SCCWRP #1479

scientific data

OPEN

DATA DESCRIPTOR

Deep metatranscriptomic sequencing data of wastewater from Los Angeles, USA, 2023–2024

Simon L. Grimm[1,2], Jason A. Rothman[3] ✉, William J. Bradshaw[1,2], Kylie Langlois[4], Joshua A. Steele[4], John F. Griffith[4], Jeff T. Kaufman[1,2] ✉ & Katrine L. Whiteson[5]

Wastewater monitoring for pathogen detection has greatly advanced over the course of the COVID-19 pandemic. While most wastewater surveillance programs only target specific pathogens using qPCR or amplicon sequencing, untargeted wastewater metatranscriptomic sequencing (W-MTS) offers broader detection capabilities. However, there is a lack of data allowing the comparison of W-MTS with more established detection methods. Here we present a dataset consisting of 13.1 terabases (43B read pairs) of untargeted Illumina W-MTS data, generated from 20 wastewater samples, with 1.4B to 2.8B 150 bp read pairs per sample. Wastewater samples were collected between December 2023 and April 2024 at the Hyperion Water Reclamation Plant (HWRP), Los Angeles, USA, serving a population of approximately 4 million residents. The resulting dataset, one of the largest W-MTS collections to date, contains bacterial, archaeal, eukaryotic, and viral taxa—including human-infecting viruses—and many sequences of unknown origin. Uploaded to the NCBI Sequence Read Archive, we expect this data to spur additional research into the viability of pathogen-agnostic wastewater epidemiology and pathogen early detection.

Background & Summary

Infectious disease outbreaks are a major cause of mortality and morbidity worldwide, both through pandemics, such as COVID-19 or 1918 influenza[1], and continuous endemic spread of pathogens like seasonal influenza[2] or respiratory syncytial virus[3]. To mitigate the impact of infectious diseases, researchers are working on ways to identify new pathogens more quickly. One type of early warning method which has seen considerable development in recent years is wastewater monitoring[1].

Many pathogens have previously been identified in wastewater, including respiratory[4], gastrointestinal[5] and other viruses[6,7] as well as pathogenic protozoa[8] and bacteria[9]. Spurred on by the COVID-19 pandemic, wastewater monitoring is now deployed across high-, middle-, and low-income countries[10], identifying the spread of pathogens like H5N1 or mpox. Most of these programs use targeted assays like qPCR, or targeted sequencing methods such as Sanger or amplicon sequencing[11]. Though useful, these methods will be of little help in detecting novel pathogens. Untargeted approaches such as wastewater metagenomic and metatranscriptomic sequencing (W-MGS/W-MTS) could, in principle, detect any pathogen, even those we have never seen before. W-MTS might be particularly promising, given its potential advantage in identifying novel RNA viruses, which pose the greatest pandemic risk[12].

While a number of valuable W-MTS datasets have been made public[13–16], the amount of available data is far less than that of W-MGS[17–22]. Furthermore, many larger W-MTS datasets were generated during the COVID-19 pandemic, when public health measures suppressed the circulation of most other RNA viruses. New W-MTS data from the post-pandemic period is thus needed to properly evaluate this new approach to pathogen early detection.

Here we present a new W-MTS dataset comprising 43B read pairs (2 × 150 bp Illumina sequencing), 5–100 times larger than previously released W-MTS datasets[13,14,16,23,24]. The 20 included influent wastewater samples were collected at the Hyperion Water Reclamation Plant, Los Angeles, California, USA, between December 2023

[1]Media Laboratory, Massachusetts Institute of Technology, Cambridge, USA. [2]SecureBio, Cambridge, USA. [3]Department of Microbiology and Plant Pathology, University of California, Riverside, Riverside, CA, USA. [4]Southern California Coastal Water Research Project, Costa Mesa, CA, USA. [5]Department of Molecular Biology and Biochemistry, University of California, Irvine, Irvine, CA, USA. ✉e-mail: jason.rothman@ucr.edu; jeff@securebio.org

Sequencing Machine	# Wastewater Samples	Date Range	# Reads	# Bases	GC Content
NovaSeq 6000	12	2023-12-03 to 2024-02-19	24.74B	7.47 T	48.38%
NovaSeq X	8	2024-02-25 to 2024-04-14	18.75B	5.66 T	48.94%

Table 1. Summary information for sequencing runs.

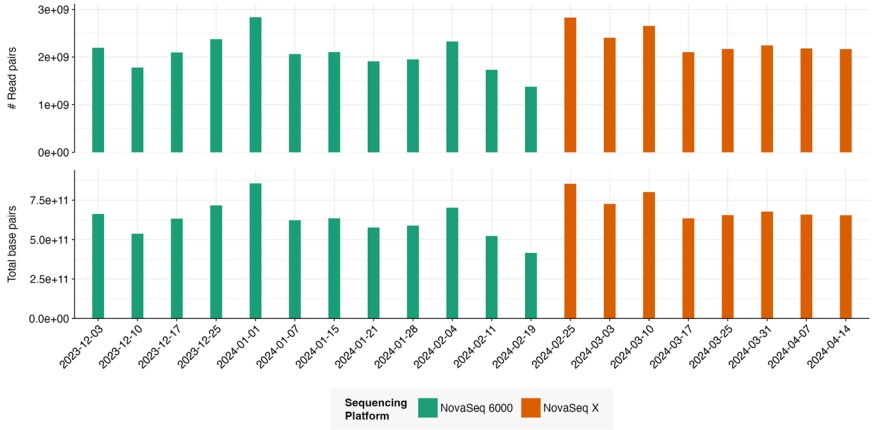

Fig. 1 Read and base counts for different dates and sequencing platforms.

and April 2024, covering most of the 2023–2024 infectious disease season. Sequencing data was generated on a NovaSeq 6000 and a NovaSeq X (Table 1). Sequencing quality was assessed by quantifying adapter presence, QC content, duplication rates, and Phred scores.

Datasets cover a large taxonomic range, with median read fractions across libraries equaling 27.42% (min = 2.16%, max = 62.44%) for bacteria, 0.46% (0.08%, 1.31%) for eukaryotes, 0.05% (0.01%, 0.45%) for archaea, and 1.72% (0.47%, 4.36%) for viruses (Table S2). Deep sequencing and sampling during the peak of the 2023–2024 disease season revealed many human-infecting virus species, including influenza A, SARS-CoV-2, and norovirus. A varying share of reads 72.16% (34.43%, 96.60%) were unclassified when compared against the 2024 Kraken2 standard database, suggesting the extent to which reference databases do not yet capture the large taxonomic diversity of wastewater sequencing data (Table S2).

Methods

Twenty total 1 L 24-h composite influent wastewater samples were collected once weekly at the Hyperion Water Reclamation Plant in Southern California between December 2023 and April 2024 (Table S2). Samples were aliquoted into 250-mL bottles and stored at 4 °C until sample processing[15]. Sixty mL of sample was filtered through a sterile 0.22-μm vacuum filter (VWR, Radnor, PA) to remove solids. The filtrate was then ultrafiltered at 3,000 × g with 10-kDa Amicon filters (MilliporeSigma, Burlington, MA). Samples were centrifuged repeatedly, discarding flowthrough until the full 60-mL sample was processed. This resulted in final volumes <500 μL for each sample, which were then stored at −80 °C until RNA extraction. Wastewater concentrates were thawed on ice. Subsequently, an Invitrogen PureLink RNA minikit with on-column DNase (Invitrogen, Waltham, MA) was used to extract RNA following the manufacturer's protocol. The RNA was subsequently stored at −80 °C.

Sample library preparation and next-generation sequencing were performed by the University of California Genomics Research and Technology Hub (GRT Hub). For library preparation, the GRT Hub used the Illumina RNA prep with enrichment kit (Illumina, San Diego, CA). The GRT Hub then sequenced paired-end libraries with 2 × 150 bp on an Illumina NovaSeq 6000 with an S4 300 cycle kit for samples collected before February 25, 2024, and a NovaSeq X with a 25B, 300 cycle flow cell for all samples collected thereafter. Per wastewater sample, 1.4B to 2.8B read pairs were generated, with an average of 2.2B.

Raw reads were screened for adapter contamination and trimmed of low-quality and low-complexity sequences with FASTP[25]. Paired-end reads were then subset to 1 million read pairs per sample for high-level taxonomic classification using Kraken2 (v2.1.3)[26] using the Standard database[27] (2024-06-05 build). Across the entire dataset, human-infecting viruses were screened using BBDUK (v39.01)[28] and then identified using Kraken2 and Bowtie2[29], Fig. 1.

Data Records

The raw sequencing reads are available on the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1198001[30]. Within the BioProject, each wastewater sample is represented by a BioSample (SAMN45825509 - SAMN45825528), containing relevant metadata (date, origin, sample type). Each BioSample contains one SRA Experiment (SRX27073143 - SRX27073162)[31]. Each SRA Experiment contains two SRA entries (which always represent the same original library, Table S1). The analysis results can be accessed in the following figshare repository: https://doi.org/10.6084/m9.figshare.28454990.v1[32].

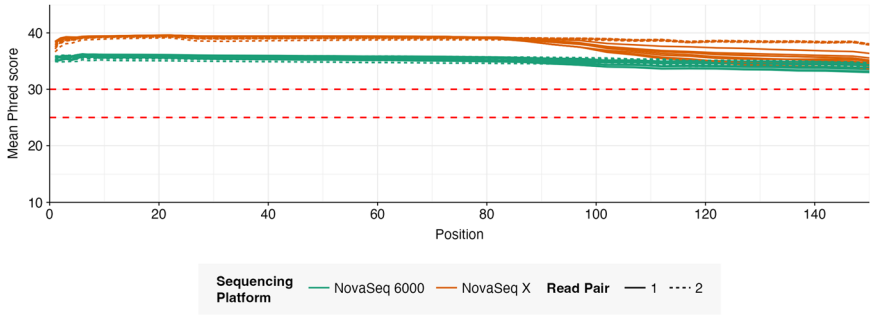

Fig. 2 FastQC-measured per-base quality scores.

Technical Validation

Sequencing quality (per-base mean Phred score, GC content, and adapter presence) was measured with FastQC. Mean Phred scores were consistently above 30 for all sequencing runs, and mean GC content ranged between 44.5% and 48.94%, Fig. 2.

Data availability

All raw metatranscriptomic sequencing data generated in this study are deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1198001[30]. The dataset contains paired-end FASTQ files (2×150 bp) for each wastewater sample, organized into BioSamples SAMN45825509–SAMN45825528 and corresponding SRA Experiments SRX27073143–SRX27073162[31]. Metadata describing sample collection dates, sample type, and sequencing platform are included in each BioSample record. BioSample metadata follows the Genomic Standards Consortium's "MIMS Environmental/Metagenome" metadata standard. The analysis results can be accessed in the following figshare repository: https://doi.org/10.6084/m9.figshare.28454990.v1[32].

Code availability

Sequencing data quality and taxonomic composition was assessed using a comprehensive computational pipeline, available under https://github.com/naobservatory/mgs-workflow/tree/2.5.1. The analysis results can be accessed in the following figshare repository: https://doi.org/10.6084/m9.figshare.28454990.v1[32]. Code for figures and tables can be accessed under https://github.com/naobservatory/w-mgs-data-paper.

Received: 7 May 2025; Accepted: 15 December 2025;
Published online: 24 December 2025

References

1. Dattani, S. & Roser, M. What were the death tolls from pandemics in history? *Our World in Data* (2023).
2. Iuliano, A. D. *et al.* Estimates of global seasonal influenza-associated respiratory mortality: a modelling study. *Lancet* **391**, 1285–1300 (2018).
3. Li, Y. *et al.* Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in children younger than 5 years in 2019: a systematic analysis. *Lancet* **399**, 2047–2064 (2022).
4. Wolfe, M. K. *et al.* 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**, e0082921 (2021).
5. Brinkman, N. E., Fout, G. S. & Keely, S. P. Retrospective Surveillance of Wastewater To Examine Seasonal Dynamics of Enterovirus Infections. *mSphere* **2**, https://doi.org/10.1128/msphere.00099-17 (2017).
6. McCall, C., Wu, H., Miyani, B. & Xagoraraki, I. Identification of multiple potential viral diseases in a large urban center using wastewater surveillance. *Water Res.* **184**, 116160 (2020).
7. Randazza-Pade, J. Pathogen Biomarkers in Wastewater, Stool and Urine. https://biobot.io/pathogen-biomarkers-in-wastewater-stool-and-urine-nearly-endless-opportunities-for-the-future-of-wbe/.
8. Diemert, S. & Yan, T. Clinically unreported salmonellosis outbreak detected via comparative genomic analysis of municipal wastewater *Salmonella* isolates. *Appl. Environ. Microbiol.* **85** (2019).
9. Zhao, Y. *et al.* Strain-level multidrug-resistant pathogenic bacteria in urban wastewater treatment plants: Transmission, source tracking and evolution. *Water Res.* **267**, 122538 (2024).
10. Pronyk, P. M. *et al.* Advancing pathogen genomics in resource-limited settings. *Cell Genom* **3**, 100443 (2023).
11. Keshaviah, A. *et al.* Wastewater monitoring can anchor global disease surveillance systems. *The Lancet Global Health* **11**, e976–e981 (2023).
12. Adalja, A. A., Watson, M., Toner, E. S., Cicero, A. & Inglesby, T. V. *The Characteristics of Pandemic Pathogens*. (2018).
13. Crits-Christoph, A. *et al.* Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *MBio* **12** (2021).
14. Spurbeck, R. R., Catlin, L. A., Mukherjee, C., Smith, A. K. & Minard-Smith, A. Analysis of metatranscriptomic methods to enable wastewater-based biosurveillance of all infectious diseases. *Frontiers in Public Health* **11** (2023).
15. Rothman, J. A. *et al.* RNA Viromics of Southern California Wastewater and Detection of SARS-CoV-2 Single-Nucleotide Variants. *Appl. Environ. Microbiol.* **87**, e01448–21 (2021).
16. Child, H. T. *et al.* Comparison of metagenomic and targeted methods for sequencing human pathogenic viruses from wastewater. *MBio* **14**, e0146823 (2023).
17. Bengtsson-Palme, J. *et al.* Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Sci. Total Environ.* **572**, 697–712 (2016).
18. Brinch, C. *et al.* Long-Term Temporal Stability of the Resistome in Sewage from Copenhagen. *mSystems* **5**, e00841–20 (2020).

19. Langenfeld, K. *et al*. Development of a quantitative metagenomic approach to establish quantitative limits and its application to viruses. *bioRxiv* https://doi.org/10.1101/2022.07.08.499345 (2022).
20. Maritz, J. M., Ten Eyck, T. A., Elizabeth Alter, S. & Carlton, J. M. Patterns of protist diversity associated with raw sewage in New York City. *ISME J.* **13**, 2750–2763 (2019).
21. Munk, P. *et al*. Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nat. Commun.* **13**, 7251 (2022).
22. Ng, C. *et al*. Metagenomic and Resistome Analysis of a Full-Scale Municipal Wastewater Treatment Plant in Singapore Containing Membrane Bioreactors. *Front. Microbiol.* **10** (2019).
23. Rothman, J. A. *et al*. Longitudinal metatranscriptomic sequencing of Southern California wastewater representing 16 million people from August 2020-21 reveals widespread transcription of antibiotic resistance genes. *Water Res.* **229**, 119421 (2023).
24. Wyler, E. *et al*. Pathogen dynamics and discovery of novel viruses and enzymes by deep nucleic acid sequencing of wastewater. *Environ. Int.* **190**, 108875 (2024).
25. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
26. Lu, J. *et al*. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* 1–25 (2022).
27. Index zone by BenLangmead. https://benlangmead.github.io/aws-indexes/k2.
28. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/ (2022).
29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
30. *NCBI BioProject* https://identifiers.org/ncbi/bioproject:PRJNA1198001 (2024).
31. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP551368 (2024).
32. Grimm, S. Output of https://github.com/naobservatory/mgs-workflow/tree/2.5.0, used for Deep wastewater metatranscriptomic sequencing data, Los Angeles, USA, 2023-2024. *figshare* https://doi.org/10.6084/M9.FIGSHARE.28454990.V1 (2025).

Acknowledgements

S.L.G., J.T.K., W.J.B., K.L.W., and J.A.R. were funded for this research project by gifts from Open Philanthropy (to SecureBio). J.A.R was supported by an allocation (#BIO240238) from the National Science Foundation Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) program. K.L.W. and J.A.R. would like to acknowledge earlier support for wastewater monitoring from the University of California Office of the President (award R00RG2814) and the Hewitt Foundation for Biomedical Research. We thank the City of Los Angeles and the Hyperion Wastewater Reclamation Plant for sampling assistance. We also thank Seung-Ah Chung and Melanie Oaks of the University of California Irvine Genomics Research and Technology Hub (GRT Hub), parts of which are supported by NIH grants to the Comprehensive Cancer Center (P30CA-062203) and the UCI Skin Biology Resource Based Center (P30AR075047) at the University of California, Irvine, as well as to the GRT Hub for instrumentation (1S10OD010794-01 and 1S10OD021718-01).

Author contributions

J.T.K. and J.A.R. conceived the study; K.L., J.A.S. and J.F.G. collected wastewater samples; J.A.R. ran sequencing experiments; J.T.K. imported sequencing data, with processing and analysis by S.L.G. and W.J.B. S.L.G. wrote the manuscript, with feedback from all authors. K.L.W. provided study design, project management, and oversight along with manuscript edits.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-06475-7.

Correspondence and requests for materials should be addressed to J.A.R. or J.T.K.