Title

Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants

Journal

mBio, 12(1)

ISSN

2161-2129

Authors

Crits-Christoph, Alexander
Kantor, Rose S
Olm, Matthew R
et al.

Publication Date

2021-02-23

DOI

10.1128/mbio.02703-20

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE
Clinical Science and Epidemiology

Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants

Alexander Crits-Christoph,[a,b] Rose S. Kantor,[c] Matthew R. Olm,[d] Oscar N. Whitney,[e] Basem Al-Shayeb,[a,b] Yue Clare Lou,[a,b] Avi Flamholz,[e*] Lauren C. Kennedy,[c] Hannah Greenwald,[c] Adrian Hinkle,[c] Jonathan Hetzel,[f] Sara Spitzer,[f] Jeffery Koble,[f] Asako Tan,[f] Fred Hyde,[j] Gary Schroth,[f] Scott Kuersten,[j] Jillian F. Banfield,[b,g,h,i] Kara L. Nelson[b,c]

[a]Department of Plant and Microbial Biology, University of California, Berkeley, California, USA
[b]Innovative Genomics Institute, Berkeley, California, USA
[c]Department of Civil and Environmental Engineering, University of California, Berkeley, California, USA
[d]Department of Microbiology and Immunology, Stanford University, Stanford, California, USA
[e]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA
[f]Illumina, San Diego, California, USA
[g]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA
[h]Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[i]Chan Zuckerberg Biohub, San Francisco, California, USA
[j]Illumina, Madison, Wisconsin, USA

ABSTRACT Viral genome sequencing has guided our understanding of the spread and extent of genetic diversity of SARS-CoV-2 during the COVID-19 pandemic. SARS-CoV-2 viral genomes are usually sequenced from nasopharyngeal swabs of individual patients to track viral spread. Recently, RT-qPCR of municipal wastewater has been used to quantify the abundance of SARS-CoV-2 in several regions globally. However, metatranscriptomic sequencing of wastewater can be used to profile the viral genetic diversity across infected communities. Here, we sequenced RNA directly from sewage collected by municipal utility districts in the San Francisco Bay Area to generate complete and nearly complete SARS-CoV-2 genomes. The major consensus SARS-CoV-2 genotypes detected in the sewage were identical to clinical genomes from the region. Using a pipeline for single nucleotide variant calling in a metagenomic context, we characterized minor SARS-CoV-2 alleles in the wastewater and detected viral genotypes which were also found within clinical genomes throughout California. Observed wastewater variants were more similar to local California patient-derived genotypes than they were to those from other regions within the United States or globally. Additional variants detected in wastewater have only been identified in genomes from patients sampled outside California, indicating that wastewater sequencing can provide evidence for recent introductions of viral lineages before they are detected by local clinical sequencing. These results demonstrate that epidemiological surveillance through wastewater sequencing can aid in tracking exact viral strains in an epidemic context.

KEYWORDS coronavirus, environmental microbiology, genomics, metagenomics

The COVID-19 pandemic caused by SARS-CoV-2 reached the United States at the start of 2020, with multiple early introduction events in the states of Washington, California, and New York (1). Since then, the total number of cases in the country has surpassed 14 million, with over 275,000 deaths and enormous implications for public health (2). While clinical viral cases have been tracked mostly with quantitative reverse transcriptase PCR (RT-qPCR), there has also been extensive whole viral genome sequencing of clinical cases, generating over 75,000 genomes globally, including 17,000 from the United States and 2,500 from California (GISAID EpiCov database as of 23 August 2020) (3).

Citation Crits-Christoph A, Kantor RS, Olm MR, Whitney ON, Al-Shayeb B, Lou YC, Flamholz A, Kennedy LC, Greenwald H, Hinkle A, Hetzel J, Spitzer S, Koble J, Tan A, Hyde F, Schroth G, Kuersten S, Banfield JF, Nelson KL. 2021. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. mBio 12:e02703-20. https://doi.org/10.1128/mBio.02703-20.

Editor Melinda M. Pettigrew, Yale School of Public Health

Address correspondence to Kara L. Nelson, karanelson@berkeley.edu.

* Present address: Avi Flamholz, Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, California, USA.

Received 21 September 2020
Accepted 15 December 2020
Published 19 January 2021

Genomic epidemiology, the analysis of viral and microbial genomes in order to make inferences about pathogen evolution, transmission, and spread, has played an important role in improving our understanding of the transmission dynamics of the SARS-CoV-2 pandemic (4). Early in the pandemic, this approach revealed multiple introduction events into California and viral lineages present at different abundances across counties in Northern California (5). Genome sequencing was also used to show that there was unexpectedly frequent community spread of a specific genotype after early introduction in Washington State (6). Genome sequencing in the New York City area identified multiple viral introduction events from Europe (7), and sequencing in the Mission district of San Francisco identified distinct viral strains in a single neighborhood, with transmission between family clusters (8).

Unlike many respiratory viruses, RNA of SARS-CoV-2 and other coronaviruses can be detected in human feces (9–11). Before the COVID-19 pandemic, members of the *Coronaviridae* had been previously identified in municipal wastewater through both RT-qPCR and shotgun metagenomic and metatranscriptomic sequencing (12, 13). Since the start of the COVID-19 pandemic, wastewater RT-qPCR has quantified the amount of SARS-CoV-2 RNA in sewage to estimate the abundance of the virus across many different municipal regions globally (14–22). Prior work showed that shotgun wastewater sequencing can provide information about many viruses simultaneously (12, 23, 24) and enable genome-resolved (25) and phylogenetic analyses (26, 27). In one study, a SARS-CoV-2 consensus genome was obtained from sewage via targeted amplification and long-read sequencing, allowing for phylogenetic analysis of the predominant lineage (27). Here, we show that sequencing of viral concentrates and RNA extracted directly from wastewater can identify multiple SARS-CoV-2 genotypes at various abundances known to be present in communities, as well as additional genotypic variants not yet observed in local clinical sequencing efforts.

RESULTS AND DISCUSSION

Metatranscriptomic detection of SARS-CoV-2 and other viruses in wastewater. Twenty-four-hour 1-liter composite samples of raw sewage were collected from wastewater treatment facilities in Alameda and Marin Counties in Northern California between 19 May 2020 and 15 July 2020 (see Table S1 in the supplemental material). We extracted nucleic acids from samples using three methods that enriched for viral particles (ultrafiltration) or total RNA (RNA silica columns or silica milk). SARS-CoV-2 viral RNA was first detected using a RT-qPCR assay (see Materials and Methods) of the N gene and C_q values ranged from 29.5 to 36.2, or an estimated ~2 to ~553 genome copies/μl of RNA. From this we estimate that there were 2.8×10^5 genome copies/liter of wastewater on average across our samples (see Table S1). For each sample, 40 to 50 μl of RNA was prepared for sequencing, implying an estimated ~4,438 viral genome copies on average were contained within each sequencing library.

After cDNA synthesis from the total RNA, samples were enriched for a panel of human respiratory viruses using a commercially available oligo-capture approach (Illumina respiratory virus panel; see Materials and Methods) and sequenced on a NextSeq 550 to produce on average 12 million 2×75 bp reads per sample. Reads were mapped to the human genome to estimate the amount of human RNA/DNA in the samples (0.7 to 16% of reads per sample). Sequencing reads were then mapped to a dereplicated set of all eukaryotic viruses contained in the RefSeq database, and stringently filtered to include only high-quality reads matching reference sequences with >97% identity (see Materials and Methods). Viral abundances and SNVs (single nucleotide variants) were then calculated using the metagenomic strain-typing program inStrain v1.12. We detected SARS-CoV-2 at various abundances of sequenced RNA/DNA (0 to 14%) across samples (Fig. 1a and b; see also Table S1). Sequencing relative abundance of SARS-CoV-2 was not strongly correlated with RT-qPCR genome copy quantification, likely due to the variability introduced by different extraction methods. Viral enrichment by ultrafiltration achieved higher relative abundances of SARS-CoV-2

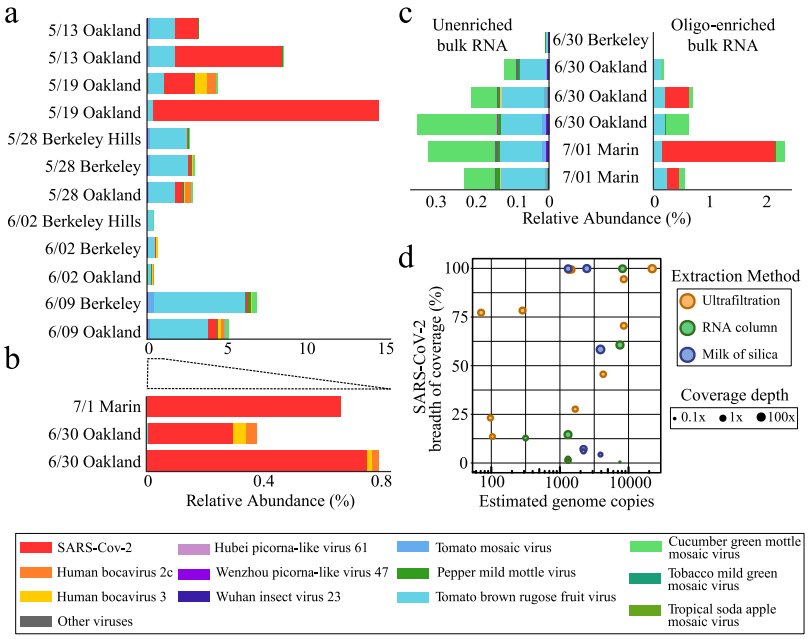

FIG 1 Characterized viruses detected in enriched and unenriched wastewater metatranscriptomes. The relative abundances of viruses with eukaryotic hosts in the RefSeq database as a percentage of total sequencing reads derived from the sample in Amicon ultrafiltration (viral fractionation) (a) and total RNA column and milk of silica samples (b). All samples were enriched with the Illumina respiratory virus panel. (c) Relative abundances of RefSeq viruses in unenriched metatranscriptomics (left) and the same samples after oligonucleotide enrichment with the Illumina respiratory virus panel. (d) The relationship between the quantity of viral genome copies in 40 μl of purified RNA and SARS-CoV-2 genome completeness (measured in breadth of coverage) for each sample. Samples are colored by extraction methodology, and the size of the point corresponds to the mean SARS-CoV-2 depth of coverage.

RNA, although these experiments were time-intensive and often had lower absolute genome copy number recovery according to RT-qPCR. In addition, we sequenced replicates from one set of samples with rRNA depletion but no viral enrichment. Without enrichment, we were able to only detect fewer than 40 total SARS-CoV-2 read pairs (Fig. 1c; see also Table S1). While this illustrates the difficulty of detecting specific viruses in wastewater in unenriched sequencing data sets, larger sequencing efforts may overcome this limitation by sequencing more deeply.

Other human viruses identified in the wastewater sequencing included Human bocaviruses 2c and 3 (Fig. 1a and b), both of which are respiratory viruses sometimes capable of causing gastroenteritis, and are included in the Illumina respiratory virus panel. Bocaviruses have been identified in sewage samples previously (28, 29). Picornavirus-like viruses were also detected (Fig. 1c). The most abundant viruses in the data were plant viruses including cucumber green mottle mosaic virus and pepper mild mottle virus (PMMoV) (Fig. 1a). These viruses are known to be highly abundant in human wastewater (30) and have been used as fecal loading controls in wastewater SARS-CoV-2 quantification (19). Near-complete (>95% breadth of coverage) genomes were obtained for SARS-CoV-2, bocavirus 3, PMMoV, and other plant viruses (see Table S2), implying that these viruses were at high enough abundance in the data set for exact genomic analysis.

Recovery of complete and nearly complete SARS-CoV-2 viral genomes from wastewater. Complete consensus viral genomes are required to perform viral lineage tracking for genomic epidemiology. We obtained complete consensus SARS-CoV-2 genomes (breadth of coverage >99%) from 7 of 22 samples (31%), while large-scale patient sequencing efforts have for example obtained genomes for ~80% of samples (31). Only samples with RT-qPCR C_T values <33 (~25 genome copies/μl) yielded complete consensus genomes (Fig. 1d), but we also recovered at least one genome using each of our three extraction methods. The mean depth of coverage for each complete

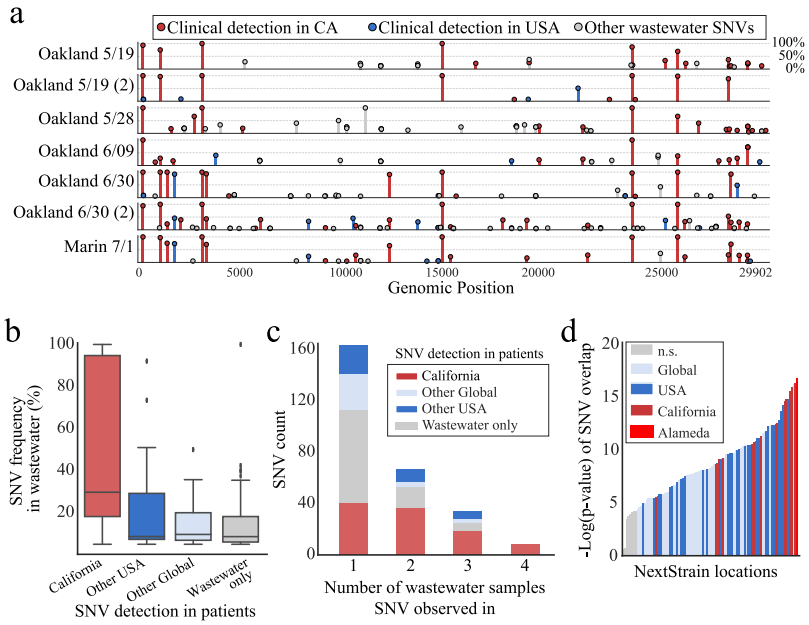

FIG 2 SARS-CoV-2 SNVs in wastewater samples. (a) Allele frequencies of SARS-CoV-2 in wastewater metatranscriptomes for each sample. Each point is a SNV by location on the SARS-CoV-2 genome (*x* axis), and the height of the bar (*y* axis) is the frequency of the alternative allele (relative to the reference genome EPI_ISL_402124) at that position. Wastewater SNVs are colored based on whether they have previously been observed in clinical samples from California, the United States, or neither. (b) Wastewater SARS-CoV-2 frequencies grouped by whether they have been observed in clinical samples from different regions. Most highly abundant SNVs have been observed previously in California or elsewhere in the United States. (c) SARS-CoV-2 SNVs grouped by the number of wastewater samples observed in (out of seven high-quality samples). Most SNVs that were observed in two or more samples have been observed clinically in California. (d) Multiple hypothesis adjusted (Bonferroni correction) *P* value distribution of hypergeometric tests for overlap between all wastewater SNVs observed and the variants clinically observed and reported in each location (a county level designation in the United States). Alameda County was the most significant comparison.

genome ranged from 7× to 107× after filtering and removal of PCR duplicates. The consensus genomes from Alameda County, and the one from Marin County, were all within 4-bp differences of each other. These consensus genomes were found to be unlikely to be chimeric, as a BLAST analysis identified SARS-CoV-2 genomes that were 100% identical at all nongapped positions (see Table S3) obtained from patients in northern California. Consensus genomes may represent predominant SARS-CoV-2 lineages in the population in the serviced areas during the summer of 2020. The results demonstrate genomic accuracy for recovery of consensus SARS-CoV-2 genomes so long as sufficient coverage is achieved in metatranscriptomic data sets.

Identification of alternative SARS-CoV-2 variants in wastewater populations recovers locally reported clinical genotypes. While consensus genotypes can describe the predominant genotype of a virus in a metatranscriptome, the strength of wastewater-based sampling and sequencing lies in the ability to identify alternative genotypes in the population being sampled. Using a recently developed pipeline for metagenomic SNV calling (32), we identified putative SNVs that are variable within the viral population sampled in each wastewater sample after read mapping to the SARS-CoV-2 reference genome EPI_ISL_402124 (Fig. 2a; see also Table S4). Due to the large-scale sequencing efforts of SARS-CoV-2 in patients in both northern California and worldwide, we established that these SNVs had also been detected in genomes from individual patients. Across all samples, 50% of SNVs observed in wastewater samples at >10% frequency were also observed in patient-derived viral genomes from California; 61% were observed in viral genomes from the United States, and 71% were observed in any viral genomes collected worldwide. SNVs that have been observed in California patients had significantly higher allele frequencies in the wastewater samples than

those that were not detected in clinical cases (mean, 48 versus 15%, respectively; $P < 0.01$ [two-sided *t* test]) (Fig. 2b). This is likely because the more abundant a SNV is in the population, the more likely it is to be sampled in wastewater and in the clinic. Further, several of the same SNVs were observed across samples, and these recurrent SNVs were, on average, 2.3× more likely to be observed in California or U.S. patient-derived genomes than SNVs observed once (Fig. 2c). Taken together, these are strong signals that deeper sequencing of wastewater and combining information across samples better recapitulates true viral genomic variation in the sampled population.

More than 75,000 patient-derived SARS-CoV-2 genomes have been sequenced and deposited into the GISAID database globally, including 2,500 genomes obtained from patients in California. To understand the context of the viral genomic variation we observed within wastewater samples, we used a hypergeometric test to calculate the likelihood of overlap by chance between the set of wastewater variants and the set of variants observed in viruses from patients in a given region. This computes the probability of observing a certain amount of overlap in variants by chance and accounts for the fact that some regions have far more sequenced patient genomes and correspondingly more alleles than others. For example, the probability of the observed overlap between wastewater variants and California clinical variants having occurred by chance was calculated to be $P < 10^{-10}$, indicating a high likelihood of nonrandom overlap. By further comparing the probabilities of SNV overlap between patient genotypes and wastewater genotypes at the NextStrain "location" level (corresponding to counties and/or cities), we found the highest likelihood of nonrandom overlap between all wastewater genotypes observed and clinical genotypes from Alameda County (Fig. 2d), the location that the wastewater samples were also derived from.

Identification of potential lineage transmission events previously undetected in local patient-based sequencing at time of sampling. Some clinical SARS-CoV-2 viral strains can be differentiated by more than one SNV. Across the wastewater data set, we observed one pair and one triplet of SNVs that were shared by clinical isolates. The pair and triplet of SNVs each occurred at similar frequencies, supporting their linkage in wastewater genomes (Fig. 3a and b). In addition to the SNVs that also have been observed clinically in California, there were four SNVs recurrent across wastewater samples that had not been previously observed in California but had been observed elsewhere in the United States (Fig. 3c). Two adjacent SNVs (14222G and 14223C) are associated with a single viral strain that has been often observed in clinical samples in Washington State. Another two SNVs (8083A and 1738T) are not linked, but both have been observed in different clinical genomes of four other states in the United States. Interestingly, these variants appear to have arisen or arrived in the United States only during the month of July, suggesting that they may be detected in clinical samples from California in the near future.

Overall, this study demonstrated that wastewater sequencing can accurately identify genotypes of viral strains that are clinically detected in a region and those not yet detected by clinical sequencing. Another key advantage of this method is that it does not rely on specific PCR primers, which can fail to detect SARS-CoV-2 strains with mutations in the primed sequence (33). With more intensive wastewater sampling, this approach also has the potential to reveal patterns of virus distribution within communities, helping to elucidate the transmission and spread of diseases during epidemics. Perhaps most significantly, the results indicate that wastewater sequencing can detect recent introductions of SARS-CoV-2 genotypes and other disease-causing viruses at a population scale.

MATERIALS AND METHODS

Sample collection and extraction. Twenty-four-hour 1-liter composite samples were collected at four different wastewater interceptors in the San Francisco Bay Area (labeled "Berkeley," "Berkeley Hills," "Oakland," and "Marin," based roughly on the municipal areas each services). The time-weighted composite samples were collected using autosamplers that draw from influent every 15 min into 24-hourly bottles, which were then combined and mixed, and subsamples were taken for analysis. Samples were immediately processed by extraction via three different methods. The first method was ultrafiltration

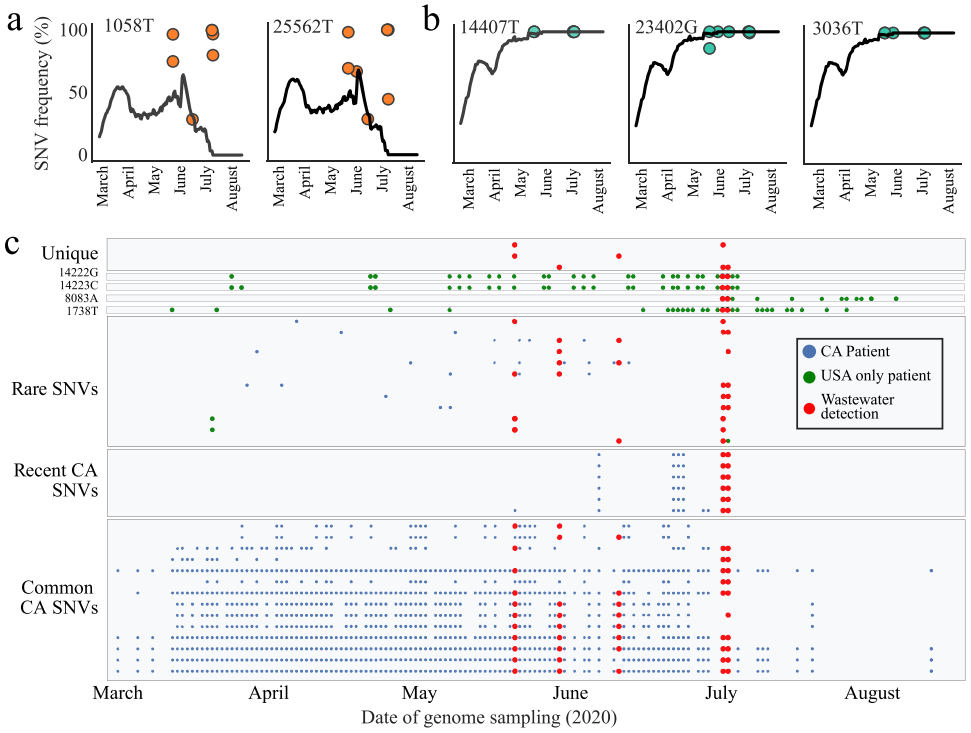

FIG 3 Time series of SARS-CoV-2 genotypes in California wastewater compared to patients. (a) Frequencies of two SNVs found in the same viral lineage across California clinical samples (black lines) and within each wastewater sample (orange points). (b) Frequencies of three SNVs found in the same viral lineage across California clinical samples (black lines) and within each wastewater sample (green points). (c) Time series of detection for recurrent wastewater genotypes in clinical samples versus wastewater samples. Each row on the *y* axis is a SNV, and the presence of a point along the *x* axis indicates when that SNV was detected in either a clinical sample or a wastewater sample.

with Amicon Ultra-15 100-kDa centrifugal filter units. Wastewater was heat inactivated in a water bath at 60°C for 90 min. Wastewater samples were then filtered on 0.22-μm SteriFlip filter units. While we found that the 0.22-μm filtration step, which was implemented to reduce clogging of the Amicon ultrafilter, did result in a loss of RNA (data not shown), we believe the methods recovered a sufficient quantity of viral RNA to adequately profile their genetic diversity. Amicon filter units were prepared by incubation with 1% bovine serum albumin in 1× phosphate-buffered saline (PBS) on ice for 1 h and then spun, loaded with 2 ml of PBS, and spun again to rinse. Amicon 100-kDa centrifugal filter units were then loaded with 15 ml of wastewater filtrate (flowthrough) and spun in a swinging-bucket rotor at 4,750 × *g* for 30 min at 4°C. Flowthrough was discarded, and amicons were reloaded with sample until all sample volume (40 ml) had been processed. For three samples (see Table S1), we processed more than 40 ml per sample but found that this did not improve the resulting SARS-CoV-2 genome quality in this specific instance. For all Amicon centrifuge-concentrated samples, the final volume of the concentrate was ~250 μl. RNA was then extracted with a Qiagen AllPrep DNA/RNA minikit. The second extraction method, direct RNA extraction with silica columns, began with viral and bacterial lysis of samples with 9.5 g of NaCl per 40 ml of wastewater and filtration on a 5-μm polyvinylidene fluoride (polyvinylidene difluoride) filter. The resulting filtrate (flowthrough) was then loaded onto a Zymo III-P silica spin column via vacuum manifold, and RNA was directly eluted from this column. Details of this protocol are available elsewhere (https://www.protocols.io/view/v-2-direct-wastewater-rna-capture-and-purification-bjr9km96). The third extraction method, "milk of silica," began with sample lysis and filtration, as in the second method. Filtered lysate is bound to free silicon dioxide particulate, eluted from the particulate, and concentrated via isopropanol precipitation. This protocol is also available online (https://www.protocols.io/view/direct-wastewater-rna-extraction-via-the-34-milk-o-biwfkfbn).

RT-qPCR and genome copy quantification. The number of viral genome copies in each sample was determined via probe-based qRT-PCR on an Applied Biosystems QuantStudio 3 real-time PCR system with the Thermo Fisher TaqPath 1-Step RT-qPCR Master Mix or TaqMan Fast Virus 1-Step Master Mix. The primer set and probe were purchased as part of the 2019-CoV RUO kit (IDT), and our quantification used the previously published CDC N1 assay (34). Either 2 or 5 μl of sample was used for each reaction (see Table S1) in a 10- or 20-μl reaction, respectively. Cycling conditions were 25°C for 2 min, 50°C for 15 min, 95°C for 2 min, and 45 cycles of 95°C for 3 s and 55°C for 30 s. A standard curve for absolute quantification of viral genome copies was generated with synthetic RNA standards of the SARS-CoV-2 genome (Twist Biosciences).

Library preparation and sequencing. Sequencing for a first set of samples was performed at the Microbial Genome Sequencing Center (Pittsburgh, PA) in three independent sequencing runs. A Maxima double-stranded cDNA RT kit (Thermo Fisher) was used to generate cDNA. An Illumina Flex for Enrichment kit paired with an Illumina Respiratory Virus Oligo Panel (Illumina, Inc.) was used to enrich for respiratory virus cDNA with 15 PCR cycles in the final step. The libraries were then sequenced on a NextSeq 550 to yield on average 119 Mbp of 2 × 75 bp paired-end sequencing reads. For a second set of samples (see Table S1), rRNA depletion was performed, and oligonucleotide capture enriched and unenriched sequencing strategies were compared. The rRNA depletion was done using RiboZero Plus supplemented with a comprehensive "Gut Microbiome" probe set. Libraries were prepared using the Illumina RNA Prep with Enrichment (L) Tagmentation protocol. The rRNA-depleted samples were amplified for 20 cycles. Enrichment was performed using the Illumina Respiratory Virus Oligo Panel.

Metatranscriptomic viral abundances. The abundances of viruses within wastewater were obtained by mapping reads with Bowtie 2 (35) to an index of all viral genomes downloaded from the RefSeq Database (release 201). For abundance calculations, mapped read pairs with MAPQ > 20 and pair percent identity to the reference >95% were retained using inStrain v1.3.2 (32). Duplicate reads were removed with the clumpify.sh dedup command from the BBTools software suite (Bushnell 2014). Only viral genomes with at least 10% breadth of genomic coverage obtained were reported.

SARS-CoV-2 variant analysis. Seven samples with nearly complete SARS-CoV-2 breadth of genomic coverage (>99%) were further investigated for a strain-resolved analysis. SNV calling was performed using inStrain v1.3.2 on all read pairs with >90% average nucleotide identity to the SARS-CoV-2 reference. An absolute minimum of two read pairs supporting a variant allele was required for any SNV to be considered in further analysis. PCR duplicates were removed with the markdup command in the Sambamba package (36). All analysis and SNV locations reported are with respect to the reference genome "hCoV-19/Wuhan/WIV04/2019|EPI_ISL_402124|2019-12-30|China." Consensus genomes from each sample were created using a custom Python script that required a minimum of three reads supporting each genomic position. A multiple sequence alignment of publicly available SARS-CoV-2 genomes and their metadata were downloaded from the GISAID (3) EpiCov database on 23 August 2020. The multiple sequence alignment was processed with a custom Python script to obtain a list of variants for each genome with respect to the WIV04 reference sequence. We removed from all analyses the genomic positions recommended to be masked from SARS-CoV-2 alignments by https://virological.org/t/masking-strategies-for-sars-cov-2-alignments/480. Hypergeometric distributions were calculated with the stats.hypergeom function in scipy (37) to compare wastewater samples to all clinical data from each NextStrain "location" with at least 20 genomes deposited. The following parameters were used for hypergeometric distribution testing: the total number of SNVs observed across all clinical SARS-CoV-2 genomes, the number of SNVs observed in wastewater, the number of clinical SNVs in a region, and the observed overlap between the two. The reproducible code is available at https://github.com/alexcritschristoph/wastewater_sarscov2.

Data availability. Sequencing data for this project has been released under NCBI BioProject ID PRJNA661613. Processed data, reproducible code, and workflows for the analyses performed are available at https://github.com/alexcritschristoph/wastewater_sarscov2.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

TABLE S1, XLSX file, 0.01 MB.

TABLE S2, XLSX file, 0.1 MB.

TABLE S3, XLSX file, 0.03 MB.

TABLE S4, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

We gratefully acknowledge the originating and submitting laboratories of SARS-CoV-2 genomes in the GISAID EpiCoV database (www.gisaid.org) that were used for our comparisons to clinical samples and in particular the Innovative Genomics Institute SARS-CoV-2 Sequencing Group for Alameda County genomes. We also gratefully acknowledge Vinson Fan for assistance with RT-qPCR and the laboratory of Robert Tjian for sharing materials.

Funding was provided to K.L.N. and J.F.B. by a Rapid Research Response grant from the Innovative Genomics Institute (IGI) and a seed grant from the Center for Information Technology Research in the Interest of Society (CITRIS) at UC Berkeley.

REFERENCES

1. Jorden MA, Rudman SL, Villarino E, Hoferka S, Patel MT, Bemis K, Simmons CR, Jespersen M, Iberg Johnson J, Mytty E, Arends KD, Henderson JJ, Mathes RW, Weng CX, Duchin J, Lenahan J, Close N, Bedford T, Boeckh M, Chu HY, Englund JA, Famulare M, Nickerson DA, Rieder MJ, Shendure J, Starita LM, Team CC-19 R, CDC COVID-19 Response Team. 2020. Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020. MMWR Morb Mortal Wkly Rep 69:680–684. https://doi.org/10.15585/mmwr.mm6922e1.

2. Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20:533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.
3. Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 1:33–46. https://doi.org/10.1002/gch2.1018.
4. Fauver JR, Petrone ME, Hodcroft EB, Shioda K, Ehrlich HY, Watts AG, et al. 2020. Coast-to-coast spread of SARS-CoV-2 in the United States revealed by genomic epidemiology. medRxiv https://doi.org/10.1101/2020.03.25.20043828.
5. Deng X, Gu W, Federman S, Du Plessis L, Pybus OG, Faria NR, Wang C, Yu G, Bushnell B, Pan C-Y, Guevara H, Sotomayor-Gonzalez A, Zorn K, Gopez A, Servellita V, Hsu E, Miller S, Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Chu HY, Shendure J, Jerome KR, Anderson C, Gangavarapu K, Zeller M, Spencer E, Andersen KG, MacCannell D, Paden CR, Li Y, Zhang J, Tong S, Armstrong G, Morrow S, Willis M, Matyas BT, Mase S, Kasirye O, Park M, Masinde G, Chan C, Yu AT, Chai SJ, Villarino E, Bonin B, Wadford DA, Chiu CY, et al. 2020. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into northern California. Science 369:582–587. https://doi.org/10.1126/science.abb9263.
6. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang M-L, Nalla A, Pepper G, Reinhardt A, Xie H, Shrestha L, Nguyen TN, Adler A, Brandstetter E, Cho S, Giroux D, Han PD, Fay K, Frazar CD, Ilcisin M, Lacombe K, Lee J, Kiavand A, Richardson M, Sibley TR, Truong M, Wolf CR, Nickerson DA, Rieder MJ, Englund JA, Hadfield J, Hodcroft EB, Huddleston J, Moncla LH, Müller NF, Neher RA, Deng X, Gu W, Federman S, Chiu C, Duchin JS, Gautom R, Melly G, Hiatt B, Dykema P, Lindquist S, Queen K, Tao Y, Uehara A, Tong S, The Seattle Flu Study Investigators, et al. 2020. Cryptic transmission of SARS-CoV-2 in Washington State. Science 370:571–575. https://doi.org/10.1126/science.abc0523.
7. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S, Kleiner G, Polanco J, Khan Z, Alburquerque B, van de Guchte A, Dutta J, Francoeur N, Melo BS, Oussenko I, Deikus G, Soto J, Sridhar SH, Wang Y-C, Twyman K, Kasarskis A, Altman DR, Smith M, Sebra R, Aberg J, Krammer F, García-Sastre A, Luksza M, Patel G, Paniz-Mondolfi A, Gitman M, Sordillo EM, Simon V, van Bakel H. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369:297–301. https://doi.org/10.1126/science.abc1917.
8. Chamie G, Marquez C, Crawford E, Peng J, Petersen M, Schwab D, Schwab J, Martinez J, Jones D, Black D, Gandhi M, Kerkhoff AD, Jain V, Sergi F, Jacobo J, Rojas S, Tulier-Laiwa V, Gallardo-Brown T, Appa A, Chiu C, Rodgers M, Hackett J, Kistler A, Hao S, Kamm J, Dynerman D, Batson J, Greenhouse B, DeRisi J, Havlir DV, CLIAhub Consortium. 2020. SARS-CoV-2 community transmission disproportionately affects Latinx population during shelter-in-place in San Francisco. Clin Infect Dis https://doi.org/10.1093/cid/ciaa1234.
9. Wölfel R, Corman VM, Guggemos W, Seilmaier M, Zange S, Müller MA, Niemeyer D, Jones TC, Vollmar P, Rothe C, Hoelscher M, Bleicker T, Brünink S, Schneider J, Ehmann R, Zwirglmaier K, Drosten C, Wendtner C. 2020. Virological assessment of hospitalized patients with COVID-2019. Nature 581:465–469. https://doi.org/10.1038/s41586-020-2196-x.
10. Jevšnik M, Steyer A, Zrim T, Pokorn M, Mrvič T, Grosek Š, Strle F, Lusa L, Petrovec M. 2013. Detection of human coronaviruses in simultaneously collected stool samples and nasopharyngeal swabs from hospitalized children with acute gastroenteritis. Virol J 10:46. https://doi.org/10.1186/1743-422X-10-46.
11. Amoah ID, Kumari S, Bux F. 2020. Coronaviruses in wastewater processes: source, fate and potential risks. Environ Int 143:105962. https://doi.org/10.1016/j.envint.2020.105962.
12. Bibby K, Peccia J. 2013. Identification of viral pathogen diversity in sewage sludge by metagenome analysis. Environ Sci Technol 47:1945–1951. https://doi.org/10.1021/es305181x.
13. Wang X-W, Li J-S, Guo T-K, Zhen B, Kong Q-X, Yi B, Li Z, Song N, Jin M, Xiao W-J, Zhu X-M, Gu C-Q, Yin J, Wei W, Yao W, Liu C, Li J-F, Ou G-R, Wang M-N, Fang T-Y, Wang G-J, Qiu Y-H, Wu H-H, Chao F-H, Li J-W. 2005. Concentration and detection of SARS coronavirus in sewage from Xiao Tang Shan Hospital and the 309th Hospital. J Virol Methods 130:165. https://doi.org/10.1016/j.jviromet.2005.08.010.
14. Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. 2020. Presence of SARS-coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in The Netherlands. Environ Sci Technol Lett 7:511–516. https://doi.org/10.1021/acs.estlett.0c00357.
15. Ahmed W, Angel N, Edson J, Bibby K, Bivins A, O'Brien JW, Choi PM, Kitajima M, Simpson SL, Li J, Tscharke B, Verhagen R, Smith WJM, Zaugg J, Dierens L, Hugenholtz P, Thomas KV, Mueller JF. 2020. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: a proof of concept for the wastewater surveillance of COVID-19 in the community. Sci Total Environ 728:138764. https://doi.org/10.1016/j.scitotenv.2020.138764.
16. Wu F, Zhang J, Xiao A, Gu X, Lee WL, Armas F, Kauffman K, Hanage W, Matus M, Ghaeli N, Endo N, Duvallet C, Poyet M, Moniz K, Washburne AD, Erickson TB, Chai PR, Thompson J, Alm EJ. 2020. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. mSystems 5:e00614-20. https://doi.org/10.1128/mSystems.00614-20.
17. Gonzalez R, Curtis K, Bivins A, Bibby K, Weir MH, Yetka K, Thompson H, Keeling D, Mitchell J, Gonzalez D. 2020. COVID-19 surveillance in southeastern Virginia using wastewater-based epidemiology. Water Res 186:116296. https://doi.org/10.1016/j.watres.2020.116296.
18. Bivins A, North D, Ahmad A, Ahmed W, Alm E, Been F, Bhattacharya P, Bijlsma L, Boehm AB, Brown J, Buttiglieri G, Calabro V, Carducci A, Castiglioni S, Cetecioglu Gurol Z, Chakraborty S, Costa F, Curcio S, de Los Reyes FL, Delgado Vela J, Farkas K, Fernandez-Casi X, Gerba C, Gerrity D, Girones R, Gonzalez R, Haramoto E, Harris A, Holden PA, Islam MT, Jones DL, Kasprzyk-Hordern B, Kitajima M, Kotlarz N, Kumar M, Kuroda K, La Rosa G, Malpei F, Mautus M, McLellan SL, Medema G, Meschke JS, Mueller J, Newton RJ, Nilsson D, Noble RT, van Nuijs A, Peccia J, Perkins TA, Pickering AJ, et al. 2020. Wastewater-based epidemiology: global collaborative to maximize contributions in the fight against COVID-19. Environ Sci Technol 54:7754–7757. https://doi.org/10.1021/acs.est.0c02388.
19. Wu F, Xiao A, Zhang J, Moniz K, Endo N, Armas F, et al. 2020. SARS-CoV-2 titers in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases. medRxiv https://doi.org/10.1101/2020.06.15.20117747.
20. Weidhaas J, Aanderud Z, Roper D, VanDerslice J, Gaddis E, Ostermiller J, et al. 2020. Correlation of SARS-CoV-2 RNA in wastewater with COVID-19 disease burden in sewersheds. Res Square https://doi.org/10.21203/rs.3.rs-40452/v1.
21. Vallejo JA, Rumbo-Feal S, Conde-Perez K, Lopez-Oriona A, Tarrio J, Reif R, et al Highly predictive regression model of active cases of COVID-19 in a population by screening wastewater viral load. medRxiv https://doi.org/10.1101/2020.07.02.20144865.
22. Peccia J, Zulli A, Brackney DE, Grubaugh ND, Kaplan EH, Casanovas-Massana A, et al SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics. medRxiv https://doi.org/10.1101/2020.05.19.20105999.
23. Fernandez-Cassi X, Timoneda N, Martínez-Puchol S, Rusiñol M, Rodriguez-Manzano J, Figuerola N, Bofill-Mas S, Abril JF, Girones R. 2018. Metagenomics for the study of viruses in urban sewage as a tool for public health surveillance. Sci Total Environ 618:870–880. https://doi.org/10.1016/j.scitotenv.2017.08.249.
24. Martínez-Puchol S, Rusiñol M, Fernández-Cassi X, Timoneda N, Itarte M, Andrés C, Antón A, Abril JF, Girones R, Bofill-Mas S. 2020. Characterization of the sewage virome: comparison of NGS tools and occurrence of significant pathogens. Sci Total Environ 713:136604. https://doi.org/10.1016/j.scitotenv.2020.136604.
25. Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, Grabe M, Hendrix RW, Girones R, Wang D, Pipas JM. 2011. Raw sewage harbors diverse viral populations. mBio 2:e00180-11. https://doi.org/10.1128/mBio.00180-11.
26. Ng TFF, Marine R, Wang C, Simmonds P, Kapusinszky B, Bodhidatta L, Oderinde BS, Wommack KE, Delwart E. 2012. High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. J Virol 86:12161–12175. https://doi.org/10.1128/JVI.00869-12.
27. Nemudryi A, Nemudraia A, Wiegand T, Surya K, Buyukyoruk M, Cicha C, Vanderwood KK, Wilkinson R, Wiedenheft B. 2020. Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. Cell Rep Med 1:100098. https://doi.org/10.1016/j.xcrm.2020.100098.
28. Iaconelli M, Divizia M, Della Libera S, Di Bonito P, La Rosa G. 2016. Frequent detection and genetic diversity of human bocavirus in urban sewage samples. Food Environ Virol 8:289–295. https://doi.org/10.1007/s12560-016-9251-7.
29. Blinkova O, Rosario K, Li L, Kapoor A, Slikas B, Bernardin F, Breitbart M, Delwart E. 2009. Frequent detection of highly diverse variants of cardiovirus, cosavirus, bocavirus, and circovirus in sewage samples collected in the United States. J Clin Microbiol 47:3507–3513. https://doi.org/10.1128/JCM.01062-09.
30. Kitajima M, Sassi HP, Torrey JR. 2018. Pepper mild mottle virus as a water quality indicator. NPJ Clean Water 1:1–9. https://doi.org/10.1038/s41545-018-0019-5.

31. Thielen PM, Wohl S, Mehoke T, Ramakrishnan S, Kirsche M, Falade-Nwulia O, et al. 2020. Genomic Diversity of SARS-CoV-2 During Early Introduction into the United States National Capital Region. medRxiv https://doi.org/10.1101/2020.08.13.20174136.
32. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek B, Morowitz MJ, Banfield JF. InStrain enables population genomic analysis from metagenomic data and rigorous detection of identical microbial strains. medRxiv https://doi.org/10.1101/2020.01.22.915579.
33. Vanaerschot M, Mann SA, Webber JT, Kamm J, Bell SM, Bell J, et al. 2020. Identification of a polymorphism in the N gene of SARS-CoV-2 that adversely impacts detection by a widely-used RT-PCR assay. medRxiv https://doi.org/10.1101/2020.08.25.265074.
34. Centers for Disease Control and Prevention. 2020. Research use only-2019: novel coronavirus (2019-nCoV) real-time RT-PCR primers and probes. Reviewed 20 May 2020. Centers for Disease Control and Prevention, Atlanta, GA.
35. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
36. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics 31:2032–2034. https://doi.org/10.1093/bioinformatics/btv098.
37. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272. https://doi.org/10.1038/s41592-019-0686-2.