Title: Comprehensive Wastewater Sequencing Reveals Community and Variant Dynamics of the Collective Human Virome.

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

Affiliations:

[1] The Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, 77030, USA

[2] Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, 77030, USA

[3] TAILOR Labs, Baylor College of Medicine, Houston, TX, 77030, USA

[4] School of Public Health, University of Texas Health Science Center at Houston, TX, 77030, USA

[5] Texas Epidemiologic Public Health Institute (TEPHI), United States

[6] Department of Epidemiology, Human Genetics and Environmental Sciences, UTHealth Houston School of Public Health, Houston, TX, 77030, USA

[7] El Paso Water Utility, El Paso, TX, USA

[8] Department of Biostatistics and Data Science, UTHealth Houston School of Public Health, Houston, TX, 77030, USA

[9] Center for Spatial-temporal Modeling for Applications in Population Sciences, UTHealth Houston School of Public Health, Houston, TX, 77030, USA

[10] Department of Pediatrics, Baylor College of Medicine, Houston, TX, 77030, USA

† These authors contributed equally

* Corresponding authors

Abstract: Wastewater is a discarded human by-product but analyzing it may help us understand the health of communities. 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. Current approaches overlook the activity of most human viruses and preclude a deeper understanding of human virome community dynamics. We conducted a comprehensive sequencing-based analysis of 363 longitudinal wastewater samples from ten distinct sites in two major cities. Over 450 distinct pathogenic viruses were detected. Sequencing reads of established pathogens and emerging viruses correlated to clinical data sets. Viral communities were tightly organized by space and time. Finally, the most abundant human viruses yielded 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.

One-Sentence Summary: Wastewater contains a trickle of circulating human viruses, and now we can sequence them to better track community transmission.

Main Text:

Wastewater-based epidemiology (WBE) refers to the specific detection and tracking of substances (1), chemicals (2), genes (3, 4), or pathogens (5) in municipal sewage or sludge to assess population health or disease risk. During the COVID-19 pandemic, the WBE field underwent significant reinvestment (6), 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 (7-9). As such, viral WBE, while initially used for environmental poliovirus surveillance nearly a century ago (10), has now been leveraged to track influenza virus (11), respiratory syncytial virus (12), enterovirus D68 (13), and monkeypox virus (14, 15) 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 (16), targeted amplification (8, 9), and/or hybrid capture methods (17-20) 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 (17, 21-23).

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 one year. The probes (TWIST Comprehensive Viral Capture Panel) are directed against a panel of 3,153 different human and animal virus genomes. As part of an initiative from the Texas Epidemic Public Health Institute (24), composite 24-hour 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

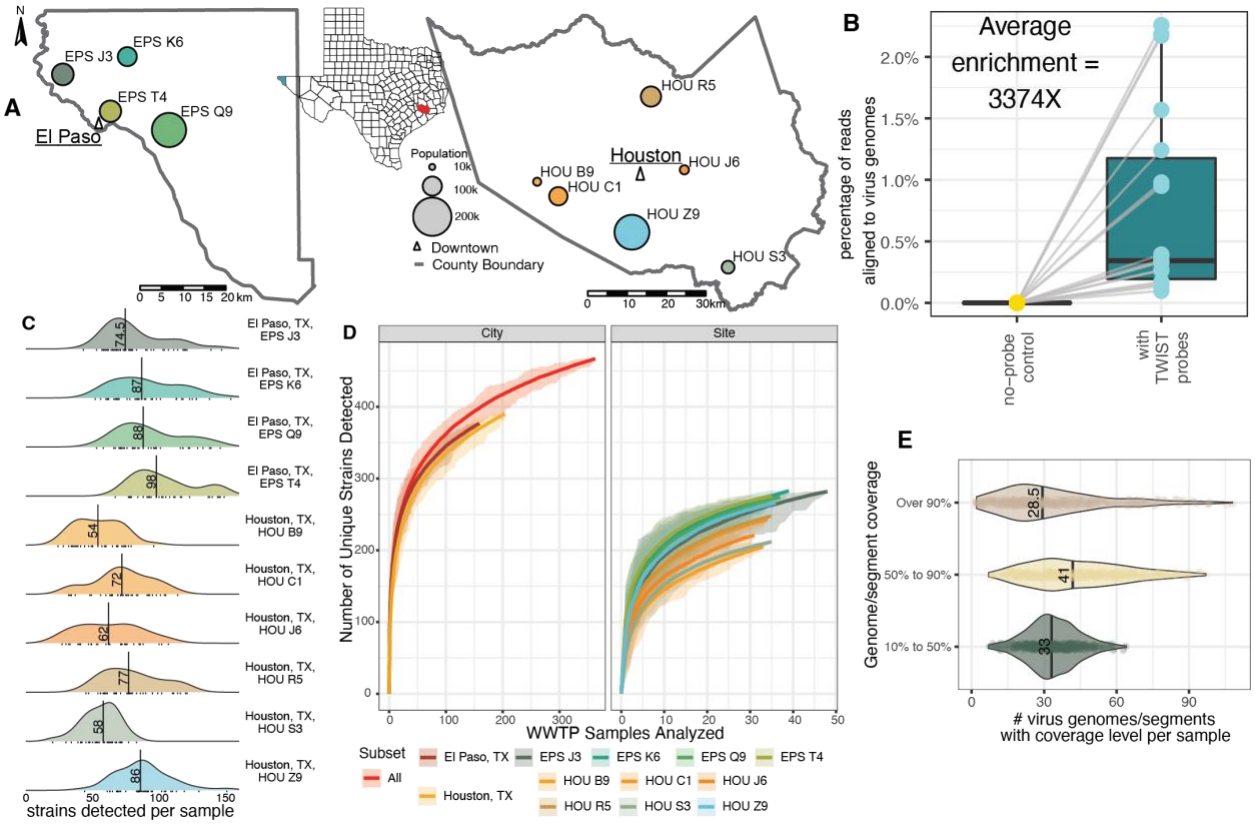
chosen because they allowed us to examine the breadth and robustness of our approach across two large cities with different characteristics. Houston and El Paso also differ in size and diversity, have contrasting climate and rainfall (El Paso dry and Houston humid), are geographically distant (almost 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 (25), 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. 1B) (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 EsViritu, a bioinformatics tool we developed for this purpose (Fig. S1). EsViritu leverages sequence information to sensitively detect mammalian viruses and filter out false positives (see materials and methods) (https://github.com/cmmr/EsViritu).

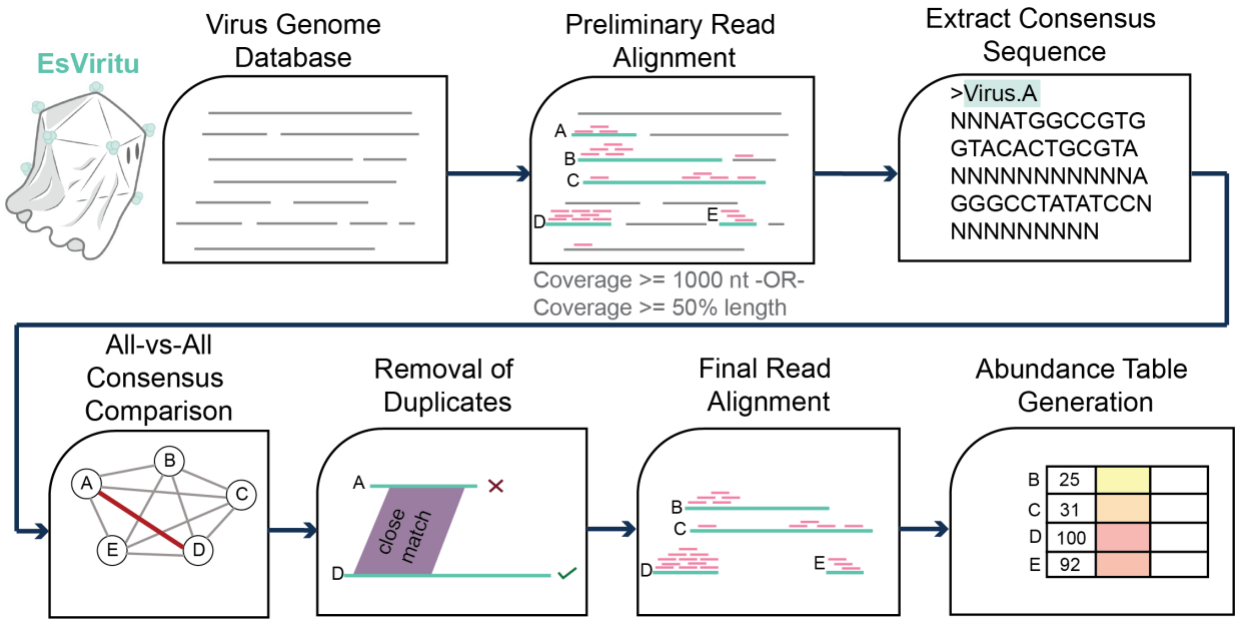
Applying these methods to 363 longitudinal samples, we detected 465 distinct virus strains in one or more samples, with a median of 54 to 98 strains detected per sample, depending on the wastewater treatment plant (Fig. 1C). Furthermore, rarefaction analysis of virus strains showed that the unique detections 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.

Fig. 1



Study sites, capture, and viral diversity. (A) Map of wastewater catchment areas in Houston and El Paso, TX. The colored areas refer to the sites in each city (EP = 4, Houston = 6) (B) Percentage of reads aligned to virus pathogen genome database in paired control (no-probe) and treatment (capture with the TWIST Comprehensive Virus Research Panel) groups. (C) Number of distinct virus strains detected per sample from each wastewater treatment plant. (D) Rarefaction curves measuring distinct virus strains detected as more samples were analyzed. (E) Genome coverage of detected virus genome/segments for each sample.

Fig. S1.



Schematic of EsViritu pipeline. EsViritu is suitable for detecting known and near-known mammalian virus genome sequences in metagenomic data. Relying on a de-replicated database of virus genomes, EsViritu aligns input reads to virus genomes with minimap2/coverM. For genomes that have read coverage of at least 1000 nucleotides or 50% of the sequence length, consensus sequences are extracted with SAMtools. Consensus sequences undergo all-vs-all pairwise comparison with BLASTn. Highly similar consensus sequences are de-replicated into the best (longest) exemplar. Reads are re-aligned to the de-replicated reference sequences with minimap2/coverM to calculate virus abundance metrics.

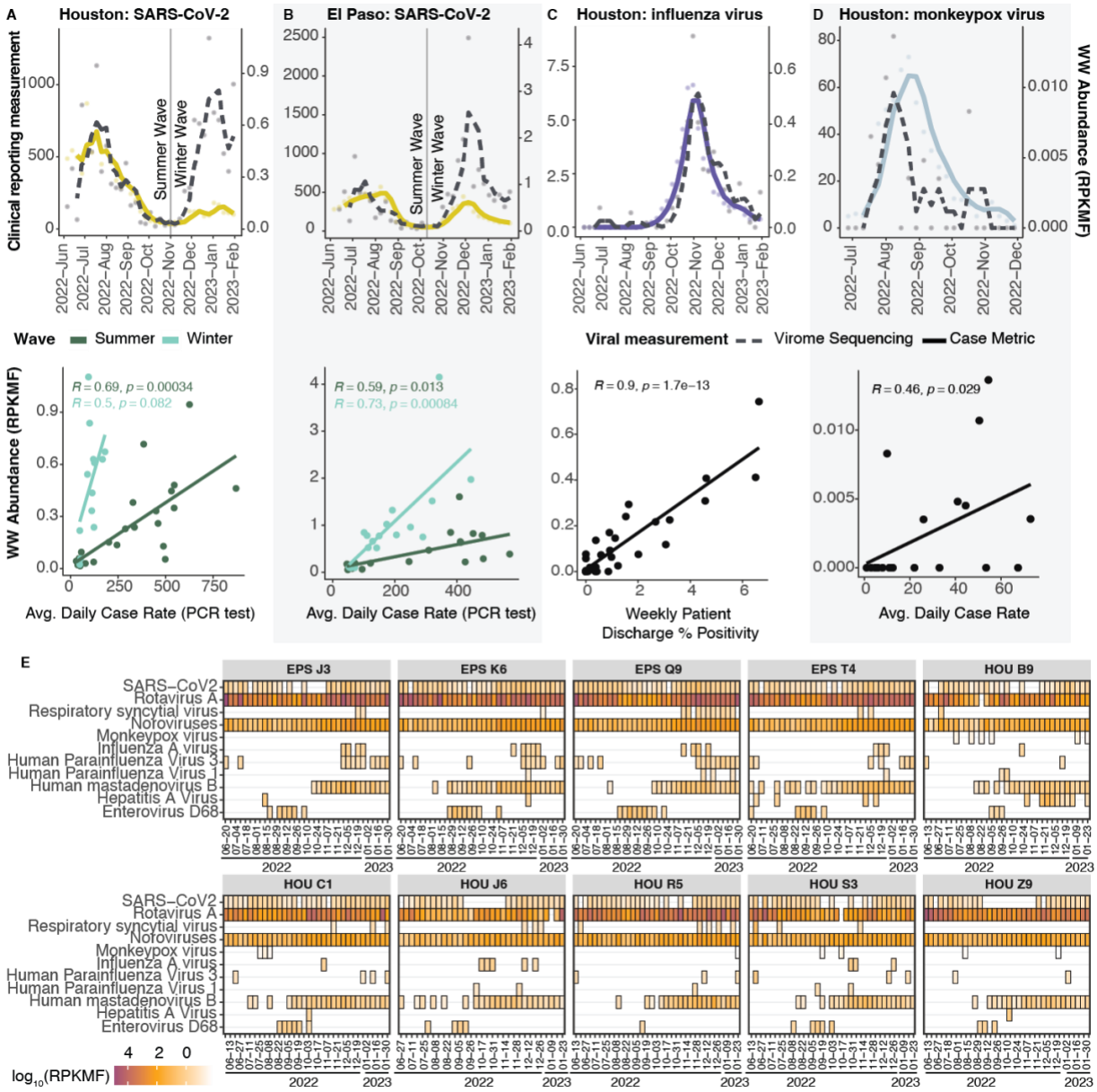
Correlation of Viral Sequencing Data with Clinical Cases:

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 (26). 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, S2A-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-CoV2 levels in wastewater (i.e., qPCR which is the current standard) was also closely correlated with the RPKMF (Fig. S2C-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. S2E-F, R = 0.57 – 0.73). Finally, a Monkeypox (Mpox) 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.

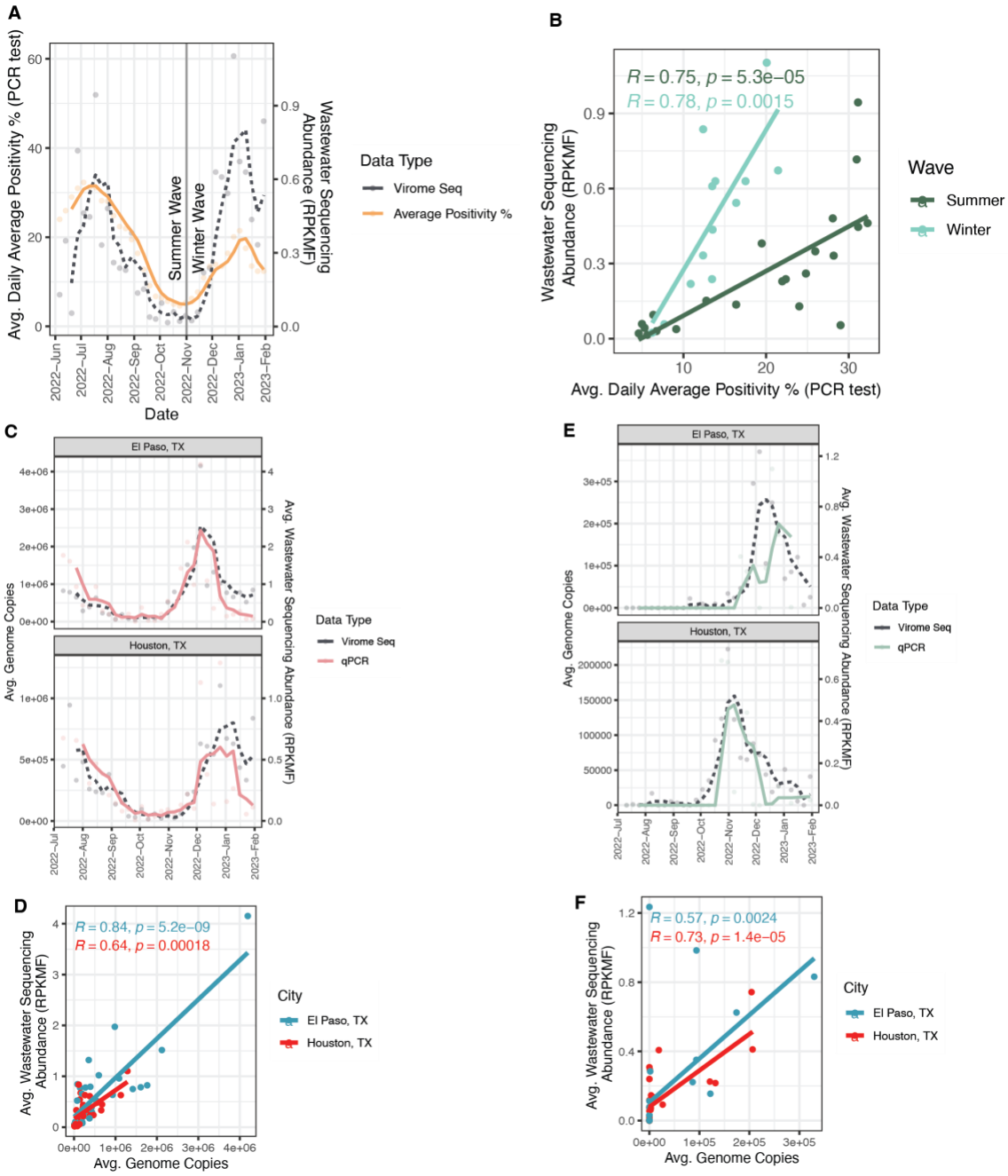
Encouraging from a detection and possibly public health standpoint, 11 categories of "major" viral pathogens were routinely detected and could be tracked over the sampling period (Fig. 2E), including noroviruses, rotavirus A, hepatitis A virus, respiratory syncytial virus (RSV), parainfluenza viruses, and enterovirus D68. Interestingly, at times, there were different trends in virus levels observed in both cities and at different periods of the year (Fig. S3).

Fig. 2



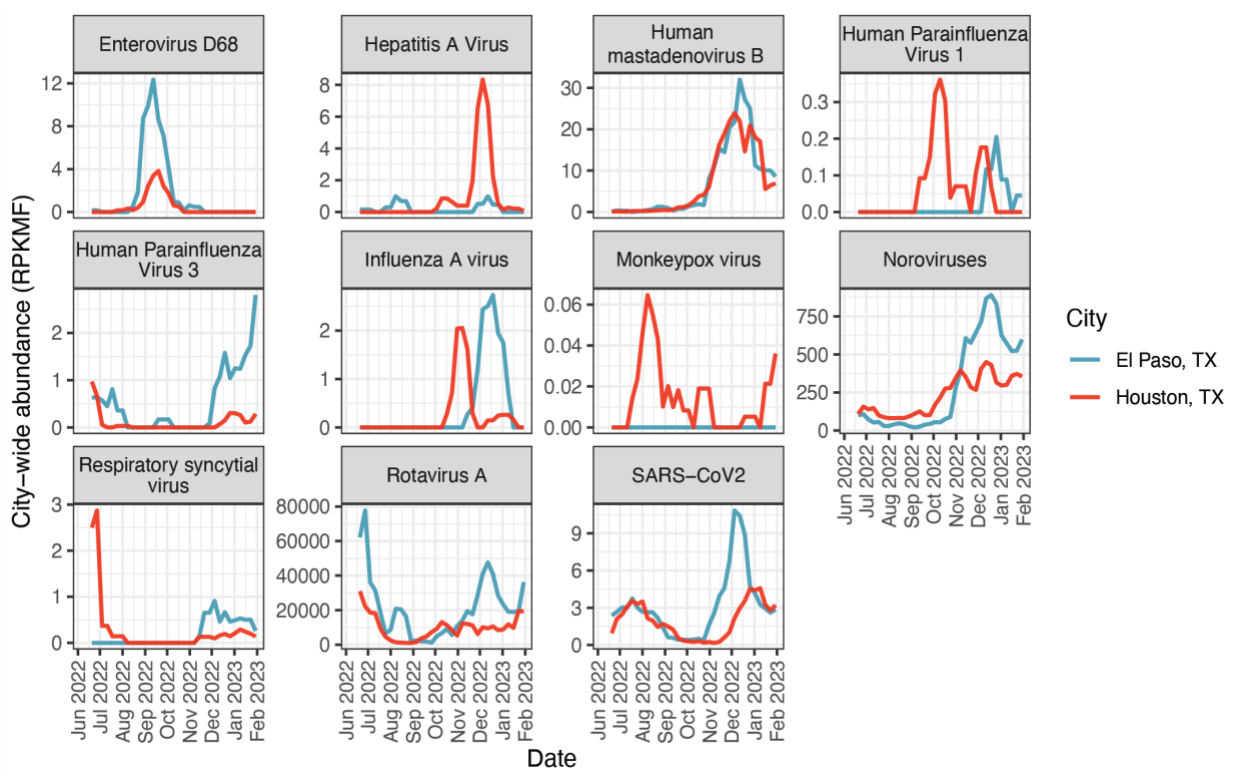
Human viruses in wastewater correlate with clinical data. (A) SARS-CoV-2 wastewater sequencing abundance compared to reported cases (top) and scatter plot with correlation between wastewater sequencing abundance compared to reported cases of SARS-CoV-2 (bottom) in Houston, TX. **(B)** SARS-CoV-2 wastewater sequencing abundance compared to reported cases (top) and scatter plot with correlation between wastewater sequencing abundance compared to reported cases of SARS-CoV-2 (bottom) in El Paso, TX. **(C)** Influenza wastewater sequencing abundance compared to reported Weekly Percentage of Visits with Discharge Diagnosed Influenza (top) and scatter plot with correlation between wastewater sequencing abundance compared to Weekly Percentage of Visits with Discharge Diagnosed of Influenza (bottom) in Houston, TX. **(D)** Monkeypox virus wastewater sequencing abundance compared to reported Mpox cases (top) and scatter plot with correlation between wastewater sequencing abundance compared to reported cases of Mpox (bottom) in Houston, TX. **(E)** Selected pathogen detection and abundance across all 10 wastewater treatment plants.

Fig. S2.



Comparison of virome sequencing to clinical data and qPCR of wastewater samples. (A) SARS-CoV-2 wastewater sequencing abundance compared to percent positivity of PCR tests in Houston, TX. **(B)** Scatter plot and correlation between wastewater sequencing abundance compared to percent positivity of PCR tests of SARS-CoV-2 in Houston, TX. **(C)** SARS-CoV-2 wastewater sequencing abundance compared qPCR values in Houston and El Paso, TX. **(D)** Scatter plot and correlation between wastewater sequencing abundance compared to qPCR values of SARS-CoV-2 in Houston and El Paso, TX. **(E)** Influenza A wastewater sequencing abundance compared qPCR values in Houston and El Paso, TX. **(F)** Scatter plot and correlation between wastewater sequencing abundance compared to qPCR values of Influenza A in Houston and El Paso, TX.

Fig. S3.



Virome sequencing Abundance of Major Pathogens. Moving average charts of eight virus pathogens in Houston and El Paso, TX.

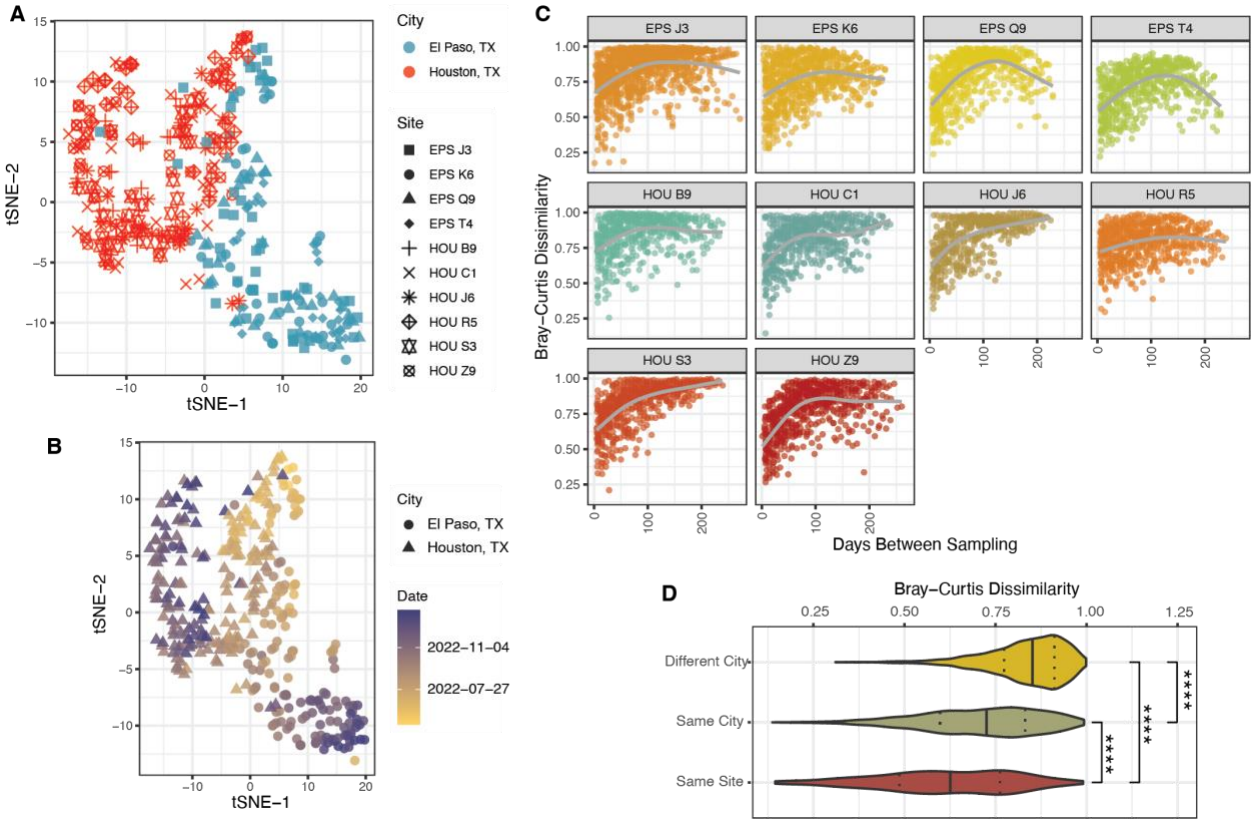
205

Pathogenic virome communities follow spatiotemporal trends:

We wished to understand how the human wastewater virome changed over space and time. Important variables in the structure of virome communities were realized by generating t-distributed stochastic neighbor embedding (t-SNE) plots from the virus abundance data of each sample. There was a stark separation of the samples by the city of collection and date of collection (Fig. 3A-B). Virus species from several families showed an uneven distribution between Houston and El Paso (Fig. S4A). For example, while we expect most viruses to have a prevalence bias towards El Paso due to higher median levels of strain detection per site (Fig. 1C), El Paso had especially strong signals from many Parvoviridae and Sedoreoviridae whereas Houston samples had higher prevalence of many Calicivirdae and Astroviridae, the reasons for which are currently unknown (Fig. S4A).

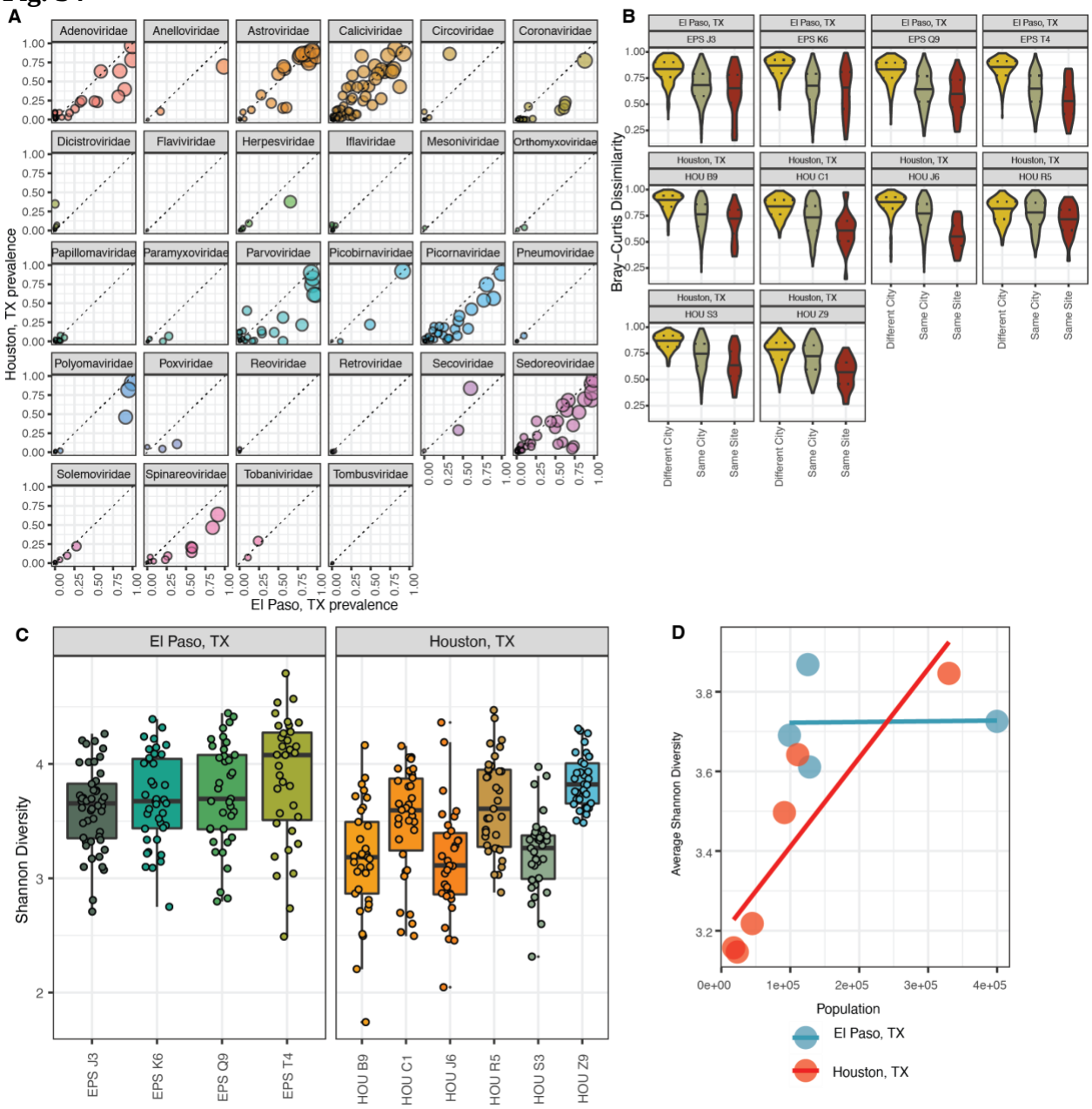
To assess community dynamics over time, all samples from each site were compared to each other using the Bray-Curtis dissimilarity statistic (Fig. 3C). In general, as time went on, the composition of the virome in samples diverged such that samples taken closer in time were quite similar, whereas those separated by many months were very different. Interestingly, a possible exception to the temporal divergence rule can be seen in samples taken from the wastewater treatment plant serving Houston's large intercontinental airport, which likely reflects a transient population of world travelers (Fig. 3C, HOU R5). Here, the compositional dissimilarity was poorly correlated with the passing of time, possibly due to flux of the virome from incoming people. On the other hand, as the data collection approaches one year and the seasons repeat, samples from 3 of 4 El Paso sites seem to be re-converging on their community structures from the previous year. In general, dissimilarity follows a pattern where sites from different cities are more different than sites within the same city, and samples from the same site are more similar than everything else (Fig. 3D, Fig. S4B). Finally, we assessed the impact of human population size on virome diversity. The alpha diversity (Shannon's statistic) was measured for each sample (Fig. S4C), and the average diversity and population of the service area for each site were plotted (Fig. S4D). Average diversity increases from catchment populations of 10,000 to 100,000 inhabitants, but the diversity values level off with greater numbers of people. Collectively, this data confirms that the structure of wastewater virome communities are substantially determined by temporal and geospatial factors.

Fig. 3



Wastewater virome community structure. (A) t-SNE of wastewater samples using virome abundance data, showing different cities/sites. **(B)** t-distributed stochastic neighbor embedding of wastewater samples using virome abundance data, showing samples over time. **(C)** Temporal analysis of intra-site community changes. Each dot is a comparison between two samples. The x-axis measures days in between sampling. The y-axis measures Bray-Curtis dissimilarity between the samples. **(D)** Bray-Curtis dissimilarity between samples taken +/- seven days apart, comparing samples from the same site, different site but same city, and different city. "****" represents p-value < 1e-04.

Fig. S4



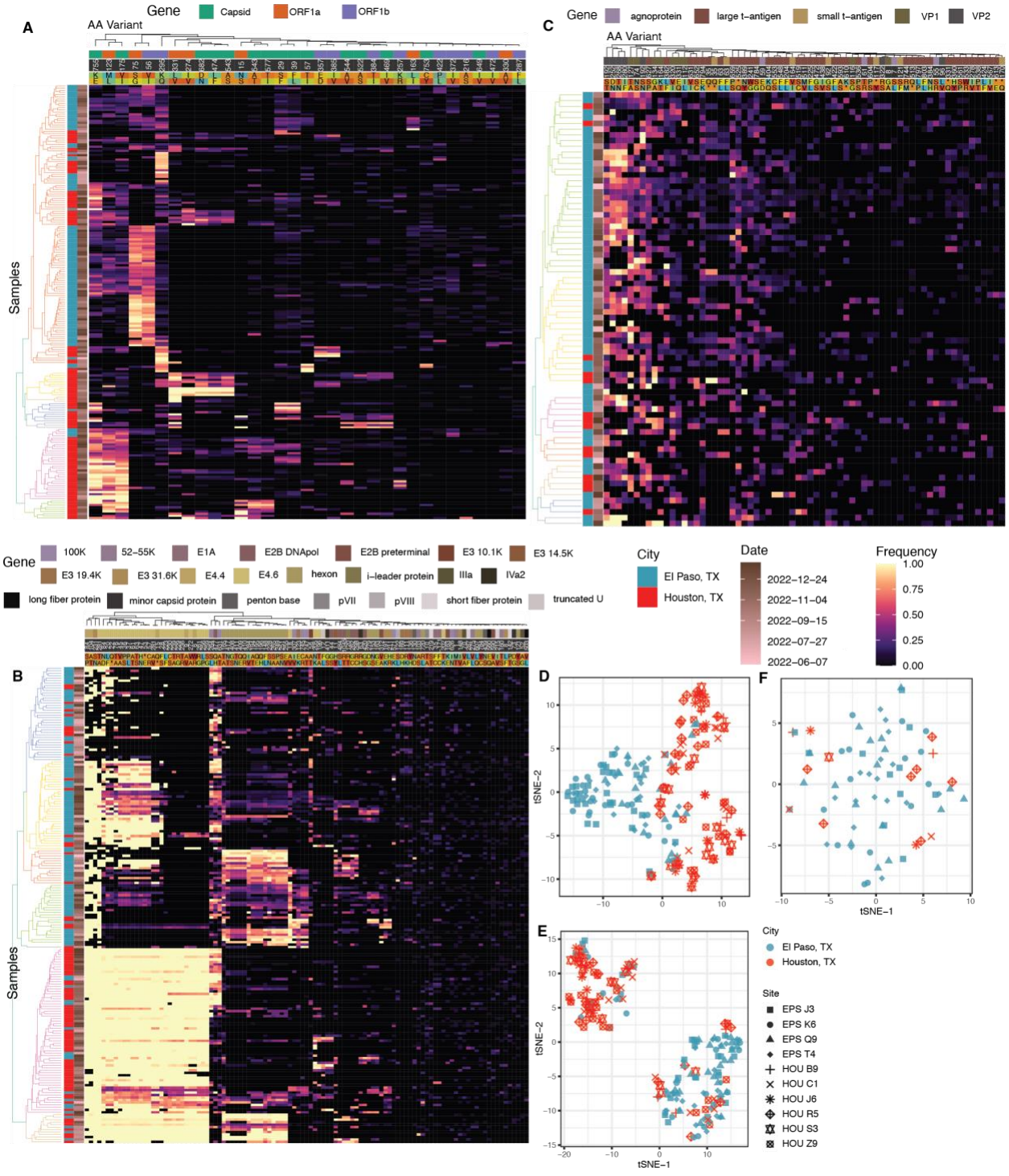
Additional wastewater virome community metrics. (A) All detected virus strains faceted by family with prevalence in Houston and El Paso. **(B)** Bray-Curtis dissimilarity between samples taken +/- seven days apart, comparing samples from the same site, different site but same city, and different city, faceted by site. **(C)** Shannon diversity by site. **(D)** Comparison of average Shannon diversity of wastewater treatment plants and population of catchment areas.

Variant trends amongst the virome backdrop:

A handful of viruses had high or complete genome coverage across many wastewater samples and were therefore suitable for variant analysis. Although a single lineage seemed to dominate the sample read abundance for some virus strains, many samples had a mix of two or more lineages. Therefore, allelic variants were measured by the frequency of non-synonymous mutations compared to the reference genome. We focused on three examples.

Astrovirus MLB1, which has a seroprevalence in Americans close to 100% (27), was the virus contained at high genome coverage in the most samples in our dataset. The variant landscape of Astrovirus MLB1 was largely dictated by the city-of-origin of the sample (Fig. 4D), with gene-specific mutations in the capsid, ORF1a, and ORF1b showing strong regional localization in time (Fig. 4A). Human Adenovirus 41 is an enteric virus associated with diarrhea and, possibly, hepatitis (28) in children and was also quite common in wastewater. This virus splits into two major lineages (Fig. 4B, E), with the hypervariable capsid (hexon) gene having a lot of diversity (29). Although both lineages dominated in samples from either city, each city had one lineage that was more common. JC Polyomavirus, which is secreted in the urine, commonly establishes long, asymptomatic infections in a high proportion of the population (30). Consistent with non-acute, rarely transmitted infections, the variant landscape of this virus seems to lack meaningful spatiotemporal structure and most samples appear to have a diversity of lineages (Fig. 4C, F).

Fig. 4



Evaluation of non-synonymous variants in prevalent wastewater viruses. (A) Dendrogram and heatmap describing frequency of common non-synonymous variants of astrovirus MLB1 by sample. **(B)** Like (A) but with Human Adenovirus 41. **(C)** Like (A) but with JC Polyomavirus. **(D)** t-SNE of non-synonymous variant frequency of astrovirus MLB1. **(E)** Like (D) but with Human Adenovirus 41. **(F)** Like (D) but with JC Polyomavirus.

Discussion:

We present here a major advance in our understanding of the human virome in wastewater. High level findings include (i) the ability of a hybrid-capture probe approach that is designed against over 3,000 human and animal viruses to, when combined with deep sequencing, significantly enrich viral detection (465 total viruses detected in 6 months); (ii) the approach's suitability as a record of aggregate community infection levels for several key viruses of concern, including SARS-CoV-2, influenza virus, enterovirus D68, noroviruses, rotaviruses, monkeypox virus, respiratory syncytial virus and many others; (iii) the clear correlation of genome-specific sequencing reads to clinical data for SARS-CoV-2, influenza virus, and monkeypox virus ; (iv) dynamics of the virome that change across space and time; and finally (v) the variant tracking of viruses with high genome coverage and prevalence.

Although this approach already offers a suitable estimation of relative virus levels, genomes of important pathogens such as SARS-CoV-2, influenza virus, and monkeypox virus were typically not sequenced at high enough coverage to assess allele and variant frequency. This was probably because these viruses do not typically carry high viral loads in the gastrointestinal or urinary tract and are not shed into wastewater in high numbers. Therefore, continued development of the enrichment methodology as well as increased sequencing depth should be considered to improve genome coverage.

The literature suggests that, while some SARS-CoV-2 variants were largely limited to certain geographies, major lineages (e.g., delta, omicron, BA.4) swept across the globe and were primarily separated by time (9), rather than geography. It is not clear whether this pattern observed for highly transmissible SARS-CoV-2 applies to other, less transmissible human viruses. Indeed, the data presented here for three highly abundant viruses suggests that this may not be the case (i.e., that geography is important).

The geographic partitioning between, for instance, allelic variants of astrovirus MLB1 but not JC polyomavirus may be due to transmission and infection modalities. Astroviruses transmit via a fecal-oral mechanism, and therefore are likely to propagate by local outbreaks (31). JC Polyomavirus is known to establish long asymptomatic infections in the large majority of people, probably being transmitted within households sporadically (30). Astrovirus variants may be introduced to an immune-naive population or city and spread outbreak-by-outbreak over the course of months, whereas different JC polyomavirus lineages acquired by different people over many years could be shed simultaneously, blurring allelic patterns seen with other viruses. It is tantalizing to consider that such an approach may help us understand the underlying infection etiology of these and related viruses as it pertains to entire population transmission and dynamics.

The future of wastewater pathogen monitoring is promising, but responsible parties will need to refine methods that provide timely, comprehensive, and useful information (6). This study shows that some of these attributes are already being attained, but additional innovation and interest will be required to get up to speed on all fronts. Post-pandemic society will require orthogonal, comprehensive viral surveillance of distinct pathogen reservoirs, including wastewater, to better equip epidemiologists with accurate information for public health action.

Materials and Methods

Sample Collection and Shipping

Between 100-500 mL of raw wastewater was collected into 500ml leak-proof prelabeled sample bottles at 6 wastewater treatment plants in Houston, TX and 4 in El Paso, TX. Treatment plants were coded upon the request of public health officials. The surface of the sample bottles was decontaminated with 10% bleach and moved to a "clean" zone, where the samples were sealed into biohazard bags in shipping boxes with absorbent pads and ice packs for overnight shipping to the Alkek Center for Metagenomics and Microbiome Research at Baylor College of Medicine, Houston, TX.

Sample Processing and Nucleic Acid Extraction

Wastewater samples were barcoded upon arrival and stored at 4°C until processing. First, 50 mL of wastewater was decanted and centrifuged, separating the solid and liquid fractions. The supernatant was then vacuum filtered using an ion-based cellulose filter paper and the virus-containing cellulose filter was placed into a bead-beating tube with lysis buffer. The tube was run on the homogenizer for 1 minute at 5 m/s, rested for 1 minute, then run on the homogenizer for 1 more minute. Following the bead beating the samples were centrifuged at 14-17x1000 RPM for 2 mins. DNA and RNA were extracted using the Qiagen QIAamp VIRAL RNA Mini Kit.

Library Preparation, probe-based virome capture and sequencing

RNA extracts were converted to cDNA using Protoscript II First Strand cDNA Synthesis Kit (New England Biolabs Inc.), NEBNext Ultra II Non-Directional RNA Second Strand Module (New England Biolabs Inc.), and Random Primer 6 (New England Biolabs Inc.). A total of 25 ng of the cDNA and DNA mix was used for library construction using Twist Library Preparation EF 2.0 Kit and Twist Universal Adaptor System (Twist Biosciences). The libraries were pooled, a maximum of 16 samples per pool, at equal mass to a total 1,500 ng per pool. The Twist Comprehensive Viral Research Panel (Twist Biosciences) was used to hybridize the probes at 70 °C for 16 hours. The post-capture pool was further PCR amplified for 12 cycles and final libraries were sequenced on Illumina NovaSeq 6000 SP flow cell, to generate 2×150 bp paired-end reads. Following sequencing, raw data files in binary base call (BCL) format were converted into FASTQs and demultiplexed based on the dual-index barcodes using the Illumina 'bcl2fastq' software.

RT-PCR of specific pathogens

Real-time RT-PCR for Influenza A and SARS-CoV-2 was performed using 5 µl of eluted RNA and 15 µl of TaqPath 1-step Multiplex Master Mix, (A28523 Applied Biosystems) under the following cycling conditions: 25 °C for 2 minutes, 50 °C for 15 minutes, 95 °C for 2 minutes, and 45 cycles of 95 °C for 3 seconds, then 55 °C for 30 seconds on a 7500 Fast Dx Real-Time PCR Instrument (4406985, Applied Biosystems) with SDS version 1.4 software. Samples were considered positive if Ct values were <40. The real-time RT-PCR included negative extraction and no template controls. A standard curve with amplicons targeting the primers and probe sequences of each target was used to determine the genomic copy numbers for each target. Primer sequences and references are available in Table 1.

Virus Pathogen Database Compilation

The TWIST Comprehensive Virus Research Panel is reported to contain over 1M unique probes that enable detection of 3,153 viral human and nonhuman pathogens. To construct the Virus Pathogen Database (v2.0.2), all available complete isolate genomes for each viral genus covered by the TWIST Comprehensive Virus Research Panel were downloaded from NCBI's GenBank using Datasets (https://www.ncbi.nlm.nih.gov/datasets/) on November 16th, 2022. Ninety seven virus genomes were determined to be of highest public health concern based on the subjective opinion of the authors. These genomes were protected from de-replication and segments were concatenated when relevant (available at https://zenodo.org/record/7876309). All other virus genomes and segments were de-replicated at 95% average nucleotide identity and 85% alignment fraction using two rounds of the anicalc/aniclust scripts from CheckV (32). Exemplar sequences for each cluster were kept in the database (https://zenodo.org/record/7876309). Hierarchical taxonomic data was obtained for each sequence from the kingdom level to the strain level using Taxonkit (33).

Sequencing Read Processing for Virus Genome & Segment Detection

Demultiplexed raw fastq sequences were processed using BBDuk to quality trim (Q25), remove Illumina adapters, and filter PhiX reads. Reads with a minimum average Phred quality score <23 and length < 50 bp after trimming were discarded. Trimmed FASTQs were mapped to a combined PhiX (standard Illumina spike in) and human reference genome database (GCF_000001405.39) using BBMap to determine and remove human/PhiX reads (34). Processed reads were run through the EsViriru pipeline (v0.1.1) with default settings. Specifically, reads were aligned to the entire Virus Pathogen Database (v2.0.2) via minimap2 (35) and alignments were filtered by CoverM (https://github.com/wwood/CoverM) to require at least 90% average identity across 90% of the read length. Virus genomes/segments with reads covering at least either 1000 nucleotides or 50% of the genome/segment length were considered preliminary detections. Even de-replicated virus sequence sets can have regions that are highly similar between two or more sequences. Consensus sequences from this preliminary set of detected virus genomes/segments were extracted using samtools consensus (36). After removing strings of ambiguous N's, the preliminary consensus sequences were compared to each other pairwise using anicalc, and clusters were made with sequences of at least 98% average identity using aniclust. The longest consensus sequence in each cluster was kept as the final sequence. The fastq reads were then aligned to a smaller database of only the references corresponding to final sequences with the same parameters for minimap2/CoverM. Virus genomes/segments from this alignment with reads covering at least either 1000 nucleotides or 50% of the genome length were considered final detections. The metric RPKMF was calculated as (Reads Per Kilobase of reference genome)/(Million reads passing Filtering). The more common RPKM was not used here and is calculated as (Reads Per Kilobase of reference genome)/(Million reads mapped to reference sequences). This metric was used because, at low levels seen in wastewater, the proportion of reads aligning to virus genomes is presumed to correspond to the proportion of viral nucleic acid molecules over background molecules in the physical sample. Finally, per sample abundance and coverage metrics and taxonomic information were gathered in tabular format and merged into a single table for downstream processing.

Virus Community and Variant Analyses

The tables reporting virus abundance in RPKMF and sample metadata were processed in R, using tidyverse packages. Alpha diversity and Bray-Curtis dissimilarity statistics were calculated using Vegan (https://github.com/vegandevs/vegan). Plots were made with ggplot (37), ggpubr (https://github.com/kassambara/ggpubr), tSNE (https://github.com/jkrijthe/Rtsne), and color packages wesanderson (https://github.com/karthik/wesanderson) and nationalparkcolors (https://github.com/katiejolly/nationalparkcolors). Reads aligning to analyzed reference genomes were processed using iVar (38) with default settings and filtered to only report non-synonymous mutations with an allele frequency of at least 3% and a minimum allele coverage of 5 reads. Dendrogram figures were drawn in R with ggtree (39), aplot and ggplot.

Wastewater Virome Sequencing Correlation with RT-PCR and Clinical Data

For each virus analyzed, the wastewater virome abundance was summarized as the mean RPKMF of the relevant genome for all wastewater treatment plants in a given city for each week. For Influenza Virus, average RPKMF values for H3N2 and H1N1 were added together. For RT-PCR, genome copies (described above) were averaged across all wastewater treatment plants in a given city for each week. SARS-CoV-2 clinical PCR-positive case data and test positivity data was averaged across each week. The data for Houston were obtained at (https://covid-harriscounty.hub.arcgis.com/). The data from El Paso were obtained from El Paso Public Health upon request. Influenza "Weekly Percentage of Visits with Discharge Diagnosed Influenza" values were manually recorded from weekly Houston Health Department Flu Reports (https://www.houstonhealth.org/services/data-reporting/flu-reports). Monkeypox case numbers were scraped from the Texas Department of State Health Services website and transformed into weekly average cases. After transformation to weekly data, Simple moving average plots with an averaging period of 3 weeks were used to visualize temporal trends with R package tidyquant (https://github.com/business-science/tidyquant). Pearson statistics were used to quantify correlations between data.

Table 1.

Target	Forward	Reverse	Probe
SARS-CoV-2	CTG CAG ATT TGG ATG ATT TCT CC	CCT TGT GTG GTC TGC ATG AGT TTA G	ATT GCA ACAATC CAT GAG CAG TGC TGA CTC
Influenza A	1. CAA GAC CAA TCY TGT CAC CTC TGA C 2. CAA GAC CAA TYC TGT CAC CTY TGA C	1. GCATTYTGGACAAAVCGT CTACG 2. GCA TTT TGG ATA AAG CGT CTA CG	TGCAGTCCTCGCTCA CTGGGCACG

Primers and probes for qPCR assays: Sourced from https://www.cdc.gov/coronavirus/2019-ncov/lab/multiplex.html

References and Notes

1. S. Castiglioni, K. V. Thomas, B. Kasprzyk-Hordern, L. Vandam, P. Griffiths, Testing wastewater to detect illicit drugs: state of the art, potential and research needs. *Sci. Total Environ.* **487**, 613-620 (2014).

2. A. K. Venkatesan, R. U. Halden, Wastewater treatment plants as chemical observatories to forecast ecological and human health risks of manmade chemicals. *Sci. Rep.* **4**, 3731 (2014).

3. A. Harrington, V. Vo, K. Papp, R. L. Tillett, C. Chang, H. Baker, S. Shen, A. Amei, C. Lockett, D. Gerrity, E. C. Oh, Urban monitoring of antimicrobial resistance during a COVID-19 surge through wastewater surveillance. *Sci. Total Environ.* **853**, 158577 (2022).

4. J. A. Rothman, A. Saghir, S. Chung, N. Boyajian, T. Dinh, J. Kim, J. Oval, V. Sharavanan, C. York, A. G. Zimmer-Faust, K. Langlois, J. A. Steele, J. F. Griffith, K. L. Whiteson, 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).

5. M. B. Diamond, A. Keshaviah, A. I. Bento, O. Conroy-Ben, E. M. Driver, K. B. Ensor, R. U. Halden, L. P. Hopkins, K. G. Kuhn, C. L. Moe, E. C. Rouchka, T. Smith, B. S. Stevenson, Z. Susswein, J. R. Vogel, M. K. Wolfe, L. B. Stadler, S. V. Scarpino, Wastewater surveillance of pathogens can inform public health responses. *Nat. Med.* **28**, 1992-1995 (2022).

6. *Wastewater-based Disease Surveillance for Public Health Action* (National Academies Press, 2023).

7. V. Vo, R. L. Tillett, K. Papp, S. Shen, R. Gu, A. Gorzalski, D. Siao, R. Markland, C. Chang, H. Baker, J. Chen, M. Schiller, W. Q. Betancourt, E. Buttery, M. Pandori, M. A. Picker, D. Gerrity, E. C. Oh, Use of wastewater surveillance for early detection of Alpha and Epsilon SARS-CoV-2 variants of concern and estimation of overall COVID-19 infection burden. *Sci. Total Environ.* **835**, 155410 (2022).

8. D. S. Smyth, M. Trujillo, D. A. Gregory, K. Cheung, A. Gao, M. Graham, Y. Guan, C. Guldenpfennig, I. Hoxie, S. Kannoly, N. Kubota, T. D. Lyddon, M. Markman, C. Rushford, K. M. San, G. Sompanya, F. Spagnolo, R. Suarez, E. Teixeiro, M. Daniels, M. C. Johnson, J. J. Dennehy, Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat. Commun.* **13**, 635-3 (2022).

9. S. Karthikeyan, J. I. Levy, P. De Hoff, G. Humphrey, A. Birmingham, K. Jepsen, S. Farmer, H. M. Tubb, T. Valles, C. E. Tribelhorn, R. Tsai, S. Aigner, S. Sathe, N. Moshiri, B. Henson, A. M. Mark, A. Hakim, N. A. Baer, T. Barber, P. Belda-Ferre, M. Chacon, W. Cheung, E. S. Cresini, E. R. Eisner, A. L. Lastrella, E. S. Lawrence, C. A. Marotz, T. T. Ngo, T. Ostrander, A. Plascencia, R. A. Salido, P. Seaver, E. W. Smoot, D. McDonald, R. M. Neuhard, A. L. Scioscia, A. M. Satterlund, E. H. Simmons, D. B. Abelman, D. Brenner, J. C. Bruner, A. Buckley, M. Ellison, J. Gattas, S. L. Gonias, M. Hale, F. Hawkins, L. Ikeda, H. Jhaveri, T. Johnson, V. Kellen, B. Kremer, G. Matthews, R. W. McLawhon, P. Ouillet, D. Park, A. Pradenas, S. Reed, L.

Riggs, A. Sanders, B. Sollenberger, A. Song, B. White, T. Winbush, C. M. Aceves, C. Anderson, K. Gangavarapu, E. Hufbauer, E. Kurzban, J. Lee, N. L. Matteson, E. Parker, S. A. Perkins, K. S. Ramesh, R. Robles-Sikisaka, M. A. Schwab, E. Spencer, S. Wohl, L. Nicholson, I. H. McHardy, D. P. Dimmock, C. A. Hobbs, O. Bakhtar, A. Harding, A. Mendoza, A. Bolze, D. Becker, E. T. Cirulli, M. Isaksson, K. M. Schiabor Barrett, N. L. Washington, J. D. Malone, A. M. Schafer, N. Gurfield, S. Stous, R. Fielding-Miller, R. S. Garfein, T. Gaines, C. Anderson, N. K. Martin, R. Schooley, B. Austin, D. R. MacCannell, S. F. Kingsmore, W. Lee, S. Shah, E. McDonald, A. T. Yu, M. Zeller, K. M. Fisch, C. Longhurst, P. Maysent, D. Pride, P. K. Khosla, L. C. Laurent, G. W. Yeo, K. G. Andersen, R. Knight, Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature.* **609**, 101-108 (2022).

10. J. L. MELNICK, Poliomyelitis virus in urban sewage in epidemic and in nonepidemic times. *Am. J. Hyg.* **45**, 240-253 (1947).

11. E. Mercier, P. M. D'Aoust, O. Thakali, N. Hegazy, J. Jia, Z. Zhang, W. Eid, J. Plaza-Diaz, M. P. Kabir, W. Fang, A. Cowan, S. E. Stephenson, L. Pisharody, A. E. MacKenzie, T. E. Graber, S. Wan, R. Delatolla, Municipal and neighbourhood level wastewater surveillance and subtyping of an influenza virus outbreak. *Sci. Rep.* **12**, 15777-z (2022).

12. H. Ando, W. Ahmed, R. Iwamoto, Y. Ando, S. Okabe, M. Kitajima, Impact of the COVID-19 pandemic on the prevalence of influenza A and respiratory syncytial viruses elucidated by wastewater-based epidemiology. *Sci. Total Environ.* **880**, 162694 (2023).

13. O. Erster, I. Bar-Or, V. Levy, R. Shatzman-Steuerman, D. Sofer, L. Weiss, R. Vasserman, I. S. Fratty, K. Kestin, M. Elul, N. Levi, R. Alkrenawi, E. Mendelson, M. Mandelboim, M. Weil, Monitoring of Enterovirus D68 Outbreak in Israel by a Parallel Clinical and Wastewater Based Surveillance. *Viruses.* **14**, 1010. doi: 10.3390/v14051010 (2022).

14. A. Tiwari, S. Adhikari, D. Kaya, M. A. Islam, B. Malla, S. P. Sherchan, A. I. Al-Mustapha, M. Kumar, S. Aggarwal, P. Bhattacharya, K. Bibby, R. U. Halden, A. Bivins, E. Haramoto, S. Oikarinen, A. Heikinheimo, T. Pitkanen, Monkeypox outbreak: Wastewater and environmental surveillance perspective. *Sci. Total Environ.* **856**, 159166 (2023).

15. F. Wu, J. Oghuan, A. Gitter, K. D. Mena, E. L. Brown, Wide mismatches in the sequences of primers and probes for monkeypox virus diagnostic assays. *J. Med. Virol.* **95**, e28395 (2023).

16. X. Fernandez-Cassi, N. Timoneda, S. Martinez-Puchol, M. Rusinol, J. Rodriguez-Manzano, N. Figuerola, S. Bofill-Mas, J. F. Abril, R. Girones, Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. *Sci. Total Environ.* **618**, 870-880 (2018).

17. A. Crits-Christoph, R. S. Kantor, M. R. Olm, O. N. Whitney, B. Al-Shayeb, Y. C. Lou, A. Flamholz, L. C. Kennedy, H. Greenwald, A. Hinkle, J. Hetzel, S. Spitzer, J. Koble, A. Tan, F. Hyde, G. Schroth, S. Kuersten, J. F. Banfield, K. L. Nelson, Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *mBio.* **12**, e02703-20 (2021).

18. E. Wyler, C. Lauber, A. Manukyan, A. Deter, C. Quedenau, L. Teixeira Alves, S. Seitz, J. Altmüller, M. Landthaler, Comprehensive profiling of wastewater viromes by genomic sequencing. *bioRxiv.* (2022).

19. M. Khan, L. Li, L. Haak, S. H. Payen, M. Carine, K. Adhikari, T. Uppal, P. D. Hartley, H. Vasquez-Gross, J. Petereit, S. C. Verma, K. Pagilla, Significance of wastewater surveillance in detecting the prevalence of SARS-CoV-2 variants and other respiratory viruses in the community - A multi-site evaluation. *One Health.* **16**, 100536 (2023).

20. Camille McCall, Ryan A. Leo Elworth, Kristine M. Wylie, Todd N. Wylie, Katherine Dyson, Ryan Doughty, Todd J. Treangen, Loren Hopkins, Katherine Ensor, Lauren B. Stadler, Targeted metagenomic sequencing for detection of vertebrate viruses in wastewater for public health surveillance. *MedRxiv.* (2023).

21. M. K. D. Dueholm, M. Nierychlo, K. S. Andersen, V. Rudkjobing, S. Knutsson, MiDAS Global Consortium, M. Albertsen, P. H. Nielsen, MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nat. Commun.* **13**, 1908-7 (2022).

22. K. E. Shannon, D. Lee, J. T. Trevors, L. A. Beaudette, Application of real-time quantitative PCR for the detection of selected bacterial pathogens during municipal wastewater treatment. *Sci. Total Environ.* **382**, 121-129 (2007).

23. J. A. Rothman, A. Saghir, S. Chung, N. Boyajian, T. Dinh, J. Kim, J. Oval, V. Sharavanan, C. York, A. G. Zimmer-Faust, K. Langlois, J. A. Steele, J. F. Griffith, K. L. Whiteson, Longitudinal metatranscriptomic sequencing of Southern California wastewater representing 16 million people from August 2020–21 reveals widespread transcription of antibiotic resistance genes. *Water research (Oxford).* **229**, 119421 (2023).

24. J. R. Clark, A. Terwilliger, V. Avadhanula, M. Tisza, J. Cormier, S. Javornik-Cregeen, M. C. Ross, K. L. Hoffman, C. Troisi, B. Hanson, J. Petrosino, J. Balliew, P. A. Piedra, J. Rios, J. Deegan, C. Bauer, F. Wu, K. D. Mena, E. Boerwinkle, A. W. Maresso, Wastewater pandemic preparedness: Toward an end-to-end pathogen monitoring program. *Front. Public. Health.* **11**, 1137881 (2023).

25. Z. W. LaTurner, D. M. Zong, P. Kalvapalle, K. R. Gamas, A. Terwilliger, T. Crosby, P. Ali, V. Avadhanula, H. H. Santos, K. Weesner, L. Hopkins, P. A. Piedra, A. W. Maresso, L. B. Stadler, Evaluating recovery, cost, and throughput of different concentration methods for SARS-CoV-2 wastewater-based epidemiology. *Water research (Oxford).* **197**, 117043 (2021).

26. A. Gitter, C. Bauer, F. Wu, R. Ramphul, C. Chavarria, K. Zhang, J. Petrosino, M. Mezzari, G. Gallegos, A. L. Terwilliger, J. R. Clark, K. Feliz, V. Avadhanula, T. Piedra, K. Weesner, A. Maresso, K. D. Mena, Assessment of a SARS-CoV-2 wastewater monitoring program in El Paso, Texas, from November 2020 to June 2022. *Int. J. Environ. Health Res.*, 1-11 (2023).

27. L. R. Holtz, I. K. Bauer, H. Jiang, R. Belshe, P. Freiden, S. L. Schultz-Cherry, D. Wang, Seroepidemiology of astrovirus MLB1. *Clin. Vaccine Immunol.* **21**, 908-911 (2014).

28. A. A. Rabaan, M. A. Bakhrebah, M. S. Nassar, Z. S. Natto, A. Al Mutair, S. Alhumaid, M. Aljeldah, M. Garout, W. A. Alfouzan, F. S. Alshahrani, T. Sulaiman, M. K. AlFonaisan, M. Alfaresi, S. A. Alshamrani, F. Nainu, S. J. Yong, O. P. Choudhary, N. Ahmed, Suspected Adenovirus Causing an Emerging HEPATITIS among Children below 10 Years: A Review. *Pathogens.* **11**, 712. doi: 10.3390/pathogens11070712 (2022).

29. L. Li, H. Shimizu, L. T. P. Doan, P. G. Tung, S. Okitsu, O. Nishio, E. Suzuki, J. K. Seo, K. S. Kim, W. E. G. Muller, H. Ushijima, Characterizations of adenovirus type 41 isolates from children with acute gastroenteritis in Japan, Vietnam, and Korea. *J. Clin. Microbiol.* **42**, 4032-4039 (2004).

30. T. Kunitake, T. Kitamura, J. Guo, F. Taguchi, K. Kawabe, Y. Yogo, Parent-to-child transmission is relatively common in the spread of the human polyomavirus JC virus. *J. Clin. Microbiol.* **33**, 1448-1451 (1995).

31. S. N. Roach, R. A. Langlois, Intra- and Cross-Species Transmission of Astroviruses. *Viruses.* **13**, 1127. doi: 10.3390/v13061127 (2021).

32. S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, N. C. Kyrpides, CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578-585 (2021).

33. W. Shen, H. Ren, TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J. Genet. Genomics.* **48**, 844-850 (2021).

34. B. Bushnell, BBMap: A Fast, Accurate, Splice-Aware Aligner. (2014).

35. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* **34**, 3094-3100 (2018).

36. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience.* **10**, giab008. doi: 10.1093/gigascience/giab008 (2021).

37. *ggplot*, 2015).

38. N. D. Grubaugh, K. Gangavarapu, J. Quick, N. L. Matteson, J. G. De Jesus, B. J. Main, A. L. Tan, L. M. Paul, D. E. Brackney, S. Grewal, N. Gurfield, K. K. A. Van Rompay, S. Isern, S. F. Michael, L. L. Coffey, N. J. Loman, K. G. Andersen, An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8-7 (2019).

39. G. Yu, Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics.* **69**, e96 (2020).

1.

Acknowledgments: The authors thank the Health Departments and Water Utilities of Houston and El Paso for their support in the contribution of wastewater samples to the project. We also acknowledge Adrien Assie for illustrating the logo for EsViritu.

Funding:

This work was supported by S.B. 1780, 87th Legislature, 2021 Reg. Sess. (Texas 2021), NIH/NIAD (Grant number U19 AI44297), and Baylor College of Medicine and Alkek Foundation Seed.

Author contributions:

Conceptualization: MJT, SJC, JRC, AT, JC, FW, JR, JD, BH, CB, KM, PP, AWM, CT, KLH, JFP, EB

Methodology: MJT, SJC, VA, PZ, TA, KF, KLH, MCR, JB, PP, DH, AG, KM

Investigation: MJT, SJC, VA, AWM

Visualization: MJT, KZ, RL, CB

Funding acquisition: JFP, EB, AWM

Project administration: SJC, MCR, JC, JB, KM, AWM

Supervision: SJC, JFP, EB, AWM

Writing – original draft: MJT, SJC, AWM

Writing – review & editing: MJT, SJC, JC, FW, JR, JD, CB, PP, JFP, EB, AWM, CT

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All analyses for this manuscript can be reproduced with the code at https://github.com/cmmr/TX_wastewater_virome using the data at Zenodo repository https://zenodo.org/record/7884454 (follow instructions on the GitHub repository). EsViritu is available at https://github.com/cmmr/EsViritu (see this GitHub repository for most recent databases). All sequencing reads are uploaded to SRA at accession SUB13231081 with any human sequences removed.

Supplementary Materials

References (32-39)