

# Wastewater sequencing reveals community and variant dynamics of the collective human virome

---

Received: 22 May 2023

---

Accepted: 25 September 2023











---

Published online: 28 October 2023

---

 Check for updates

---

Michael Tisza <sup>1,2,11</sup>, Sara Javornik Cregeen<sup>1,2,11</sup>, Vasanthi Avadhanula<sup>2</sup>, Ping Zhang<sup>1,2</sup>, Tulin Ayvaz<sup>1,2</sup>, Karen Feliz<sup>2</sup>, Kristi L. Hoffman<sup>1,2</sup>, Justin R. Clark<sup>2,3</sup>, Austen Terwilliger<sup>2,3</sup>, Matthew C. Ross <sup>1,2</sup>, Juwan Cormier<sup>1,2</sup>, Hannah Moreno<sup>1,2</sup>, Li Wang<sup>1,2</sup>, Katelyn Payne<sup>1,2</sup>, David Henke <sup>2</sup>, Catherine Troisi<sup>4</sup>, Fuqing Wu<sup>4,5,6</sup>, Janelle Rios<sup>4,5</sup>, Jennifer Deegan<sup>4,5</sup>, Blake Hansen <sup>4,5,6</sup>, John Balliew<sup>7</sup>, Anna Gitter<sup>4,5,6</sup>, Kehe Zhang <sup>4,8,9</sup>, Runze Li<sup>4,8,9</sup>, Cici X. Bauer <sup>4,5,8,9</sup>, Kristina D. Mena<sup>4,5,6</sup>, Pedro A. Piedra <sup>2,10</sup>, Joseph F. Petrosino<sup>1,2</sup> , Eric Boerwinkle<sup>4,5,6</sup>  & Anthony W. Maresso<sup>2,3</sup> 

---

Wastewater is a discarded human by-product, but its analysis may help us understand the health of populations. Epidemiologists first analyzed wastewater to track outbreaks of poliovirus decades ago, but so-called wastewater-based epidemiology was reinvigorated to monitor SARS-CoV-2 levels while bypassing the difficulties and pit falls of individual testing. Current approaches overlook the activity of most human viruses and preclude a deeper understanding of human virome community dynamics. Here, we conduct a comprehensive sequencing-based analysis of 363 longitudinal wastewater samples from ten distinct sites in two major cities. Critical to detection is the use of a viral probe capture set targeting thousands of viral species or variants. Over 450 distinct pathogenic viruses from 28 viral families are observed, most of which have never been detected in such samples. Sequencing reads of established pathogens and emerging viruses correlate to clinical data sets of SARS-CoV-2, influenza virus, and monkeypox viruses, outlining the public health utility of this approach. Viral communities are tightly organized by space and time. Finally, the most abundant human viruses yield sequence variant information consistent with regional spread and evolution. We reveal the viral landscape of human wastewater and its potential to improve our understanding of outbreaks, transmission, and its effects on overall population health.

Wastewater-based epidemiology (WBE) refers to the specific detection and tracking of substances<sup>1</sup>, chemicals<sup>2</sup>, genes<sup>3,4</sup>, or pathogens<sup>5</sup> in municipal sewage or sludge to assess population health or disease risk. During the COVID-19 pandemic, the WBE field underwent significant reinvestment<sup>6</sup>, wherein PCR-based detection of SARS-CoV-2 was used

as a proxy for community infection levels, and amplicon sequencing facilitated the resolution of SARS-CoV-2 variants well before clinical detection<sup>7-9</sup>. As such, viral WBE, while initially used for environmental poliovirus surveillance nearly a century ago<sup>10</sup>, has now been leveraged to track influenza virus<sup>11</sup>, respiratory syncytial virus<sup>12</sup>, enterovirus

---

A full list of affiliations appears at the end of the paper.  e-mail: [jpetrosi@bcm.edu](mailto:jpetrosi@bcm.edu); [Eric.Boerwinkle@uth.tmc.edu](mailto:Eric.Boerwinkle@uth.tmc.edu); [maresso@bcm.edu](mailto:maresso@bcm.edu)

D68<sup>13</sup>, and monkeypox virus<sup>14,15</sup> using modern PCR-based methods. Although delivering high sensitivity and specificity, these methods are limited as they cannot provide a comprehensive assessment of human virus levels, community diversity, and variant compositions in a heterogeneous sample.

Recent approaches have investigated the use of virus-like particle enrichment<sup>16</sup>, targeted amplification<sup>8,9</sup>, and/or hybrid capture methods<sup>17–20</sup> to enrich for rare viral sequences amongst the backdrop of what is mostly nucleic acid from bacteria and mammalian hosts. These studies have seen mixed success, not been applied at scale, and the extent to which the levels of virus sequences corresponded to community infection levels is unclear. In any case, even the most prevalent wastewater viruses, e.g., human astroviruses and rotaviruses, comprise a very small fraction of the total biomatter in wastewater, especially compared to other microorganisms, such as bacteria<sup>4,17,21,22</sup>.

Clinical reporting of infectious diseases is extremely valuable for understanding potential sources of outbreaks and disease burden on those with co-morbidities and certain population demographics, but it is constrained by resources, changes in human behavior, and trends in clinical practice. We demonstrate that WBE employing virome sequencing provides insights into aggregate community loads of specific pathogens, viral evolution, dynamics between different viral species or variants, and is presumably agnostic to clinical reporting biases. Specifically, we apply a probe-based capture method accounting for thousands of human and animal viruses followed by deep sequencing to wastewater samples from two major cities whose combined populations reach nearly 3 million people. We reveal the dynamics of the human virome in space and time from hundreds of pathogenic viruses, correlate some of this activity to established detection platforms and clinical data sets, and identify widespread allelic variants of specific viruses for evolutionary tracking.

## Results

### Probe-based capture drives viral enrichment

We developed a comprehensive viral capture approach using a diverse probe set across ten different sites on a weekly basis for nearly 1 year. The probes (TWIST Comprehensive Viral Capture Panel) are directed against a panel of 3153 different human and animal virus genomes. As part of an initiative from the Texas Epidemic Public Health Institute<sup>23</sup>, composite 24-h wastewater influent was collected from six treatment plants in Houston, Texas, USA and four plants El Paso, Texas, USA from May 2022 through February 2023 (Fig. 1A). Wastewater treatment plant catchment areas varied between 10,000 and 400,000 people (estimated 618,148 people served in Houston and 751,982 in El Paso County). These sites were chosen because they allowed us to examine the breadth and robustness of our approach across two large cities with different characteristics. Houston and El Paso also differ in size and diversity, have contrasting climate and rainfall (El Paso dry and Houston humid), are geographically distant (almost 1200 kilometers), and have different patterns of human travel (El Paso a border city with thousands of daily cross-border commuters, Houston a coastal city with one of the largest ports in the world).

The efficacy of probe-based enrichment methods was tested on 18 pilot samples. Following clearance of solids and nucleic acid extraction using methods designed for SARS-CoV-2 detection<sup>24</sup>, we first sequenced and examined viral read numbers from unenriched samples. Low proportions of viral reads were derived from these unenriched samples (4 – 78 aligned reads out of 9.8 – 18.0 million total reads), with 0 to 1 total mammalian viruses detected. In contrast, utilizing the TWIST Comprehensive Viral Research Panel probes on the same extractions, a 3,374-fold enrichment in the proportion of virus reads was observed (Fig. 1F) (14.9 thousand – 407.0 thousand aligned reads out of 11.6 – 24.2 million total reads), with 42 to 128 total mammalian viruses detected).

Read mapping-based virus detection and abundance measurement was conducted using EsVirtu, a bioinformatics tool we developed for this purpose (Fig. S1). EsVirtu leverages sequence information to sensitively detect mammalian viruses and filter out false positives (see materials and methods) (<https://github.com/cmmr/EsVirtu>).

Applying these methods to 363 longitudinal wastewater samples, we detected 28 viral families, 77 genera, 191 species, and 465 distinct virus strains in total (Fig. 1B), with a median of 54 to 98 strains detected per sample, depending on the wastewater treatment plant (Fig. 1C). Furthermore, rarefaction analysis of virus strains showed that the unique detections were not saturated, and additional virus strains are likely to be detected in future samples (Fig. 1D). A median of 28.5 reference genomes or segments had sequencing reads aligning to over 90% of their length with an additional 41 (median) genomes or segments with over 50% alignment (Fig. 1E). From a methodological standpoint, this emphasizes the potential for in-depth analysis of circulating viruses beyond abundance measurements.

To infer the quantitative dynamic range for pathogen detection of this assay, we added in lab-grown respiratory syncytial virus A (RSV) virions to real wastewater samples (samples were previously determined to have no detectable RSV). Based on a stepwise dilution series, we could accurately detect and quantify RSV from a spike-in of 51 genome copies to 4 million genome copies with a Pearson correlation of at least  $R = 0.975$  (Fig. S2).

### Correlation of viral sequencing data with clinical cases

Having established a capture-based approach that offers the prospect of a comprehensive virome analysis of complex wastewater samples, we next asked whether signals generated from sequencing data mirror trends observed from publicly available clinical datasets. Case data from select viral pathogens, namely SARS-CoV-2, influenza virus, and monkeypox virus, were obtained for Houston and, when available El Paso, from local or state government sources. We started first with SARS-CoV-2, as wastewater levels have previously been correlated with case data<sup>25</sup>. Using the reads per kilobase of transcript per million filtered reads (RPKMF) as a proxy for relative virus levels in a given sample, there was a positive correlation between case data and positivity rate for SARS-CoV-2 summer and winter waves and the wastewater signal in Houston (Fig. 2A, S3A, B,  $R = 0.5–0.78$ ) and case data from El Paso (Fig. 2B,  $R = 0.59–0.73$ ). This finding is strengthened by the fact that a second orthogonal technique to measure SARS-CoV-2 levels in wastewater (i.e., qPCR which is the current standard) was also closely correlated with the RPKMF (Fig. S3C, D) for both Houston ( $R = 0.64$ ) and El Paso ( $R = 0.84$ ).

Similarly, Influenza A Virus abundance in the virome sequencing data was highly concordant with reporting of “Weekly Percentage of Visits with Discharge Diagnosed Influenza” in the Houston area (Fig. 2C,  $R = 0.9$ ). Influenza variants H3N2 and H1N1 were also resolved in our data, concordant with clinical subtyping of this flu season in Texas (see Data and Materials Availability). Once more, the virome sequencing data was highly correlated with qPCR measurements from the same samples (Fig. S3E, F,  $R = 0.57–0.73$ ). Finally, a Monkeypox (Mpx) outbreak occurred in the summer of 2022 in several U.S. cities. Rather strikingly, monkeypox virus was detected numerous times at low abundance in Houston wastewater samples (Fig. 2D,  $R = 0.46$ ) in our virome dataset, even though only 1,050 cases were reported in the entire Houston area between July and November 2022. Meanwhile, no detection events of monkeypox virus were recorded from El Paso wastewater samples, consistent with only 10 total reported clinical cases in this metro area.

Encouraging from a detection and possibly public health standpoint, 11 categories of “major” viral pathogens were routinely detected and could be tracked over the sampling period (Fig. 2E), including noroviruses, rotavirus A, hepatitis A virus, RSV, parainfluenza viruses,