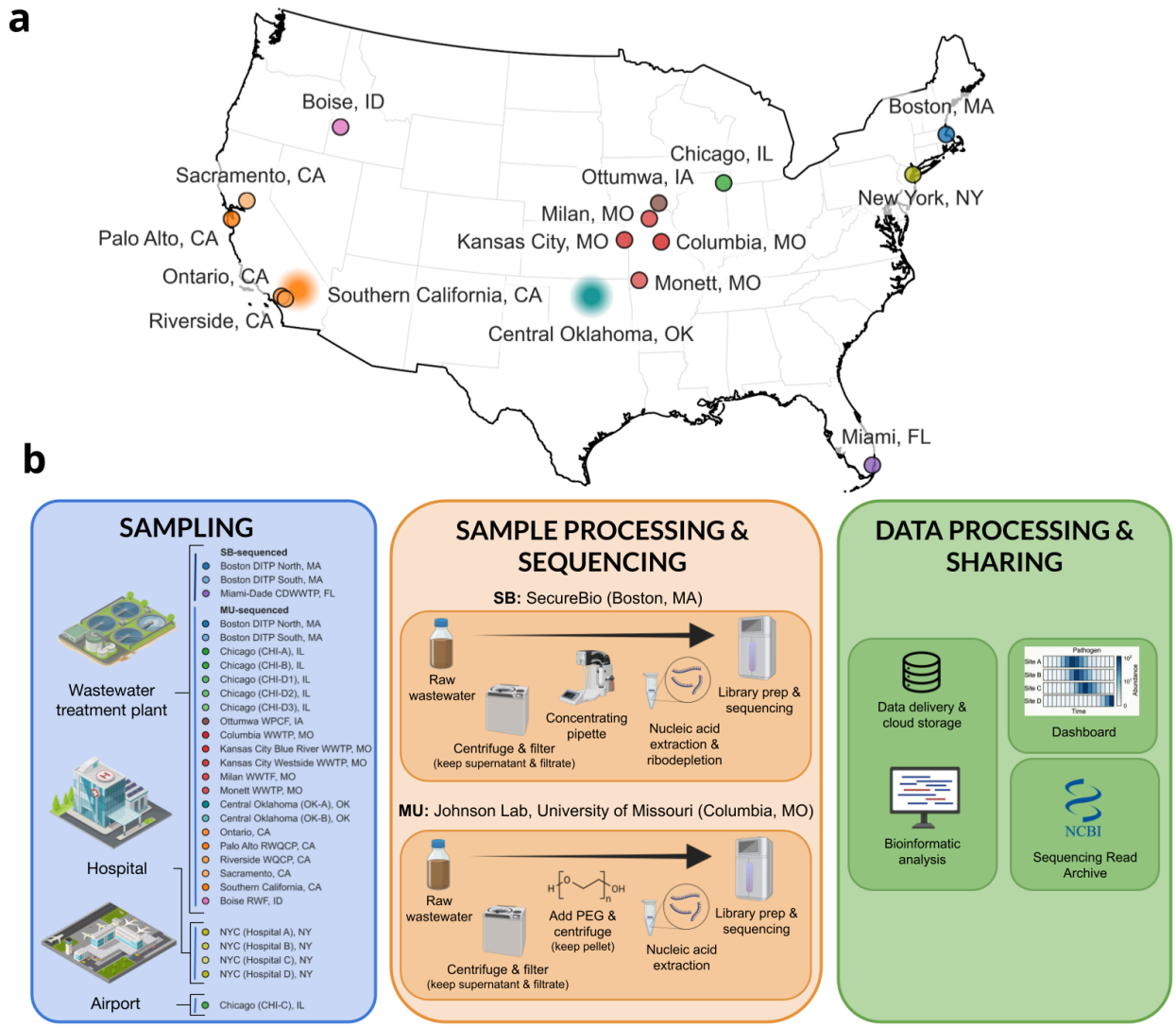

Figure 1. Overview of CASPER sampling network and workflow. (a) Geographic distribution of CASPER sampling sites included in this data release. Sites span nine U.S. states and 27 distinct wastewater catchments; some metropolitan areas (Chicago, Kansas City, Boston, and New York City) include multiple sampling sites. Marker locations for Southern California and Central Oklahoma are approximate to preserve site anonymity. Wastewater treatment plants labeled with a particular city may serve surrounding municipalities and communities. (b) Schematic illustrating the flow of samples from wastewater collection through laboratory processing, sequencing and public data release.

Most samples consist of 24-hour composite influent collected from municipal wastewater treatment plants (WWTP) at approximately weekly intervals, although sampling frequency varies by site (Figure 2a). A small number of sites, including Riverside WQCP, CA and Ontario, CA, collect samples twice weekly, whereas most others collect weekly. Samples from New York City (NYC) represent grab samples from hospital sewage, with one site collected from an indoor service hatch within the facility and three others from external manholes serving the hospitals exclusively. NYC samples were collected at more irregular intervals and often stored at 4°C for several months prior to sequencing. In future releases, we expect NYC collections to occur more regularly and with shorter turnaround times. Samples from the CHI-C site consist of Moore swabs (Sikorski and Levine 2020) collected from a manhole serving an airport. We selected sites based on established partnerships and to maximize geographic and population coverage across U.S. regions. Some sites, including Milan WWTF, MO and Ottumwa WPCF, IA receive animal agricultural input, enabling monitoring of agricultural and animal disease signals. Site names reflect preferences of local partners, and some sites opted not to disclose treatment plant names, locations, or other metadata.

All samples undergo deep, untargeted RNA sequencing on Illumina NovaSeq instruments, generating 2x150 bp paired-end reads and targeting approximately one billion read pairs per sample. NYC hospital samples sequenced December 2025 are the exception, targeting approximately 150 million read pairs per sample. Across the 1,206 samples included in this release, median sequencing depth is 986 million read pairs (range: 127M–3.25B; Figure 2b; Supplementary Figure S1), corresponding to a median of 279 gigabases (Gb) per sample (Figure 2e) and totaling 1.2 trillion read pairs (347 terabases) (Figure 2c, Supplementary Figure S2). This depth substantially exceeds that of other publicly available untargeted WW-MGS datasets, with median bases per sample approximately 25-fold higher than the next-deepest study identified in our literature survey (Figure 2e; Supplementary Table S1). Rapid declines in sequencing costs over the past decade have made deep sequencing at this scale feasible. Sequencing a typical CASPER sample would have cost approximately $4,000 USD in 2016, compared to approximately $500 USD in 2026 based on National Human Genome Research Institute (NHGRI) cost tracking data and internal estimates (Wetterstrand 2019; Kaufman 2025) (Figure 2e; Supplementary Table S2).

Median raw read quality score is Q39 (range: Q35–Q39) for both MU- and SB-processed samples. Median turnaround time from sample collection to delivery of sequencing data is 19 days (range: 9–657 days), varying by site and sequencing laboratory (Supplementary Figure S2). The longest turnaround times reflect NYC hospital grab samples collected during early pilot sampling and banked at 4°C prior to integration into routine processing; turnaround for these sites has since improved (Methods).

Following sequencing, data is transferred to Amazon Web Services (AWS) cloud storage and analyzed by MGS analysis pipelines developed by SB (Bradshaw et al. 2026) and other CASPER partners (O'Connor et al. 2026). Partner reports and raw sequencing data are made available immediately to CASPER members and visualized on a public dashboard ("Wastewater Surveillance Dashboards," n.d.). Raw data is periodically deposited in the NCBI SRA (Figure 1b). Public data from CASPER represent the largest publically available repository of untargeted

wastewater metagenomic sequencing data to date (Supplementary Figure S2). When combined with a previously released Los Angeles dataset (PRJNA1198001; 13 terabases Tb; Grimm, Rothman, et al. 2025), CASPER data account for more than 66% of untargeted wastewater sequencing on SRA (Supplementary Figure S2), and approximately 50% of all wastewater sequencing data including amplicon- and hybrid-capture-based datasets (Figure 2b).

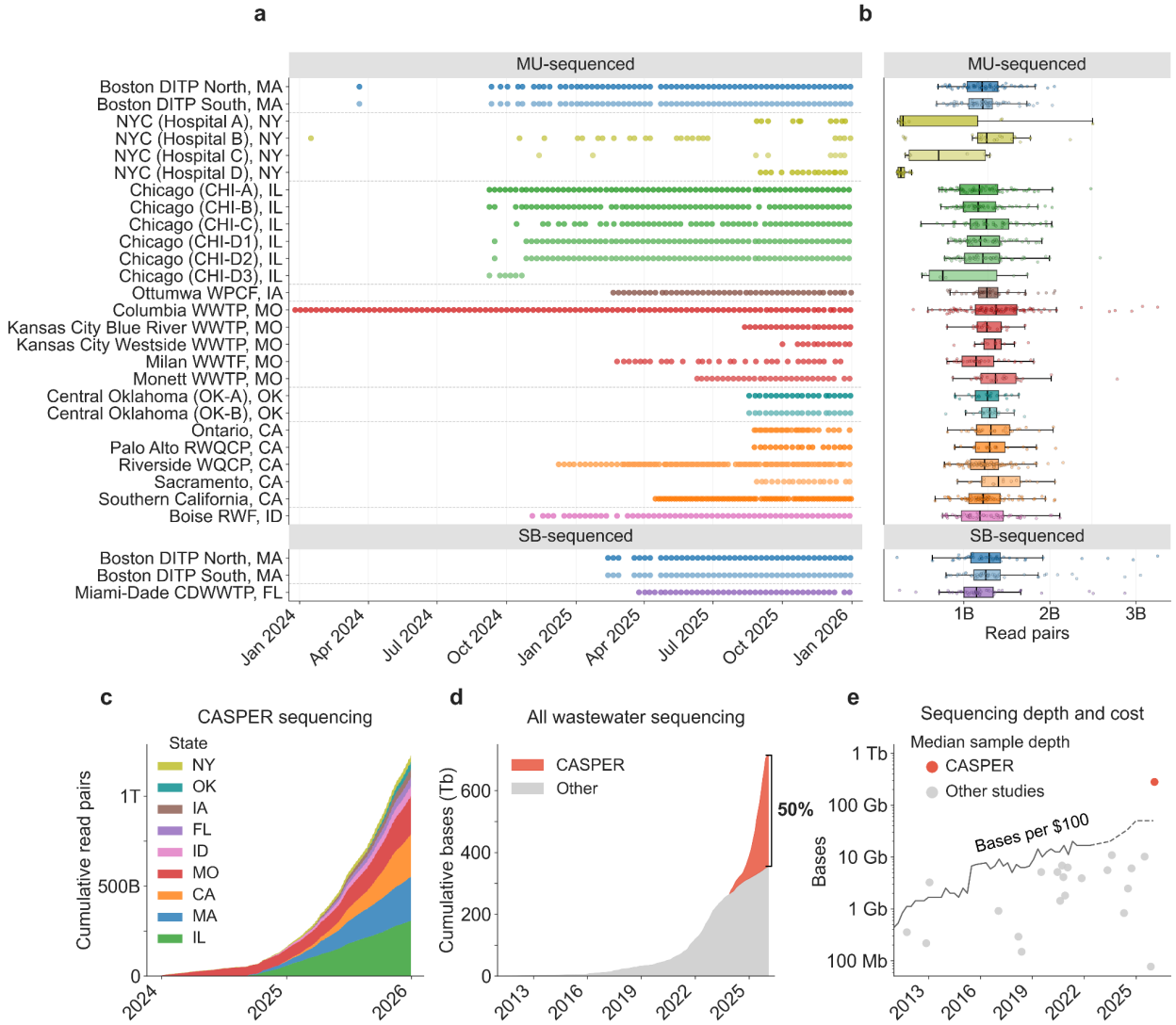

Figure 2. CASPER sequencing effort and scale. (a) Sample collection dates by site. (b) Sequencing depth per sample by site. (c) Cumulative read pairs generated by CASPER in this release over time. (d) CASPER contribution to all wastewater sequencing data deposited on the NCBI Sequence Read Archive (SRA). Includes data from both the current release (PRJNA1247874; 347 Tb) and the previous Los Angeles data release (PRJNA1198001; 13 Tb). (e) Median sequencing depth (bases per sample) across studies with publicly available untargeted wastewater metagenomic sequencing data, compared to CASPER samples in this release, alongside the number of bases purchasable for $100 USD over time (line). Sequencing cost data through 2022 are from NHGRI (Wetterstrand 2019; solid line); costs from 2023–present are estimated from observed sequencing output and pricing (Kaufman 2025; dashed line).

Taxonomic composition of sequencing data

To rapidly characterize the high-level composition of CASPER sequencing data, we perform shallow taxonomic profiling on a subset of reads from each sequencing library. For each library, we randomly sample one million read pairs per sequencing lane, resulting in approximately one million read pairs for SB-processed samples and eight million for MU-processed samples. Reads are partitioned into ribosomal and non-ribosomal fractions and classified using Kraken2 with the Standard reference database (Methods).

Across all samples, ribosomal RNA (rRNA) sequences, predominantly bacterial in origin, constitute the largest fraction of reads. The median rRNA fraction is 61.1% (range: 1.2–99.5%) (Figure 3b-c). Among the remaining non-ribosomal reads, sequences that could not be classified by Kraken dominate (median 28.9%, range 0.14–91.9%) followed by bacteria (median 5.7%, range 0.39–48.1%). Viruses (median 1.6%, range 0.00–14.3%), eukaryotes (median 0.08%, range 0.00–3.1%), and archaea (median 0.03%, range 0.00–0.87%) contribute smaller fractions (Figure 3b). The high proportion of unclassified reads likely reflects the limited taxonomic breadth of the Kraken Standard database, as well as the large fraction of wastewater microbiota that remains uncharacterized.

High-level taxonomic composition varies across sites and between sequencing laboratories (Figure 3b; Supplementary Figure S4), motivating the use of normalization strategies. Among SB-processed samples, a transient increase in rRNA fraction during mid-2025 reflects a temporary ribodepletion reagent failure (Figure 3c–d; Methods). Because metagenomic sequencing data is compositional and rRNA constitutes a large fraction of total reads, apparent pathogen abundance is heavily influenced by variation in ribosomal content. Several normalization strategies can account for this compositional variation, including normalization to non-rRNA reads, or to fecal indicator viruses such as pepper mild mottle virus (PMMoV) or tomato brown rugose fruit virus (ToBRFV). Here, we use PMMoV normalization for time-series pathogen abundance (Figures 6 and 7) because PMMoV is a widely established fecal strength indicator in wastewater surveillance (Symonds et al. 2018; Maal-Bared et al. 2023; Wolfe et al. 2021) and is consistent with CDC's NWSS normalization approach, facilitating comparison with PCR-based measurements. PMMoV abundance over time is shown in Figure 3e and demonstrates strong Pearson correlation with ToBRFV abundance (Supplementary Figure S5; MU: $R=0.87$, $p < 0.001$, SB: $R=0.98$, $p < 0.001$), indicating similar normalization outcomes.

Among viral reads, plant-infecting viruses, many associated with human diet, constitute the majority (58.6%), followed by viruses with unknown or ambiguous host assignment (34.7%), and bacteriophages (6.0%) (Figure 3a). The large unknown fraction likely reflects limited coverage of environmental bacteriophage lineages in Virus-Host DB, and true bacteriophage abundance is likely higher than reported here. Vertebrate-infecting viruses (VV) represent approximately 0.01% of all reads based on the Kraken2 shallow taxonomic profiling.

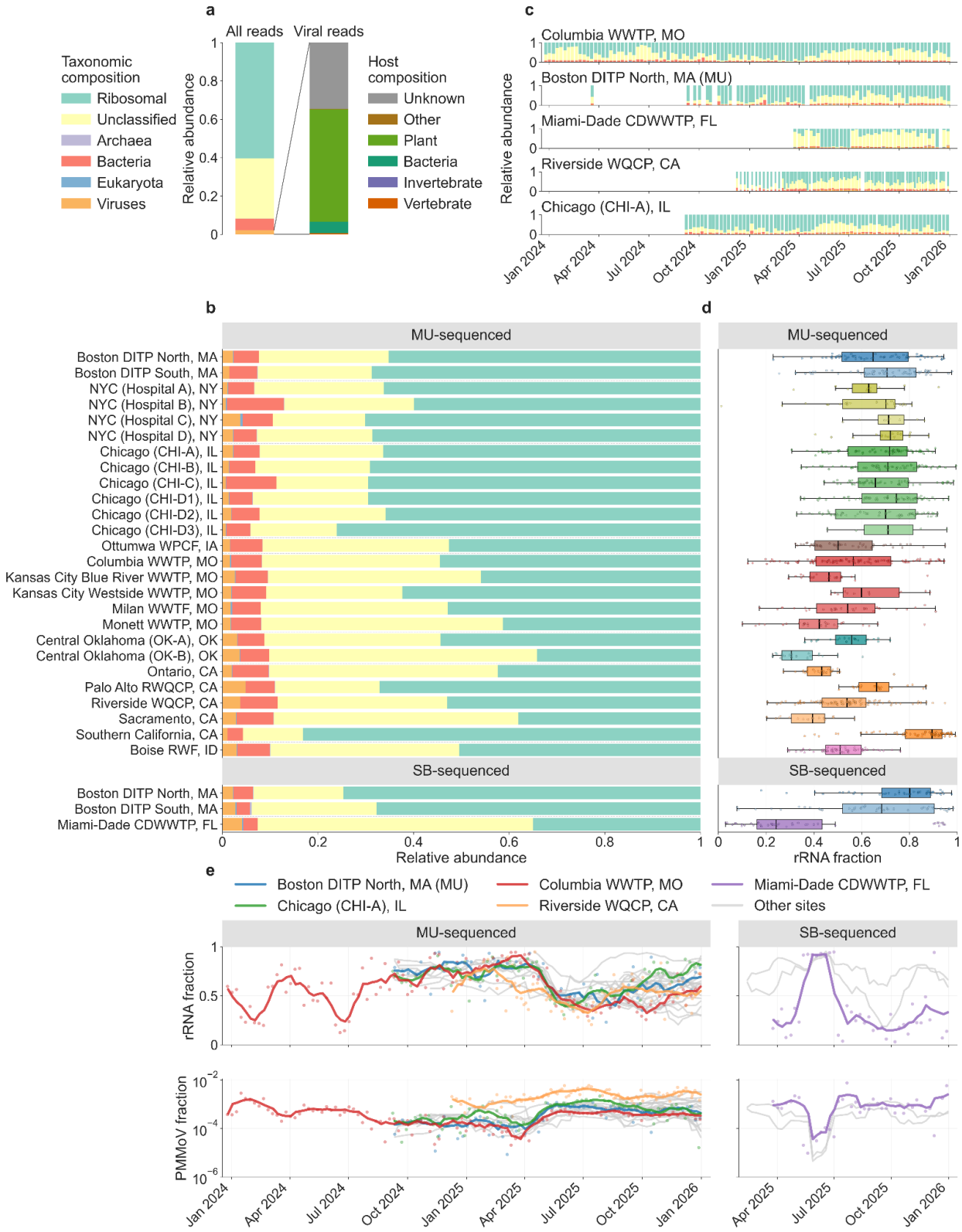

Figure 3. Taxonomic composition of CASPER sequencing data. (a) Aggregate composition across all samples, showing read classification by major taxonomic categories (left) and viral reads stratified by host category (right). (b) Aggregate composition by site and sequencing partner. (c) Temporal variation in taxonomic composition for selected sites. (d) Distribution of ribosomal RNA (rRNA) fraction by site. (e) rRNA fraction (top) and pepper mild mottle virus (PMMoV) relative abundance (bottom) over time, with select sites highlighted. The mid-2025 increase in rRNA fraction among SecureBio-processed samples reflects a transient ribodepletion reagent failure (see Methods). PMMoV is used as a normalization factor for pathogen time series in Figures 4 and 5.

Tracking vertebrate-infecting viruses

To sensitively identify vertebrate-infecting viruses (VV) reads across the full dataset, we analyze all sequencing data using SB's Viral Metagenomics Pipeline (v3.0.1; index version 20250825; github.com/securebio/nao-mgs-workflow; Bradshaw et al. 2026). This Nextflow-based workflow, executed on AWS Batch, is designed to screen large sequencing datasets efficiently for VV sequences. We first screen raw read papers against a masked VV genome database using exact k-mer matching and retain any read pair with at least one match. We then adapter-trim and quality-filter candidate reads and screen them against common contaminant genomes, including human, cow, pig, carp, mouse, *E. coli*, and genetic engineering vectors. We align remaining reads to the VV genome database using Bowtie2, filter alignments by a length-normalized score threshold, and assign NCBI taxonomy identifiers using a custom lowest common ancestor (LCA) algorithm (Methods). We construct the VV genome database from Virus-Host DB and NCBI GenBank, excluding transgenic, recombinant, and unverified sequences, and masking reference genomes to reduce spurious alignments. This dedicated, more sensitive VV-pipeline identifies 0.03% of total reads across all samples as VV, approximately three times the number detected by the Kraken2 shallow taxonomic profiling.

Among samples, VV reads constitute a median of 0.01% of total reads (range: 0.00–1.08%), varying by site and sequencing partner (Figure 4b; Supplementary Figure S6). Enteric virus families dominate, including *Sedoreoviridae* (rotaviruses), *Caliciviridae* (noroviruses, sapoviruses), *Astroviridae*, and *Picornaviridae* (enteroviruses, rhinoviruses), consistent with the fecal origin of wastewater (Figure 4a). Respiratory virus families, including *Orthomyxoviridae* (influenza viruses), *Coronaviridae* (coronaviruses), *Paramyxoviridae* (parainfluenza viruses) and *Pneumoviridae* (RSV, metapneumovirus), occur at lower relative abundance but remain readily identifiable (Figure 4c).

Hospital sites and sites receiving agricultural inputs show distinctive VV composition signatures. Ottumwa, IA, which receives agricultural wastewater, shows elevated *Tobaniviridae* and *Coronaviridae* consistent with porcine viruses (buldecovirus, pedacovirus, torovirus). Hospital samples representing grab collections taken closer to the source, with less dilution from other sewershed inputs, and from populations enriched for sick or immunocompromised individuals (Renfro et al. 2026), also diverge from municipal influent patterns, with NYC Hospital A and C showing elevated *Picobirnaviridae* (human picobirnavirus) and *Picornaviridae* (enterovirus) abundance.

We next examined time-series trends in PMMoV-normalized abundance for respiratory and enteric viruses in CASPER data (Figure 4c). Across geographically distant locations, abundance trends for both respiratory and enteric viruses show broad concordance, with synchronized seasonal dynamics and observable leads and lags between sites likely reflecting regional outbreak timing, consistent with patterns observed in large-scale PCR-based wastewater monitoring programs (Chan and Boehm 2025; Duvallet et al. 2022). Among respiratory viruses, influenza A and B, RSV, human metapneumovirus, and seasonal human coronaviruses display consistent winter and spring. Parainfluenza viruses exhibit more variable patterns. At Columbia, MO, one of the few sites with multi-year coverage, parainfluenza 3 shows spring peaks in both 2024 and 2025, whereas parainfluenza 2 is detected in fall 2025 but absent the previous year.

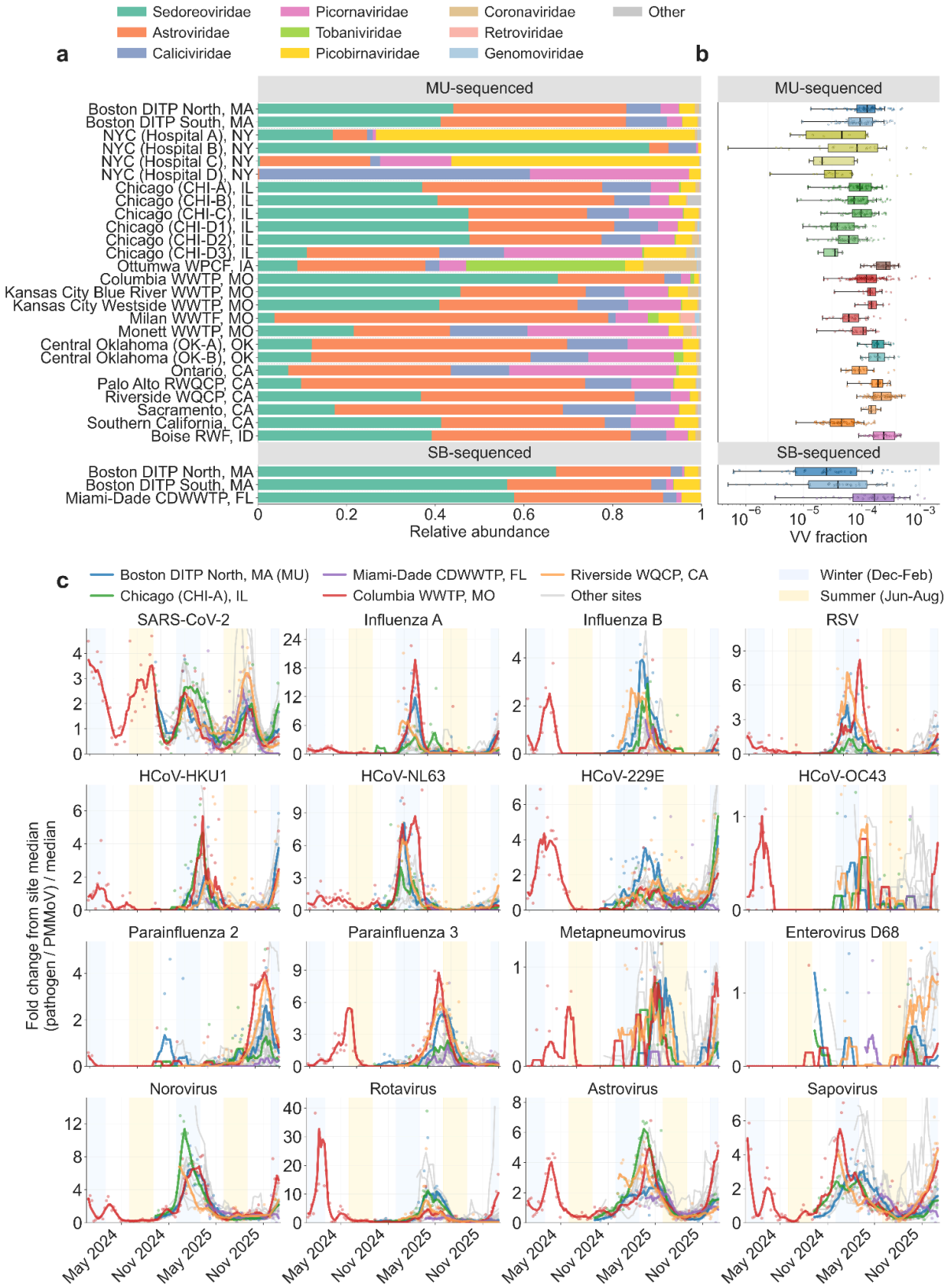

Figure 4. Vertebrate-infecting virus composition and temporal dynamics. (a) Family-level composition of vertebrate-infecting virus (VV) reads by site. (b) Total VV read fraction by site. (c) Temporal patterns of select human-infecting viruses across CASPER sites. For each pathogen, abundance is normalized by pepper mild mottle virus (PMMoV) and expressed as fold change relative to the site-specific median of non-zero PMMoV-normalized values. Respiratory viruses (top three rows) and gastrointestinal viruses (bottom row) are shown over time, with five representative sites highlighted in color and remaining sites shown in gray. Y-axis limits are scaled such that the highest smoothed peak among highlighted sites falls at 75% of the axis height; values exceeding this range are clipped. MU: University of Missouri, SB: SecureBio.

Enteric viruses including norovirus, rotavirus, astrovirus, and sapovirus show similar cross-site concordance, with synchronized winter and spring peaks (Figure 4c) and generally higher relative abundance than respiratory pathogens (Supplementary Figure S7). Enterovirus D68 shows elevated abundance across California sites in fall 2025, consistent with regional wastewater PCR data (WastewaterSCAN, n.d.). Rhinovirus serotype dynamics are tracked via a public dashboard ("Rhinovirus Serotypes Dashboard," n.d.). CASPER data enables tracking of hundreds of additional taxa and supports genomic analyses including detecting shifts in viral lineages.

Beyond routine pathogen tracking, CASPER data has contributed to detection of unexpected and emerging threats. Analysis of Columbia, MO wastewater revealed highly pathogenic avian influenza H5N1 (genotype B3.13) during March to May 2024, concurrent with initial USDA reports of dairy cattle infections despite no confirmed infected herds in Missouri (Rushford et al. 2025). The same site showed an unanticipated influenza C surge during winter 2023–2024 and wastewater detections of West Nile virus in fall 2025 that helped inform public health messaging (Missouri DHSS 2025; Rushford et al. 2025). The same deep sequencing workflow, applied to samples from Hawaii (not included in this public data release), identified measles virus in Kaua'i County wastewater in September 2025 with no identified clinical cases (Hawaii DOH 2025).

To assess concordance with established pathogen monitoring methods, we compared CASPER WW-MGS measurements with publicly available wastewater PCR data from the CDC NWSS data portal, which aggregates data generated by CDC, state and local health departments, and by WastewaterSCAN (Boehm et al. 2024, 2026), at sites with overlapping coverage (Supplementary Table S3). For SARS-CoV-2, influenza A, and RSV, WW-MGS and PCR measurements show weak to strong positive correlations at sites with extended time series (Figure 5a; Spearman R = 0.24–0.91). Correlations for all sites are reported in Supplementary Figure S8 and Supplementary Table S4; NWSS PCR assay metadata is reported in Supplementary Table S5.

For Boston, we compared WW-MGS-derived abundance with clinical case counts from testing performed at the Massachusetts General Hospital (MGH), a large academic medical center whose patient catchment substantially overlaps the DITP sewershed (Figure 5b). WW-MGS measurements from both the DITP North and South systems correlate with MGH positive test counts for all three viruses (R = 0.67–0.84; Supplementary Table S6). Together, these comparisons indicate that CASPER wastewater sequencing data tracks population-level pathogen dynamics comparably to established monitoring methods, while enabling simultaneous detection across a broader range of pathogens and providing genomic information.

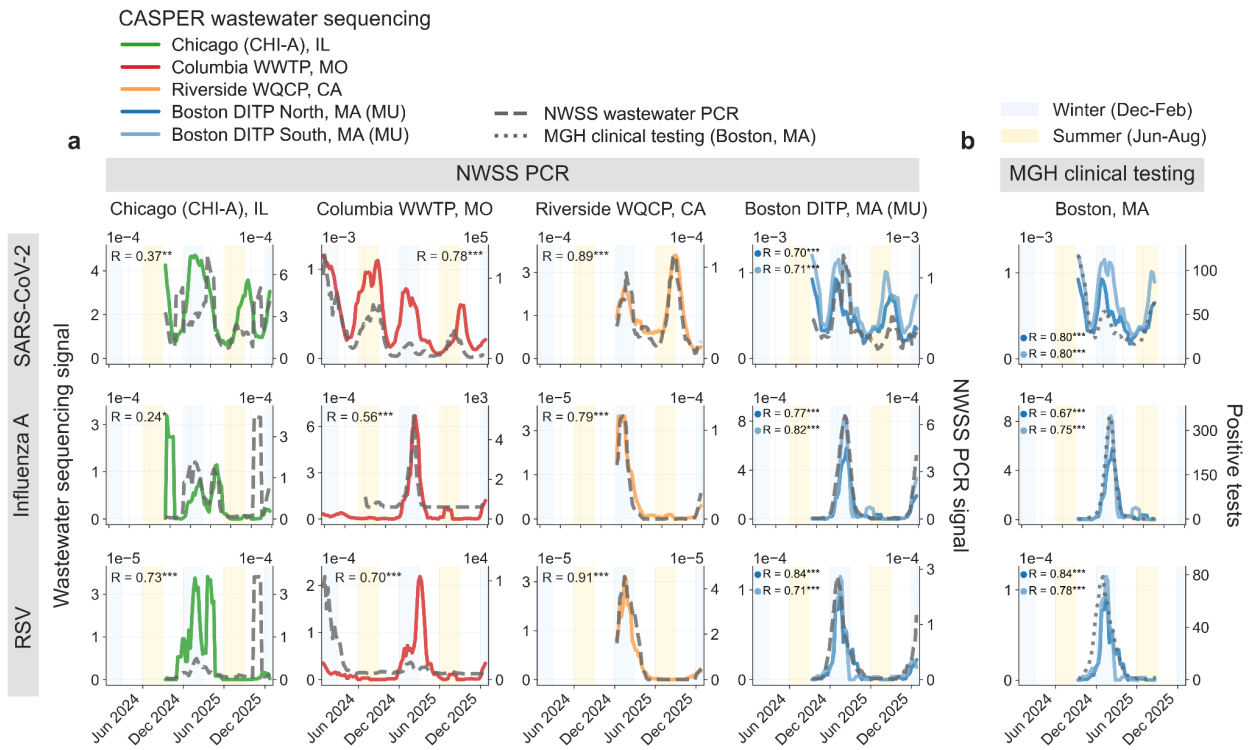

Figure 5. Agreement between wastewater metagenomic sequencing, wastewater PCR data, and clinical diagnostic testing. (a) Wastewater metagenomic sequencing (WW-MGS) abundance compared with PCR measurements from the National Wastewater Surveillance System (NWSS) data portal (generated by state and local health departments and WastewaterSCAN; Boehm et al. 2024, 2026). All sites except Columbia, MO show pepper mild mottle virus (PMMoV)-normalized sequencing and PCR abundance; Columbia, MO shows raw relative abundance and concentration because PMMoV-normalized PCR data were unavailable. PCR assay details and data sources are described in Supplementary Table S5. (b) WW-MGS abundance from Boston Deer Island Treatment Plant (DITP) compared with clinical testing data from Massachusetts General Hospital (MGH). Clinical data represent positive counts from multiplex nucleic acid amplification tests performed on symptomatic patients as part of routine clinical care. For both panels, Spearman correlation coefficients (R) were calculated using week-aggregated data with centered 5-week smoothing (Supplementary Tables S4 and S6). *p < 0.05, **p < 0.01, ***p < 0.001. MU: University of Missouri-sequenced.

Discussion

CASPER represents the largest publicly available resource for untargeted wastewater metagenomic sequencing. Encompassing two years, 13 million people, 1,206 samples, 347 terabases, and 27 sites across nine U.S. states, this dataset substantially expands the scale of available wastewater sequencing data and demonstrates the feasibility of scalable workflows for deep wastewater sequencing. The resulting time-series data enable broad detection and tracking of pathogens. By adopting a pathogen-agnostic approach, CASPER extends monitoring beyond common targets such as SARS-CoV-2, influenza, and RSV, and supports detection of rare, novel, or unexpected pathogens, including the highly pathogenic avian influenza H5N1, West Nile virus, and measles virus identified in this data, independent of clinical cases. CASPER data is not static; sampling, sequencing, and periodic data release will

continue, incorporating additional sites over time. We intend for CASPER to serve as a growing resource for pathogen monitoring, public health, and biosecurity, as well as for research into urban microbiomes and viral ecology.

A key distinguishing feature of CASPER is its sequencing depth. Most existing untargeted WW-MGS studies have used substantially lower per-sample depth, limiting their ability to resolve rare viral targets. Some have concluded from this that untargeted WW-MGS is ineffective for viral detection altogether (Bellekom et al. 2026). However, at shallow depths (e.g., ~200k reads per sample in Bellekom et al. 2026), the expected number of reads from low-abundance targets approaches zero regardless of whether the pathogen is present (Grimm, Kaufman, et al. 2025). By sequencing approximately 5,000-fold deeper, CASPER demonstrates that the limitation lies in depth, not in untargeted approaches themselves. Even so, sensitivity constraints remain (Grimm, Kaufman, et al. 2025), particularly for pathogens with low prevalence or low fecal shedding, and genome coverage for most pathogens remains partial, limiting strain-level characterization. Despite these constraints, abundance trends show strong concordance with both wastewater PCR measurements from established monitoring programs and clinical testing data, supporting the use of WW-MGS as a quantitative pathogen monitoring tool.

We anticipate that CASPER will have broad utility for pathogen surveillance and environmental microbiology. For pathogen surveillance, these data can support the development and evaluation of novel detection approaches, including anomalous growth detection (The Nucleic Acid Observatory Consortium 2021), genetic engineering detection (Tay et al. 2024; Kaufman 2024; Crook et al. 2022), AI/ML methods (Liu et al. 2025), and reference-, marker-, and functional-based approaches (Luebbert et al. 2025; Balaji et al. 2022; Wittmann et al. 2025). CASPER's depth and scale also enable large-scale metagenomic assembly, supporting discovery and characterization of novel sequences and exploration of microbial dark matter (Fiamenghi et al. 2025; Edgar et al. 2022; Neri et al. 2022; Worp et al. 2025). The data further support development of tools for viral evolution and outbreak dynamics suited to low-coverage data (Levy et al. 2025), and evaluation of alternative pathogen monitoring approaches including air sampling (Justen et al. 2025; Minor et al. 2023), airport and aircraft wastewater surveillance (St-Onge et al. 2025; Friedman et al. 2025; McLaren et al. 2026), and hybrid-capture sequencing (Tisza et al. 2023; Kantor and Jiang 2024).

To develop CASPER as a valuable real-time pathogen surveillance resource, we are improving sample processing and data analysis techniques, establishing data use agreements and procedures for sharing findings with stakeholders, and reducing turnaround time. Within the consortium, partners share data as results become available and present them through a public dashboard. Bulk data uploads to the SRA currently occur on a delayed schedule.

In sharing this resource, we wish to highlight that interpreting WW-MGS data for public health and biosecurity requires careful consideration of nucleic acid sources beyond active human infection. These sources include sample contamination, laboratory and industrial waste, stormwater intrusion, animal infections, diet-associated contributions, and vaccine shedding. This complexity can make it challenging to determine whether detected sequences originate from human infections or other sources, especially when limited sequencing coverage permits only partial strain characterization. Investigation of flagged sequences, often based on only a

small number of unique reads, is therefore nuanced, and determining whether a signal represents an actionable finding requires careful consideration of the environmental context. This perspective guides our communications with stakeholders, and we encourage others to bear it in mind.

We invite researchers, public health agencies, and other interested parties to engage with this growing resource. As the CASPER network continues to expand, we welcome inquiries from potential collaborators at casper-inquiries@securebio.org.

Methods

Sample processing and sequencing at the University of Missouri

Sample collection and transport

Composite samples. Influent wastewater was collected from most locations as a 24-hour composite sample. Samples were aliquoted into three 50 mL conicals (150 mL total), shipped overnight with ice packs in insulated containers, and stored at 4°C until processing. Samples were processed in batches, typically within one week of receipt. Chicago composite samples CHI-D1, D2, and D3 represent different catchments served by the same wastewater treatment plant.

Moore swab samples. CHI-C samples were collected using a Moore swab deployed for 24 hours in a manhole serving an airport. Wastewater was recovered by manual compression of the swab into a collection bag, yielding approximately 30–40 mL of sample which was transferred to a 50 mL conical for shipment.

Grab samples. NYC hospital samples were collected via grab sampling from four facilities, with approximately 100 mL collected per sample and shipped in two 50 mL conicals. Hospital A samples were collected from an indoor service hatch within the facility at approximately 5:00 AM. Hospitals B, C, and D samples were collected from external manholes exclusively serving the hospital facilities: Hospital B at approximately 12:00 PM, Hospital C at approximately 9:00 AM, and Hospital D between 8:00–11:00 AM. The Hospital B manhole also receives stormwater from an adjacent dock area; Hospitals C and D manholes have minimal rainwater input. These samples were initially collected as an exploratory effort, which accounts for extended processing times: samples were stored at 4°C prior to shipping to MU, with a median time from collection to sequencing of 62 days (range: 21–657 days). We expect future turnaround times for these sites to align with other locations in the network.

Viral concentration and nucleic acid extraction

Samples were centrifuged at 2000 xg for 5 minutes at 4°C and the supernatant was filtered through 0.22 µm filters (Millipore, SCGP00525). Filtrates (35 mL) were mixed with 12.5 mL of 50% (w/vol) PEG (Research Products International, P48080-1000.0) and 1.2 M NaCl and incubated for 1 h at 4°C. The mixture was then centrifuged at 12,000 xg for 2 h at 4°C to pellet viruses. After decanting the supernatant, nucleic acids were extracted from viral pellets using the QIAamp Viral RNA Mini Kit (Qiagen, 52906) according to the manufacturer's instructions, omitting carrier RNA and DNase treatment. Extractions prior to 2025-06-04 were performed using a Qiagen QIAcube; subsequent extractions were performed manually in a biosafety cabinet.

A subset of Columbia, MO samples collected between 2024-01-02 and 2024-03-19 underwent enzymatic pretreatment prior to filtration to reduce bacterial ribosomal RNA. Samples were

incubated with 10,000 ng RNase A (ThermoFisher, R1253), 2 units of TURBO DNase (Invitrogen, AM1907), and 4.75 mL of buffer containing 5 mM $CaCl_2$ and 25 mM $MgCl_2$. Some samples additionally received 30 mM EDTA to promote dissociation of rRNA from ribosomal complexes. For all nuclease-treated samples, 400 units of RNaseOUT™ Recombinant Ribonuclease Inhibitor (Invitrogen, 10777019) were added prior to RNA extraction. Treatment conditions varied over time: samples from 2024-01-02 through 2024-01-09 were processed in triplicate under three conditions (no treatment, nuclease only, nuclease plus EDTA); samples from 2024-01-16 through 2024-02-20 were processed under two conditions (nuclease only, nuclease plus EDTA); and samples from 2024-02-27 through 2024-03-19 were processed under two conditions (no treatment, nuclease only). These treatments did not substantially affect rRNA content or viral detection (Rushford et al. 2025); all samples collected after 2024-03-19 were therefore processed without enzymatic pretreatment. For samples processed under multiple conditions, sequencing data were merged prior to public release on SRA.

To evaluate algorithms designed to detect engineered sequences, we spiked 1 µL of a previously described NL4-3–derived HIV construct (containing a CMV-driven puromycin resistance gene and lacking *env*, *nef*, *vif*, and *vpr*) into the Columbia, MO sample collected on 2024-07-09, prior to filtration (Robinson et al. 2022).

RT-qPCR quantification of PMMoV

Successful concentration of viral RNA were evaluated prior to library preparation using comparative RT-qPCR targeting PMMoV. RT-qPCR assays utilized PMMoV specific forward (5'-GGCGTAGATCCATTGGTGC-3') and reverse (5'-CGAACCTTCCTCCTTTGATG-3') primers, each prepared at a stock concentration of 100 µM, along with a TaqMan probe (VIC-5' GCTGTGGTTTCAAATGAGAGTGG 3'-QSY) prepared at 100 µM. Each 20 µL RT-qPCR reaction consisted of 5 µL of TaqPath 1-Step RT-qPCR Master Mix GC (Applied Biosystems, A15299), forward and reverse primers at a final concentration of 500 nM each, the TaqMan probe at 125 nM, and 5 µL of the RNA template. Amplification was performed on an Applied Biosystems 7500 Fast Real-Time PCR System using the following cycling conditions: reverse transcription at 48 °C for 15 min, initial denaturation at 95 °C for 10 min, and 40 amplification cycles of 95 °C for 15 s followed by 60 °C for 1 min.

Library preparation and sequencing

Total RNA concentrations were measured using the Qubit RNA High Sensitivity Assay prior to library construction. RNA input amounts were normalized based on Qubit quantification, targeting 10–15 ng of RNA in a final volume of 9.5 µL. For samples with RNA concentrations below the detection limit, the full 9.5 µL was used as input. Libraries were prepped using the Illumina Stranded mRNA Prep (Illumina, 20040534) and index adapters sets A–D (Illumina, 20091655, 20091657, 20091659, and 20091661). Sequencing libraries were prepped according to the manufacturer's protocol with minor modifications listed below.

The mRNA capture step was omitted, and library preparation began directly with RNA fragmentation followed by cDNA synthesis. Each normalized RNA sample was divided into two equal fractions with one fraction undergoing fragmentation to reduce longer RNA molecules to a

size range suitable for sequencing, and the second fraction left unfragmented to preserve RNA already within the target size range. Following fragmentation, both fractions were recombined prior to cDNA synthesis.

Final indexed libraries were size-selected during the post-PCR cleanup using a two-step AMPure XP bead protocol (Beckman Coulter, A63881). Initially, 30 µL of AMPure XP beads were added to 50 µL of post-indexing PCR products to bind larger library fragments. After a 5-minute incubation, the supernatant was transferred to a new tube and combined with an additional 20 µL of AMPure XP beads to capture shorter fragments. These shorter bead-bound libraries were washed, eluted in 30 µL of resuspension buffer, and used for final pooling.

Sequencing was performed using Illumina NovaSeq platforms with paired-end 2×150 bp reads. Libraries generated through 2024-04-16 were sequenced on a NovaSeq 6000, while libraries prepared after were sequenced on a NovaSeq X system. The transition to NovaSeq X was associated with improved quality scores and a transient decrease in GC content for Columbia, MO samples (Supplementary Figures S10–S11), which resolved as the sequencing core optimized performance on the new instrument.

While most libraries were sequenced at 2x150 bp, a small number of samples were sequenced with different read lengths due to pilot runs or inadvertent sequencing core errors (Supplementary Figure S12): (1) Columbia, MO samples from 2023-12-26, 2024-01-02, and 2024-01-09 include a small fraction of reads at 2x75 bp from initial shallow sequencing, with the majority at 2x150 bp from subsequent deep sequencing; (2) Columbia, MO samples from 2024-06-11 through 2024-07-02 include approximately 900M reads at 2x100 bp alongside 2x150 bp reads, as part of diagnostic sequencing to investigate the transient GC bias; (3) Columbia, MO samples from 2024-08-13 through 2024-09-10, Boston DITP North and DITP South from 2024-09-11, and CHI-A, CHI-B, and CHI-D3 from 2024-09-08 were inadvertently sequenced at 2x100 bp.

Libraries from samples collected prior to 2024-09-24 were generated by the University of Missouri's Genomics and Technology Core while samples collected after were processed in Marc Johnson's lab at the University of Missouri.

Sample processing and sequencing at SecureBio

Sample collection and transport

Wastewater samples consisting of 1 L 24-hour composite influent from Miami-Dade CDWWTP and Boston DITP North and South were shipped to SecureBio for processing and sequencing. Boston samples were transported via courier in coolers containing ice packs. Miami samples were shipped overnight via FedEx in insulated containers to maintain temperature stability. Upon receipt, all samples were processed either immediately or within 24 hours. A 200 mL aliquot from each bulk sample was used for downstream processing.

Viral concentration, nucleic acid extraction, and rRNA depletion

Viral concentration and nucleic acid extraction were performed as described in (Machtinger et al. 2025). To facilitate dissociation of viral particles from solid matter, Tween 20 (Sigma Aldrich, P7949) was added to the 200 mL aliquot to a final concentration of 0.1%. Samples were vortexed at medium speed for approximately 1 minute, followed by sonication in a water bath at 40 kHz for 1 minute.

Following physical dissociation, samples were clarified by centrifugation at 10,000 xg for 5 minutes at 4°C using an Eppendorf 5920R centrifuge with a FA-6x250 fixed-angle rotor. The supernatant was decanted and filtered through a 0.45 µm polyethersulfone (PES) vacuum filtration unit (VWR, 10040-470) to remove remaining bacterial cells and large debris.

Viral concentration was performed using the InnovaPrep Concentrating Pipette Select with an Ultra Concentrating Pipette Tip (CPT; InnovaPrep, CC08004-100), following the manufacturer's recommended protocol adjustments for wastewater processing (*Concentrating Pathogens from Raw and Primary Wastewater Using the InnovaPrep® Concentrating Pipette, Protocol (Revision E)*, n.d.)). Samples were eluted using InnovaPrep Buffer-Tris fluid (InnovaPrep, HC08005) directly into a collection tube containing 400 µL of DNA/RNA Shield (Zymo Research, R1200-125) to achieve a 1:1 ratio of eluate to shield, resulting in a final concentrate volume of approximately 800 µL.

Total nucleic acids were extracted from the resulting concentrate using the Quick-Viral™ RNA/DNA Miniprep Kit (Zymo Research, D7021) according to the manufacturer's instructions, scaled for the volume of the concentrated sample. To prepare samples for RNA sequencing, extracts were treated with the TURBO DNA-free™ Kit (Thermo Fisher Scientific, AM2238) to remove genomic DNA. Bacterial ribosomal RNA was subsequently depleted using the Pan-Bacterial RiboPOOLs kit (siTOOLs Biotech, dp-K096-00026). Due to a reagent preparation error, ribodepletion probes were omitted from samples processed between May 13 and July 1, 2025 (inclusive). Sequencing data from these samples exhibit a higher fraction of ribosomal reads and reduced effective depth for non-ribosomal targets (Figure 3).

Library preparation and sequencing

Total RNA concentrations were quantified using the Qubit RNA High Sensitivity Assay (ThermoFisher Scientific, Q32855). RNA sequencing libraries were constructed using 5 µL of ribodepleted RNA as input. Library preparation was performed using the xGen RNA Library Prep Kit (Integrated DNA Technologies, 10009814) with IDT Unique Dual Index (UDI) PCR primers (Integrated DNA Technologies, 10005922) according to the manufacturer's instructions. Final amplified libraries were pooled at equimolar concentrations (2–4 nM) and submitted to the Broad Clinical Labs (Broad Institute, Cambridge, MA, USA) for sequencing on the NovaSeq X platform (Illumina, San Diego, CA, USA) at 2x150 bp.

Bioinformatic processing

Sequencing data was analyzed using SecureBio's Viral Metagenomics Pipeline (v3.0.1; index version 20250825; github.com/securebio/nao-mgs-workflow; Bradshaw et al. 2026), implemented in Nextflow (v25.10.0; Di Tommaso et al. 2017) and executed on AWS Batch. For this manuscript, the pipeline performs two primary analyses: (1) shallow taxonomic profiling on a subset of reads, and (2) deep VV profiling on all reads.

Preprocessing and quality control

Total read counts were computed from raw demultiplexed paired-end FASTQ files. Sequencing libraries were typically sequenced across multiple lanes and sometimes multiple sequencing runs; we term each lane-barcode combination an "index group." For quality assessment and taxonomic profiling, read pairs were randomly subsampled to approximately one million per index group using Seqtk (v1.5; Li 2013). Because MU-processed samples were typically sequenced across more lanes than SB-processed samples, the number of reads profiled per library was higher for MU (median 8.0M, range 2.0M–31.9M) than for SB (median 942K, range 563K–3.9M).

Subsampled reads were adapter-trimmed and quality-filtered with FASTP (v0.23.4; parameters: `--cut_front --cut_tail --correction --detect_adapter_for_pe --trim_poly_x --cut_mean_quality 20 --average_qual 20 --qualified_quality_phred 20 --low_complexity_filter`; Chen et al. 2018). Quality metrics such as mean GC content (Supplementary Figure S10) and quality scores (Supplementary Figure S11) were generated using FASTQC (v0.12.1; Andrews 2010) and aggregated with MultiQC (v1.21; Ewels et al. 2016).

Shallow taxonomic profiling

Subsampled, adapter-trimmed reads were partitioned into ribosomal and non-ribosomal fractions using BBDuk (BBTools v39.01; parameters: `k=27 minkmerfraction=0.4`; Bushnell 2014) against the SILVA ribosomal RNA databases (v138.2; Chuvochina et al. 2025): the SSU (Small Subunit; 16S/18S rRNA) NR99 and LSU (Large Subunit; 23S/28S rRNA) NR99 reference sets. Both fractions were independently classified with Kraken2 (v2.1.3; Wood et al. 2019) using the Standard reference database (k2_standard_20250714), which includes RefSeq complete genomes for bacteria, archaea, and viruses, the human reference genome (GRCh38), and UniVec_Core vector sequences. Ribosomal and non-ribosomal classifications were reported separately.

For aggregate taxonomic composition analysis (Figure 4), we combined ribosomal and non-ribosomal Kraken2 classifications into six categories: (1) Ribosomal—all rRNA reads regardless of taxonomic classification; (2) Bacteria (non-rRNA only); (3) Archaea (non-rRNA reads only); (4) Eukaryota (non-rRNA only); (5) Viruses (non-rRNA only); and (6) Unclassified—non-rRNA reads that were either unclassified by Kraken2 or assigned directly to the root or cellular organisms without more specific classification. Reads assigned to NCBI's "unclassified entries" or "other entries" were excluded as ambiguous. Site-level compositions were calculated by summing estimated read counts (category fraction × total reads) across all

libraries at each site, then normalizing to proportions, effectively weighting each library's contribution by its sequencing depth.

Vertebrate-infecting virus identification

All reads from each index group underwent deep profiling for vertebrate-infecting viruses:

1. **K-mer screening.** Raw read pairs were screened against a masked database of VV genomes using BBDuk (parameters: `k=24 minkmerhits=1`), retaining any read pair containing at least one 24-mer exact match to any VV genome. This rapid initial filter discards the vast majority of non-viral reads while retaining putative VV reads for more computationally intensive downstream analysis.
2. **Read cleaning.** Putative VV reads were adapter-trimmed and quality-filtered with FASTP (parameters as above), followed by additional adapter trimming with Cutadapt (v5.0; parameters: `-b -B -m 20 -e 0.33 --action=trim`), which trims adapters from either end of either read and discards read pairs where either read falls below 20 bp after trimming.
3. **Alignment to VV database.** Cleaned reads were aligned to the VV genome database using Bowtie2 (v2.5.4; parameters: `--local --very-sensitive-local --score-min G,0.1,19 -k 10`; Langmead and Salzberg 2012), retaining up to 10 alignments per read to enable subsequent lowest common ancestor (LCA) analysis.
4. **Contaminant removal.** Reads aligning to the VV database were then screened against common contaminant genomes using Bowtie2 (parameters: `--local --very-sensitive-local`), sequentially filtering against human (GRCh38), cow, pig, carp, mouse, *E. coli*, and common genetic engineering vectors. Reads mapping to any contaminant were discarded.
5. **Score filtering.** Surviving alignments were filtered by a length-normalized alignment score threshold (alignmentScore / ln(readLength) ≥ 20) to remove low-confidence alignments.
6. **Taxonomic assignment via LCA.** Reads passing score filtering were assigned NCBI taxids using a custom LCA algorithm. For reads with multiple alignments, the algorithm computes the LCA across all aligned taxa, with special handling for unclassified taxa (those whose names contain "unclassified" or "sp.") and artificial sequences. Unclassified taxa are included in LCA calculation only if they have the highest alignment score and are not tied with any classified taxon. LCA assignments are computed separately for natural sequences, artificial sequences, and all sequences combined; natural-only LCA assignments are used for downstream analysis.

Vertebrate-infecting virus genome database construction

The VV genome database was constructed by the pipeline's INDEX workflow as follows: (1) vertebrate-infecting virus taxids were obtained from Virus-Host DB (daily release, 2025-08-25) and NCBI Taxonomy (downloaded 2025-08-25); (2) this list was expanded to include all descendant taxonomy IDs; (3) all corresponding viral genome sequences were downloaded from NCBI GenBank (downloaded 2025-08-25) using ncbi-genome-download (v0.3.3; Blin 2023); (4) sequences were filtered to exclude entries with header keywords indicating

transgenic, mutant, recombinant, unverified, or draft sequences, as well as specific accessions previously identified as contaminated or erroneous; (5) remaining genomes were masked using BBDuk in two passes to reduce spurious alignments: first, adapter sequences were masked (parameters: `k=20 hdist=3 mink=8 hdist2=1 mask=N`) along with low-entropy regions (parameters: `entropy=0.5 entropymask=t`); second, homopolymer runs of ≥10 bp were masked (parameters: `k=10 hdist=0 mask=N`).

Host annotation

Viral taxa were annotated with host infection status using a custom algorithm that propagates host information through the NCBI viral taxonomy based on Virus-Host DB annotations (Mihara et al. 2016). The algorithm assigns each viral taxon an infection status for each host group of interest: MATCH (affirmatively infects the host), INCONSISTENT (does not infect the host), CONSISTENT (likely infects based on related taxa), or UNCLEAR (ambiguous due to conflicting descendant annotations). Taxa designated as "vertebrate-infecting" for inclusion in the VV database are those with MATCH or CONSISTENT status for the vertebrate host group (NCBI taxid 7742). Implementation details are available in the pipeline repository (github.com/securebio/nao-mgs-workflow).

For viral host composition analysis of Kraken2 results (Figure 4), we applied the same algorithm using an extended set of host categories grouped into six display categories: Vertebrate, Invertebrate (Metazoa minus Vertebrata), Bacteria (bacteriophages), Plant (Viridiplantae), Other (archaea, fungi, other eukaryotes, satellite viruses), and Unknown or ambiguous (metagenome-associated sequences, unclassified entries, and viruses not in the annotation database). Virus-Host DB has limited coverage of environmental bacteriophage lineages, likely resulting in underestimation of bacteriophage abundance.

For downstream analysis and figures (Figure 5, Supplementary Figure S6), we excluded the Dicistroviridae family from vertebrate-infecting virus counts. Although two constituent taxa are annotated as vertebrate-infecting in Virus-Host DB, both appear to represent insect viruses detected in vertebrate feces rather than true vertebrate infections.

Relative abundance and normalization

We report VV abundances using two complementary metrics depending on the analysis context:

Raw relative abundance. For cross-site comparisons of overall VV abundance (Figure 5, Supplementary Figure S6), we calculated raw relative abundance as VV read counts divided by total read pairs in the sequencing library.

PMMoV-normalized abundance. For pathogen time-series analysis (Figures 6 and 7), we normalized VV abundances by PMMoV read counts to account for variation in fecal input concentration, viral recovery efficiency, and sample processing across samples and sites. PMMoV is a plant virus shed in human feces at high and relatively stable concentrations, making it a widely used fecal strength indicator in wastewater surveillance (Wolfe et al. 2021; Maal-Bared et al. 2023; Symonds et al. 2018). This normalization approach is consistent with CDC's NWSS, which reports PMMoV-normalized viral concentrations for ddPCR-based

measurements. PMMoV read counts were obtained from Kraken2 taxonomic profiling (non-rRNA fraction) and scaled to total library read depth (PMMoV_scaled = PMMoV_counts × total_read_pairs / n_reads_profiled) to extrapolate from the subsampled profiling reads to the full library. PMMoV-normalized abundance was then calculated as VV_counts / PMMoV_scaled.

To validate our normalization approach, we compared PMMoV and ToBRFV temporal abundance and their correlation across samples (Supplementary Figure S5). ToBRFV is another plant virus commonly detected in wastewater that has been proposed as an alternative fecal indicator (Rushford et al. 2025). We computed ToBRFV-normalized pathogen abundances using the same methodology for comparison with wastewater PCR data (Supplementary Table S3).

Wastewater PCR data

We compared WW-MGS-derived pathogen abundances with wastewater PCR measurements obtained from the CDC NWSS data portal for sites where matching data was available. The NWSS portal aggregates wastewater PCR data from multiple sources, including state and local health departments, CDC, and WastewaterSCAN (Boehm et al. 2024, 2026) (Supplementary Table S5). Data was downloaded via the CDC Socrata API from the public datasets for SARS-CoV-2 (dataset j9g8-acpt; CDC 2025c), Influenza A (dataset ymmh-divb; CDC 2025a), and RSV (dataset 45cq-cw4i; CDC 2025b). NWSS sewershed_id values were matched to CASPER sampling sites with assistance from sampling partners. Population served values reported in the NWSS data were used to inform population estimates in Table 1. NWSS sewershed IDs are shown in Supplementary Table S3. WastewaterSCAN data are shared under a CC BY-NC 4.0 license.

For Boston, NWSS maintains separate sewershed IDs for the DITP North (743) and South System (752); however, current NWSS data are only available for the combined system (sewershed 742), which we used for all Boston comparisons (Figure 7, Supplementary Figure S8). For Boise, WW-MGS samples represent composites from two treatment plants (sewershed 372 and 374), each reporting separately to NWSS. We used data from sewershed 372 for comparison in Supplementary Figure S8, as it had more complete PMMoV-normalized measurements across all three pathogens.

For each matched site-pathogen combination, we extracted PMMoV-normalized pathogen concentration (reported as pathogen copies per PMMoV copy) as the primary comparison metric where available, with raw pathogen concentration (copies/L) as a fallback for sites without PMMoV normalization (Columbia, Monett and Kansas City Westside WWTP, MO). To ensure consistent comparison, for sites where wastewater PCR data was PMMoV-normalized, we compared PMMoV-normalized WW-MGS abundance; for sites with raw wastewater PCR concentrations, we compared raw WW-MGS relative abundance. Wastewater PCR assays are further detailed in Supplementary Table S5.

Clinical respiratory virus testing data

We compared WW-MGS-derived, PMMoV-normalized pathogen abundances from Boston DITP with clinical respiratory virus testing data from the Massachusetts General Hospital and its

affiliated outpatient health centers. Clinical testing data comprised aggregated weekly test counts (positive and negative) from nucleic acid amplification test (NAAT) assays performed by the Massachusetts General Hospital Clinical Microbiology Laboratory, covering the period from the start of Boston wastewater sequencing through August 31, 2025. The hospital network's catchment overlaps geographically with the sewershed served by Boston DITP.

Clinical respiratory virus testing followed a seasonal algorithm in emergency department (ED) and inpatient (IP) settings: during fall and winter months, symptomatic patients received multiplex molecular testing that included at least influenza, SARS-CoV-2, and RSV; during spring and summer months, SARS-CoV-2-only testing was the default, with multiplex molecular testing reserved for special populations (immunocompromised and pediatric patients being admitted to the hospital). In outpatient (OP) settings, a different order set was used where test selection was at clinician discretion, with embedded language within the order set describing appropriate indications for testing. Testing for SARS-CoV-2, influenza A/B, and RSV was available using this order set; testing for additional respiratory viruses required microbiology laboratory approval. All available multiplex molecular assays across all settings were either cleared or authorized for emergency use by the Food and Drug Administration and had been verified for clinical use by the clinical laboratory.

For our comparison, we applied test filtering that mirrored this seasonal testing algorithm. During fall and winter months, we used positive counts from multiplex molecular assays only, excluding SARS-CoV-2-specific assays to avoid repeat testing of known positives. During spring and summer months, we included SARS-CoV-2-specific assays, as this was the default testing strategy for symptomatic patients in ED and IP settings and the recommended approach in OP settings. Influenza and RSV counts used multiplex molecular assays only year-round, as no single-pathogen assays exist for these pathogens. Test volume and positive rate by pathogen are shown in Supplementary Figure S12.

Clinical data was collected under a waiver of consent as non-identifiable aggregate variables with oversight from Massachusetts General Brigham Institutional Review Board (IRB: 2023P001716).

Time series smoothing and correlation

To enable comparison across data sources with different sampling frequencies, all smoothed time series plots in this manuscript use a consistent methodology. WW-MGS and NWSS time series were aggregated to MMWR (Morbidity and Mortality Weekly Report) epidemiological weeks using the epiweeks Python package (v2.3.0). Within each MMWR week, values were aggregated using the arithmetic mean. Clinical data was already provided as weekly aggregates aligned to MMWR weeks (reported by week start date, Sunday). For all data sources, a 5-week centered moving average was applied to produce smoothed time series. For visualization, smoothed values were plotted at the MMWR week midpoint (Wednesday).

For time series visualizations, NYC Hospital C was excluded entirely due to sparse sampling, the first sample from NYC Hospital B was dropped, and the first sample from each Boston DITP site was dropped to avoid long flat lines in plots caused by isolated early samples.

Spearman correlation coefficients were calculated between MMWR-week-aligned smoothed time series for each site-pathogen and wastewater PCR versus clinical combination, restricted to the overlapping date range where both data sources had measurements.

Comparison with public SRA wastewater sequencing data

To examine the CASPER dataset as a fraction of publicly available wastewater sequencing data (Figure 2d, Supplementary Figure S2a), we queried the NCBI SRA metadata using Google BigQuery (table `nih-sra-datastore.sra.metadata`). We defined two reference categories for comparison:

1. **All wastewater sequencing** (Figure 2d): SRA runs with `organism` = "wastewater metagenome", representing publicly deposited wastewater metagenomic sequencing regardless of sequencing methodology.
2. **Untargeted wastewater sequencing** (Supplementary Figure S2a): A subset of wastewater metagenome runs filtered to `assay_type` IN ("RNA-Seq", "WGS") and `librarysource` IN ("METAGENOMIC", "METATRANSCRIPTOMIC"), representing untargeted shotgun or metatranscriptomic sequencing approaches comparable to the CASPER methodology.

Both filters require `organism` = "wastewater metagenome" and therefore do not capture studies deposited under other organism labels (e.g., "viral metagenome", "metagenomes").

For CASPER data, we queried runs associated with BioProject PRJNA1247874 (this data release) and PRJNA1198001 (the previously published Los Angeles dataset from Grimm, Rothman, et al. 2025). Monthly sequencing volumes were aggregated by sample collection date for each category. SRA metadata queries were performed on 2026-03-04.

To compare CASPER data with prior WW-MGS efforts (Figure 3c, Supplementary Figure S2b), we compiled a non-systematic collection of published studies. For each study with publicly deposited data, we queried the associated NCBI SRA BioProject for sample-level base pair statistics. We reviewed each SRA record and classified samples as untargeted metagenomic sequencing, excluding amplicon or hybrid-capture samples when computing per-sample median sequencing depth and total sequencing effort. For studies without public SRA data, we estimated these metrics from statistics reported in the manuscript text, tables, or supplementary materials, excluding studies where such estimation was not tractable. The full annotated literature review table is provided in Supplementary Table S1.

Data availability

Sequencing data generated by the CASPER consortium included in this release are available on the NCBI SRA under BioProject PRJNA1247874. Prior to public release, human-derived reads were masked using NCBI's Human Read Removal Tool (Katz et al. 2021). Additional deep sequencing data from Los Angeles, collected under CASPER, are available under BioProject PRJNA1198001 (Grimm, Rothman, et al. 2025).

Near real-time visualization of human-infecting virus trends from CASPER sampling sites is available through a public dashboard ("Wastewater Surveillance Dashboards," n.d.), which uses a different bioinformatic pipeline (github.com/dhoconno/nvd; O'Connor et al., 2026) than described here.

Wastewater PCR data used for validation are publicly available through the CDC NWSS data portal: SARS-CoV-2 (dataset j9g8-acpt), Influenza A (dataset ymmh-divb), and RSV (dataset 45cq-cw4i). A subset of these data were generated by WastewaterSCAN (Boehm et al. 2024, 2026) and are shared under a CC BY-NC 4.0 license; see https://data.wastewaterscan.org/about for terms of use.

Tabular data underlying manuscript figures, PMMoV RT-qPCR measurements, and clinical respiratory virus testing data from Massachusetts General Hospital are available in the paper's GitHub repository (github.com/securebio/casper-paper).

Funding

L.J.J. was supported by the Draper Scholar program at The Charles Stark Draper Laboratory. J.K., O.S.H., R.F-O., S.L.G., W.J.B., H.B., D.P.R., K.S., J.D.F., and M.R.M. received support for this work from Coefficient Giving via a gift to SecureBio. C.R., A.T-M., E.E.C., M.C.J., and D.H.O. were supported by Inkfish and Heart of Racing. L.J.J., J.P., and P.C.S. were supported by the CDC Pathogen Genomics Centers of Excellence (contract INTF5104H78W22195346) and a CDC Broad Agency Announcement (contract 75D30123C17983). J.E.L. and G.A. were supported by a subcontract under CDC Broad Agency Announcement contract 75D30123C17983. H.M.S-G. and A.A. were supported in part by the National Institute on Drug Abuse of the National Institutes of Health under Award Number U01DA053941, and by the University of Miami Initiative on Virology and Infectious Disease and SecureBio. R.P. was supported by the Illinois Department of Public Health and the Chicago Department of Public Health. This work used Expanse at the San Diego Supercomputer Center through allocation BIO240238 to J.A.R. from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Acknowledgements

We thank the University of Missouri Genomics Technology Core for sequencing services, technical support in weekly sequencing efforts, and optimization of library preparation protocols. We thank the Broad Clinical Labs at the Broad Institute for sequencing support. We thank the staff of the Riverside Water Quality Control Plant; Gabriel Holguin and Rodney Brees at Inland Empire Utilities Agency Regional Water Recycling Plant No. 1; Steve Rhode and the Massachusetts Water Resources Authority Central Laboratory Violet Team; the operations staff at NYC Health + Hospitals; the Sacramento Area Sewer District and its Environmental Laboratory; and Tami Hansen at the Columbia, MO wastewater treatment plant for assistance with sample collection. We thank Adam Horton and Dolores Sanchez Gonzalez for coordinating

sample shipments from Chicago. Figure 1b was created with BioRender (https://BioRender.com/1o0o1uk).

Prepared by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

The conclusions, opinions, or recommendations in this paper are those of the authors and not of the Illinois Department of Public Health, the Chicago Department of Public Health, or the Centers for Disease Control and Prevention.

Author contributions

Conceptualization: Lennart Justen, Clayton Rushford, Jeff Kaufman, Marc Johnson, Rose Kantor. **Methodology:** Lennart Justen, Clayton Rushford, Olivia Hershey, William Bradshaw, Mike McLaren, Jeff Kaufman, Marc Johnson. **Software:** Lennart Justen, Simon Grimm, William Bradshaw, Harmon Bhasin, Daniel Rice, Katherine Stansifer, Jo Faraguna. **Validation:** Lennart Justen, Clayton Rushford, William Bradshaw, Harmon Bhasin. **Formal analysis:** Lennart Justen. **Investigation:** Lennart Justen, Clayton Rushford, Olivia Hershey, Róisín Floyd-O'Sullivan, Alejandro Tovar-Mendez, Emma Copen. **Resources:** Lennart Justen, Kristen Shelton, Ayaaz Amirali, Sherin Kannoly, Sofia Pesantez, Aiden Stanciu, Iñigo Caballero Quiroga, Leopolda Silvera, Nicole Greenwood, Barbra Bongiovi, Austin Walkins, Ryan Love, Scott Lening, Kaylyn Patterson, Theresa Johnston, Sandra Hernandez, Aymara Benitez, Billie Jo McCarley, Samantha Engelage, Suguna Pillay, Cindy Calender, Brent Herring, Carey Robinson, Monett Wastewater Treatment Plant, Columbia Missouri Wastewater Treatment Plant, Gordon Adams, Jillian Paull, Jamie Devlin, Vamsi Thiriveedhi, Sarah Turbett, Jacob E. Lemieux, John Dennehy, Rachel Poretsky, Jason Rothman, Helena Solo-Gabriele, Jason Vogel, Pardis Sabeti, Jeff Kaufman, Marc Johnson. **Data curation:** Lennart Justen, Clayton Rushford, Róisín Floyd-O'Sullivan, Simon Grimm, Daniel Rice, Katherine Stansifer, Jo Faraguna, Alessandro Zulli, Alejandro Tovar-Mendez, Emma Copen, Aiden Stanciu, Gordon Adams, Jillian Paull, Jamie Devlin, Vamsi Thiriveedhi, Sarah Turbett, John Dennehy, Rachel Poretsky, Jason Rothman, Helena Solo-Gabriele, Jason Vogel. **Writing – original draft:** Lennart Justen, Clayton Rushford, Olivia Hershey. **Writing – review & editing:** Lennart Justen, Clayton Rushford, William Bradshaw, Katherine Stansifer, Michael McLaren, Alessandro Zulli, Daniel Cunningham-Bryant, Gordon Adams, Jillian Paull, Sarah Turbett, Rose Kantor, David O'Connor, John Dennehy, Jason Rothman, Helena Solo-Gabriele, Pardis Sabeti, Jeff Kaufman, Marc Johnson. **Visualization:** Lennart Justen. **Supervision:** Olivia Hershey, William Bradshaw, Daniel Cunningham-Bryant, Sarah Turbett, Jacob E. Lemieux, John Dennehy, Rachel Poretsky, Helena Solo-Gabriele, Pardis Sabeti, Jeff Kaufman, Marc Johnson. **Project administration:** Lennart Justen, Clayton Rushford, Jeff Kaufman, Marc Johnson. **Funding acquisition:** David O'Connor, Jason Rothman, Helena Solo-Gabriele, Pardis Sabeti, Jeff Kaufman, Marc Johnson.

Competing interests

D.H.O. received support for this project from Inkfish and Heart of Racing. D.H.O. is a managing partner of Pathogenuity LLC, a consultancy that advises on topics including environmental monitoring for pathogens. P.C.S. hold several patents related to diagnostic and surveillance

technologies and is a co-founder and equity holder in Delve Biosciences and Lyra Labs, a board member and equity holder in Polaris Genomics, and an equity holder of NextGenJane. P.C.S was formerly a co-founder of Sherlock Biosciences and board member of Danaher Corporation, until December 2024. All potential conflicts are managed in accordance with institutional policy.

Code availability

The NAO Viral Metagenomics Pipeline used to process sequencing data is available at github.com/securebio/nao-mgs-workflow. Code for generating figures and analyses presented in this paper is available at github.com/securebio/casper-paper.

References

Adalja, Amesh A., Matthew Watson, Eric S. Toner, Anita Cicero, and Thomas V. Inglesby. 2018. *The Characteristics of Pandemic Pathogens*. Center for Health Security, Johns Hopkins Bloomberg School of Health. https://www.centerforhealthsecurity.org/our-work/pubs_archive/pubs-pdfs/2018/180510-pandemic-pathogens-report.pdf.

Adams, Carly, Megan Bias, Rory M. Welsh, et al. 2024. "The National Wastewater Surveillance System (NWSS): From Inception to Widespread Coverage, 2020-2022, United States." *The Science of the Total Environment* 924 (171566): 171566.

Adams, Carly, Amy E. Kirby, Megan Bias, et al. 2024. "Detecting Mpox Cases through Wastewater Surveillance - United States, August 2022-May 2023." *MMWR. Morbidity and Mortality Weekly Report* 73 (2): 37–43.

Andrews, S. 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. https://github.com/s-andrews/FastQC.

Balaji, Advait, Bryce Kille, Anthony D. Kappell, et al. 2022. "SeqScreen: Accurate and Sensitive Functional Screening of Pathogenic Sequences via Ensemble Learning." *Genome Biology* 23 (1): 133.

Bellekom, Ben, Catherine Troman, Shannon Fitz, Joyce Odeke Akello, Nicholas C. Grassly, and Alexander G. Shaw. 2026. "Comparison of the Sensitivity of Targeted and Untargeted (metagenomic) Methods for the Detection of Viral Pathogens in Wastewater." *The Science of the Total Environment* 1013 (181333): 181333.

Blin, Kai. 2023. *Ncbi-Genome-Download*. Zenodo. https://doi.org/10.5281/ZENODO.8192486.

Boehm, Alexandria B., Marlene K. Wolfe, Amanda L. Bidwell, et al. 2024. "Human Pathogen Nucleic Acids in Wastewater Solids from 191 Wastewater Treatment Plants in the United States." *Scientific Data* 11 (1): 1141.

Boehm, Alexandria B., Marlene K. Wolfe, Amanda L. Bidwell, et al. 2026. "Pathogen Nucleic Acids Data in Wastewater Solids from 147 Treatment Plants in the United States: 2024-2025." *Data in Brief* 65 (112503): 112503.

Bradshaw, William J., Harmon Bhasin, Simon L. Grimm, et al. 2026. *Securebio/nao-Mgs-Workflow: 3.0.1.9*. Zenodo. https://doi.org/10.5281/ZENODO.18434553.

Bushnell, Brian. 2014. "BBMap: A Fast, Accurate, Splice-Aware Aligner." Paper presented 9th Annual Genomics of Energy & Environment Meeting, Walnut Creek, CA. March 17. https://www.osti.gov/servlets/purl/1241166.

CDC. 2025a. "CDC Wastewater Data for Influenza A." July 30. https://data.cdc.gov/Public-Health-Surveillance/CDC-Wastewater-Data-for-Influenza-A/ymmh-divb.

CDC. 2025b. "CDC Wastewater Data for RSV." July 30. https://data.cdc.gov/Public-Health-Surveillance/CDC-Wastewater-Data-for-RSV/45cq-cw4i.

CDC. 2025c. “CDC Wastewater Data for SARS-CoV-2.” July 29. https://data.cdc.gov/Public-Health-Surveillance/CDC-Wastewater-Data-for-SARS-CoV-2/j9g8-acpt.

Chan, Elana M. G., and Alexandria B. Boehm. 2025. “Respiratory Virus Season Surveillance in the United States Using Wastewater Metrics, 2023-2024.” *ACS ES&T Water* 5 (2): 985–992.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” *Bioinformatics (Oxford, England)* 34 (17): i884–i890.

Chuvochina, Maria, Jan Gerken, Martinique Frentrup, et al. 2025. “SILVA in 2026: A Global Core Biodata Resource for rRNA within the DSMZ Digital Diversity.” *Nucleic Acids Research*, no. gkaf1247 (November): gkaf1247.

Committee on Community Wastewater-based Infectious Disease Surveillance, Water Science and Technology Board, Division on Earth and Life Studies, Board on Population Health and Public Health Practice, Health and Medicine Division, and National Academies of Sciences, Engineering, and Medicine. 2024. “Increasing the Utility of Wastewater-Based Disease Surveillance for Public Health Action: A Phase 2 Report.” Preprint, National Academies Press, December 6. https://doi.org/10.17226/27516.

Concentrating Pathogens from Raw and Primary Wastewater Using the InnovaPrep® Concentrating Pipette, Protocol (Revision E). n.d. InnovaPrep.

Corrin, Tricia, Prakathesh Rabeenthira, Kaitlin M. Young, et al. 2024. “A Scoping Review of Human Pathogens Detected in Untreated Human Wastewater and Sludge.” *Journal of Water and Health* 22 (2): 436–449.

Crook, Oliver M., Kelsey Lane Warmbrod, Greg Lipstein, et al. 2022. “Analysis of the First Genetic Engineering Attribution Challenge.” *Nature Communications* 13 (1): 7374.

Diamond, Megan B., Aparna Keshaviah, Ana I. Bento, et al. 2022. “Wastewater Surveillance of Pathogens Can Inform Public Health Responses.” *Nature Medicine* 28 (10): 1992–1995.

Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. “Nextflow Enables Reproducible Computational Workflows.” *Nature Biotechnology* 35 (4): 316–319.

Duvallet, Claire, Fuqing Wu, Kyle A. McElroy, et al. 2022. “Nationwide Trends in COVID-19 Cases and SARS-CoV-2 RNA Wastewater Concentrations in the United States.” *ACS ES&T Water* 2 (11): 1899–1909.

Edgar, Robert C., Brie Taylor, Victor Lin, et al. 2022. “Petabase-Scale Sequence Alignment Catalyses Viral Discovery.” *Nature* 602 (7895): 142–147.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics (Oxford, England)* 32 (19): 3047–3048.

Fiamenghi, Mateus B., Antonio Pedro Camargo, Iro N. Chasapi, et al. 2025. “Meta-Virus Resource (MetaVR): Expanding the Frontiers of Viral Diversity with 24 Million Uncultivated Virus Genomes.” *Nucleic Acids Research*, no. gkaf1283 (November): gkaf1283.

Friedman, Cindy R., Robert C. Morfino, and Ezra T. Ernst. 2025. “Leveraging a Strategic Public-Private Partnership to Launch an Airport-Based Pathogen Monitoring Program to Detect Emerging Health Threats.” *Emerging Infectious Diseases* 31 (13): 35–38.

Grassly, Nicholas C., Alexander G. Shaw, and Michael Owusu. 2025. “Global Wastewater Surveillance for Pathogens with Pandemic Potential: Opportunities and Challenges.” *The Lancet. Microbe* 6 (1): 100939.

Grimm, Simon L., Jeff T. Kaufman, Daniel P. Rice, Charles Whittaker, William J. Bradshaw, and Michael R. McLaren. 2025. “Inferring the Sensitivity of Wastewater Metagenomic Sequencing for Early Detection of Viruses: A Statistical Modelling Study.” *The Lancet. Microbe*, October 14, 101187.

Grimm, Simon L., Jason A. Rothman, William J. Bradshaw, et al. 2025. “Deep Metatranscriptomic Sequencing Data of Wastewater from Los Angeles, USA, 2023-2024.” *Scientific Data* 13 (1): 158.

Hawaii DOH. 2025. “DOH Monitoring First Wastewater Detection of Measles in Kaua’i County.” News Releases from Department of Health, October 22. https://health.hawaii.gov/news/newsroom/doh-monitoring-first-wastewater-detection-of-measles-in-kaua%ca%bbi-county/.

Hellmér, Maria, Nicklas Paxéus, Lars Magnius, et al. 2014. “Detection of Pathogenic Viruses in Sewage Provided Early Warnings of Hepatitis A Virus and Norovirus Outbreaks.” *Applied and Environmental Microbiology* 80 (21): 6771–6781.

Justen, Lennart J., Simon L. Grimm, Kevin M. Esvelt, and William J. Bradshaw. 2025. “Indoor Air Sampling for Detection of Viral Nucleic Acids.” *Journal of Aerosol Science*, no. 106549 (February): 106549.

Kantor, Rose S., and Minxi Jiang. 2024. “Considerations and Opportunities for Probe Capture Enrichment Sequencing of Emerging Viruses from Wastewater.” *Environmental Science & Technology* 58 (19): 8161–8168.

Karthikeyan, Smruthi, Joshua I. Levy, Peter De Hoff, et al. 2022. “Wastewater Sequencing Reveals Early Cryptic SARS-CoV-2 Variant Transmission.” *Nature* 609 (7925): 101–108.

Katz, Kenneth S., Oleg Shutov, Richard Lapoint, Michael Kimelman, J. Rodney Brister, and Christopher O’Sullivan. 2021. “STAT: A Fast, Scalable, MinHash-Based K-Mer Tool to Assess Sequence Read Archive next-Generation Sequence Submissions.” *Genome Biology* 22 (1): 270.

Kaufman, Jeff. 2024. “Detecting Genetically Engineered Viruses With Metagenomic Sequencing.” June 27. https://naobservatory.org/blog/detecting-genetically-engineered-viruses.

Kaufman, Jeff. 2025. “How Much Data From a Sequencing Run?” SecureBio, June 25. https://data.securebio.org/jefftk-notebook/how-much-data-from-a-sequencing-run.

Kidd, Michael, Alex Richter, Angus Best, et al. 2021. “S-Variant SARS-CoV-2 Lineage B1.1.7 Is Associated with Significantly Higher Viral Load in Samples Tested by TaqPath Polymerase Chain Reaction.” *The Journal of Infectious Diseases* 223 (10): 1666–1670.

Kirby, Amy E., Rory M. Welsh, Zachary A. Marsh, et al. 2022. “Notes from the Field: Early

Evidence of the SARS-CoV-2 B.1.1.529 (omicron) Variant in Community Wastewater - United States, November-December 2021." *MMWR. Morbidity and Mortality Weekly Report* 71 (3): 103–105.

Lambisia, Arnold W., Khadija S. Mohammed, Timothy O. Makori, et al. 2022. "Optimization of the SARS-CoV-2 ARTIC Network V4 Primers and Whole Genome Sequencing Protocol." *Frontiers in Medicine* 9 (February): 836728.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–359.

Levy, Joshua I., Praneeth Gangavarapu, Dylan A. Pilz, et al. 2025. "Real-Time, Multi-Pathogen Wastewater Genomic Surveillance with Freyja 2." In *medRxiv*. July 27. https://doi.org/10.1101/2025.07.26.25332245.

Li, H. 2013. *Seqtk*. https://github.com/lh3/seqtk.

Liu, Ollie, Sami Jaghouar, Johannes Hagemann, et al. 2025. "METAGENE-1: Metagenomic Foundation Model for Pandemic Monitoring." In *arXiv [q-bio.GN]*. January 3. arXiv. http://arxiv.org/abs/2501.02045.

Luebbert, Laura, Delaney K. Sullivan, Maria Carilli, et al. 2025. "Detection of Viral Sequences at Single-Cell Resolution Identifies Novel Viruses Associated with Host Gene Expression Changes." *Nature Biotechnology*, ahead of print, April 22. https://doi.org/10.1038/s41587-025-02614-y.

Maal-Bared, Rasha, Yuanyuan Qiu, Qiaozhi Li, et al. 2023. "Does Normalization of SARS-CoV-2 Concentrations by Pepper Mild Mottle Virus Improve Correlations and Lead Time between Wastewater Surveillance and Clinical Data in Alberta (Canada): Comparing Twelve SARS-CoV-2 Normalization Approaches." *The Science of the Total Environment* 856 (Pt 1): 158964.

Machtinger, Ari N., Olivia S. Hershey, William J. Bradshaw, Daniel P. Rice, and Michael R. McLaren. 2025. "Concentration and Nucleic Acid Extraction of Viruses from Wastewater Influent." Protocols.io, May 26. https://www.protocols.io/view/concentration-and-nucleic-acid-extraction-of-virus-gzzhbx737.

McLaren, Michael R., Olivia S. Hershey, Ari N. Machtinger, et al. 2026. "Metagenomic Sequencing of Composite Airplane Wastewater for Surveillance of Emerging Viruses." In *medRxiv*. MedRxiv, January 30. https://doi.org/10.64898/2026.01.29.26343714.

Mihara, Tomoko, Yosuke Nishimura, Yugo Shimizu, et al. 2016. "Linking Virus Genomes with Host Taxonomy." *Viruses* 8 (3): 66.

Minor, Nicholas R., Mitchell D. Ramuta, Miranda R. Stauss, et al. 2023. "Metagenomic Sequencing Detects Human Respiratory and Enteric Viruses in Air Samples Collected from Congregate Settings." *Scientific Reports* 13 (1): 21398.

Missouri DHSS. 2025. "Missouri Health Officials Urge Vigilance as West Nile Virus Activity Increases in 2025." October 8. https://content.govdelivery.com/accounts/MODHSS/bulletins/3f64540.

Morfino, Robert C., Stephen M. Bart, Andrew Franklin, et al. 2023. "Notes from the Field: Aircraft Wastewater Surveillance for Early Detection of SARS-CoV-2 Variants - John F. Kennedy

International Airport, New York City, August-September 2022." *MMWR. Morbidity and Mortality Weekly Report* 72 (8): 210–211.

Neri, Uri, Yuri I. Wolf, Simon Roux, et al. 2022. "Expansion of the Global RNA Virome Reveals Diverse Clades of Bacteriophages." *Cell* 185 (21): 4023–4037.e18.

O'Connor, David, Nick Minor, and William Gardner. 2026. *Dhoconno/nvd*. Zenodo. https://doi.org/10.5281/ZENODO.18319592.

Peccia, Jordan, Alessandro Zulli, Doug E. Brackney, et al. 2020. "Measurement of SARS-CoV-2 RNA in Wastewater Tracks Community Infection Dynamics." *Nature Biotechnology* 38 (10): 1164–1167.

Public Health England. 2020. *Investigation of Novel SARS-COV-2 Variant: Variant of Concern 202012/01*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/959438/Technical_Briefing_VOC_SH_NJL2_SH2.pdf.

Rebecca Falender, Dvm*, Md* Melissa Sutton, Paul Cieslak, et al. 2026. "Notes from the Field: Retrospective Analysis of Wild-Type Measles Virus in Wastewater during a Measles Outbreak — Oregon, March 24–September 22, 2024." *MMWR. Morbidity and Mortality Weekly Report* 75. https://doi.org/10.15585/mmwr.mm7502a1.

Renfro, Zachary T., Alessandro Zulli, Julie Parsonnet, Alexandria Boehm, and Christopher L. Bennett. 2026. "Flush with Data: Harnessing Emergency Department Wastewater as an Innovative Approach for Surveillance of Infectious Diseases." *American Journal of Epidemiology*, no. kwag019 (January): kwag019.

"Rhinovirus Serotypes Dashboard." n.d. Accessed December 16, 2025. https://dholab.github.io/public_viz/004-rhinovirus-serotypes-dashboards/.

Robinson, Carolyn A., Hsin-Yeh Hsieh, Shu-Yu Hsu, et al. 2022. "Defining Biological and Biophysical Properties of SARS-CoV-2 Genetic Material in Wastewater." *The Science of the Total Environment* 807 (Pt 1): 150786.

Rothman, Jason A., Theresa B. Loveless, Joseph Kapcia 3rd, et al. 2021. "RNA Viromics of Southern California Wastewater and Detection of SARS-CoV-2 Single-Nucleotide Variants." *Applied and Environmental Microbiology* 87 (23): e0144821.

Rushford, Clayton, Devon A. Gregory, Emma E. Copen, et al. 2025. "Untargeted Longitudinal Ultra Deep Metagenomic Sequencing of Wastewater Provides a Comprehensive Readout of Expected and Unexpected Viral Pathogens." In *medRxiv*. October 28. https://doi.org/10.1101/2025.10.27.25338874.

Schuele, Leonard, Leandre Murhula Masirika, Jean Claude Udahemuka, et al. 2024. "Real-Time PCR Assay to Detect the Novel Clade Ib Monkeypox Virus, September 2023 to May 2024." *Euro Surveillance : Bulletin Europeen Sur Les Maladies Transmissibles [Euro Surveillance : European Communicable Disease Bulletin]* 29 (32): 2400486.

Sikorski, Michael J., and Myron M. Levine. 2020. "Reviving the 'Moore Swab': A Classic Environmental Surveillance Tool Involving Filtration of Flowing Surface Water and Sewage Water to Recover Typhoidal Salmonella Bacteria." *Applied and Environmental Microbiology* 86 (13): e00060–20.

Spurbeck, Rachel R., Lindsay A. Catlin, Chiranjit Mukherjee, Anthony K. Smith, and Angela Minard-Smith. 2023. “Analysis of Metatranscriptomic Methods to Enable Wastewater-Based Biosurveillance of All Infectious Diseases.” *Frontiers in Public Health* 11 (March). https://doi.org/10.3389/fpubh.2023.114527510.3389/fpubh.2023.1145275.s001.

St-Onge, Guillaume, Jessica T. Davis, Laurent Hébert-Dufresne, et al. 2025. “Pandemic Monitoring with Global Aircraft-Based Wastewater Surveillance Networks.” *Nature Medicine* 31 (3): 788–796.

Symonds, E. M., Karena H. Nguyen, V. J. Harwood, and M. Breitbart. 2018. “Pepper Mild Mottle Virus: A Plant Pathogen with a Greater Purpose in (waste)water Treatment Development and Public Health Management.” *Water Research* 144 (November): 1–12.

Tay, Aidan P., Kieran Didi, Anuradha Wickramarachchi, Denis C. Bauer, Laurence O. W. Wilson, and Maciej Maselko. 2024. “Synsor: A Tool for Alignment-Free Detection of Engineered DNA Sequences.” *Frontiers in Bioengineering and Biotechnology* 12 (July): 1375626.

The Nucleic Acid Observatory Consortium. 2021. “A Global Nucleic Acid Observatory for Biodefense and Planetary Health.” In *arXiv [q-bio.GN]*. August 5. arXiv. http://arxiv.org/abs/2108.02678.

Tisza, Michael, Sara Javornik Cregeen, Vasanthi Avadhanula, et al. 2023. “Wastewater Sequencing Reveals Community and Variant Dynamics of the Collective Human Virome.” *Nature Communications* 14 (1): 6878.

WastewaterSCAN. n.d. “EVD68, California.” WastewaterSCAN Dashboard. Accessed December 31, 2025. https://data.wastewaterscan.org/tracker?charts=CvYDEAAgAUgAUgY3NGQwMjZSBjA3Y2VkN1IGYjcyM2FlUgYzMDMyYzhSBjI0ZDAzMFIGMTEwN2I2UgY2NTdlYTVSBjY0NzNjMFIGN2NhOTA1UgYxZDYxNDlSBjMyNDllZIIGNWFhNjk5UgYzYTNmYTFSBjc0YjM5YVIGZmEyZDYzUgYwMzI1ZGRSBmI5MjVlN1IGMWE0M2E3UgZhMGZiY2RSBjI1NDgxOVIGNTc1NzM4UgY4YTliNGJSBjI3ODQ3MFIGY2RjZWFkUgYwMmQyNDJSBjcxYTJmNFIGYmNiYjg1UgZiYzc5ZjlSBmU5ZTg3ZVIGZWVjMmMyUgYyZTA3ZTZSBjU3OWRhM1IGYTEwODZmUgY3YWExMzZSBmMyMjE5N1IGYjljMDJkUgY2MjkyMmZSBjM3MzcwMlIGZGRlODhmUgYzMTlhYTZSBjg0ZDQ4OVIGMGM4MDkxUgZhZDg2YTlSBmM4ZDM1N1IGMzE3NDU0UgZlZDlhMWZSBmE4MmNlOVIGY2IwZTFjUgYzNzQzMGFSBjY3YzJlYlIGZGQzNmZiUgZmZWQ5N2ZSBjE4OTFlMFIGZWRlZmI0UgYyOTNiMjVSBjYwOWZjOFIGYzk1ZTY0WgVFVkQ2OHIKMjAyNS0wNy0yNXIKMjAyNS0xMi0yOloBBjI0Y2RmMA%3D%3D&selectedChartId=24cdf0.

“Wastewater Surveillance Dashboards.” n.d. Accessed December 1, 2025. https://dholab.github.io/public_viz/001-make-by-city-and-by-virus-dashboards/index.html.

Wetterstrand, K. A. 2019. “DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP).” Genome.gov, NHGRI, March 13. https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Wittmann, Bruce J., Tessa Alexanian, Craig Bartling, et al. 2025. “Strengthening Nucleic Acid Biosecurity Screening against Generative Protein Design Tools.” *Science (New York, N.Y.)* 390 (6768): 82–87.

Wolfe, Marlene K., Dorothea Duong, Bridgette Hughes, Vikram Chan-Herur, Bradley J. White, and Alexandria B. Boehm. 2022. “Detection of Monkeypox Viral DNA in a Routine

Wastewater Monitoring Program." In *medRxiv*. MedRxiv, July 26. https://doi.org/10.1101/2022.07.25.22278043.

Wolfe, Marlene K., Aaron Topol, Alisha Knudson, et al. 2021. "High-Frequency, High-Throughput Quantification of SARS-CoV-2 RNA in Wastewater Settled Solids at Eight Publicly Owned Treatment Works in Northern California Shows Strong Association with COVID-19 Incidence." *mSystems* 6 (5): e0082921.

Wongprommoon, Arin, Chalita Chomkatekaew, and Claire Chewapreecha. 2024. "Monitoring Pathogens in Wastewater." *Nature Reviews. Microbiology* 22 (5): 261.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20 (1): 257.

Worp, Nathalie, David F. Nieuwenhuijse, Ray W. Izquierdo-Lara, et al. 2025. "Unveiling the Global Urban Virome through Wastewater Metagenomics." *Nature Communications* 16 (1): 10707.

Wu, Fuqing, Jeremiah Oghuan, Anna Gitter, Kristina D. Mena, and Eric L. Brown. 2023. "Wide Mismatches in the Sequences of Primers and Probes for Monkeypox Virus Diagnostic Assays." *Journal of Medical Virology* 95 (1): e28395.

Zulli, Alessandro, Kate R. Bowie, Madelena Ruedaflores, et al. 2025. "Probe-Based Enrichment Sequencing Applied to Wastewater Surveillance Accurately Tracks Multiple Viral Respiratory Outbreaks." *ACS ES&T Water* 5 (8): 4575–4583.