Title

Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission

Authors
Karthikeyan, Smruthi
Levy, Joshua I
De Hoff, Peter
et al.

Publication Date
2021

DOI
10.1101/2021.12.21.21268143

Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission

Smruthi Karthikeyan[1,#], Joshua I Levy[2,#], Peter De Hoff[3], Greg Humphrey[1], Amanda Birmingham[4], Kristen Jepsen[5], Sawyer Farmer[1], Helena M. Tubb[1], Tommy Valles[1], Caitlin E Tribelhorn[1], Rebecca Tsai[1], Stefan Aigner[3], Shashank Sathe[3], Niema Moshiri[6], Benjamin Henson[5], Abbas Hakim[3], Nathan A Baer[3], Tom Barber[3], Pedro Belda-Ferre[3], Marisol Chacón[3], Willi Cheung[3], Evelyn S Cresini[3], Emily R Eisner,[3] Alma L Lastrella[3], Elijah S Lawrence[3], Clarisse A Marotz[3], Toan T Ngo[3], Tyler Ostrander[3], Ashley Plascencia[3], Rodolfo A Salido[3], Phoebe Seaver[3], Elizabeth W Smoot[3], Daniel McDonald[1], Robert M Neuhard[7], Angela L Scioscia[8,9], Alysson M. Satterlund[10], Elizabeth H Simmons[11] , Christine M Aceves[2], Catelyn Anderson[2], Karthik Gangavarapu[2], Emory Hufbauer[2], Ezra Kurzban[2], Justin Lee[2], Nathaniel L Matteson[2], Edyth Parker[2], Sarah A Perkins[2], Karthik S Ramesh[2], Refugio Robles-Sikisaka[2], Madison A Schwab[2], Emily Spencer[2], Shirlee Wohl[2], Laura Nicholson[2], Ian H Mchardy[2], David P Dimmock[13], Charlotte A Hobbs[13], Omid Bakhtar[14], Aaron Harding[14], Art Mendoza[14], Alexandre Bolze[15], David Becker[15], Elizabeth T Cirulli[15], Magnus Isaksson[15], Kelly M Schiabor Barrett[15], Nicole L Washington[15], John D Malone[16], Ashleigh Murphy Schafer[16], Nikos Gurfield[16], Sarah Stous[16], Rebecca Fielding-Miller[17,18], Richard Garfein[17], Tommi Gaines[17], Cheryl Anderson[17], Natasha K Martin[18] , Robert Schooley[18] , Brett Austin[16], Stephen F Kingsmore[13] , William Lee[16], Seema Shah[16], Eric McDonald[16], Mark Zeller[2] , Kathleen M Fisch[4], Louise Laurent[3,9,19], Gene W Yeo[3,19,20], Kristian G Andersen[2,*], Rob Knight[1,6,21,*]

[#]equal contribution
*Senior author

[1] Department of Pediatrics, University of California San Diego, La Jolla, CA, USA
[2] Department of Immunology and Microbiology, The Scripps Research Institute, La Jolla, CA, USA
[3] Expedited COVID Identification Environment (EXCITE) Laboratory, Department of Pediatrics, University of California San Diego, La Jolla, CA, USA
[4] Center for Computational Biology and Bioinformatics, University of California San Diego, La Jolla, CA, USA
[5] Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA
[6] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
[7] Operational Strategic Initiatives, University of California San Diego, La Jolla, CA, USA
[8] Student Health and Well-Being, University of California San Diego, La Jolla, CA, USA
[9] Department of Obstetrics, Gynecology, and Reproductive Sciences, University of California San Diego, La Jolla, CA, USA
[10] Student Affairs, University of California San Diego, La Jolla, CA, USA
[11] Academic Affairs, University of California San Diego, La Jolla, CA, USA
[12] Scripps Health, San Diego, La Jolla, CA, USA
[13] Rady Children's Institute for Genomic Medicine, San Diego, CA, USA
[14] Sharp Healthcare, San Diego, CA, USA
[15] Helix, San Mateo, CA, USA
[16] County of San Diego Health and Human Services Agency, San Diego, CA, USA

[17] Herbert Wertheim School of Public Health and Human Longevity Science, University of California San Diego, La Jolla, CA
[18] Division of Infectious Disease and Global Public Health, University of California San Diego, La Jolla, CA, USA
[19] Sanford Consortium of Regenerative Medicine, University of California San Diego, La Jolla, CA
[20] Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA
[21] Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

Address correspondence to:

Rob Knight
Department of Pediatrics
University of California San Diego
9500 Gilman Drive, MC 0763
La Jolla, CA 92093, USA
robknight@ucsd.edu
+1 858-246-1184

Summary

As SARS-CoV-2 becomes an endemic pathogen, detecting emerging variants early is critical for public health interventions. Inferring lineage prevalence by clinical testing is infeasible at scale, especially in areas with limited resources, participation, or testing/sequencing capacity, which can also introduce biases. SARS-CoV-2 RNA concentration in wastewater successfully tracks regional infection dynamics and provides less biased abundance estimates than clinical testing. Tracking virus genomic sequences in wastewater would improve community prevalence estimates and detect emerging variants. However, two factors limit wastewater-based genomic surveillance: low-quality sequence data and inability to estimate relative lineage abundance in mixed samples. Here, we resolve these critical issues to perform a high-resolution, 295-day wastewater and clinical sequencing effort, in the controlled environment of a large university campus and the broader context of the surrounding county. We develop and deploy improved virus concentration protocols and deconvolution software that fully resolve multiple virus strains from wastewater. We detect emerging variants of concern up to 14 days earlier in wastewater samples, and identify multiple instances of virus spread not captured by clinical genomic surveillance. Our study provides a scalable solution for wastewater genomic surveillance that allows early detection of SARS-CoV-2 variants and identification of cryptic transmission.

Introduction

As SARS-CoV-2 transitions to endemicity, it continues to evolve, producing diverse new lineages[1]. Emerging variants of concern (VOCs) and variants of interest (VOIs) demonstrate increased transmissibility, disease severity, and/or immune escape[2]. Timely and accurate quantification of local prevalence of SARS-CoV-2 variants is thus essential for effective public health measures. However, existing strategies for variant detection based on virus genome sequencing of biospecimens obtained from clinical testing ("clinical genomic surveillance") are expensive, inefficient, and have sampling bias because of systemic healthcare disparities, particularly in poor and underserved communities[3,4].

In contrast, PCR-based wastewater surveillance of SARS-CoV-2 RNA is not subject to clinical testing biases and can track temporal changes in overall SARS-CoV-2 prevalence in a region [5–7], but cannot identify epidemiological transmission links or monitor lineages in the population. Virus genome sequencing from wastewater ("wastewater genomic surveillance") has the potential to cost-effectively capture community virus spread[8,9], acting as a surrogate to elucidate lineage geospatial distributions and track emerging SARS-CoV-2 variants (including new variants for which targeted assays do not yet exist), and provide genome sequence data needed for transmission network analysis and interpretation[10].

However, wastewater genomic surveillance is technically challenging[9]. Low viral loads, heavily fragmented RNA, and PCR inhibitors in complex environmental samples lead to poor sequencing coverage/quality[11]. Additionally, tools for SARS-CoV-2 lineage classification, such as pangolin[12] and UShER[13], were designed for clinical samples containing a single dominant variant, and cannot estimate relative abundances of multiple SARS-CoV-2 lineages in samples with virus mixtures such as wastewater.

Here, we report a high-resolution approach to study community virus transmission using wastewater genomic surveillance, leveraging several technical advances in wastewater virus concentration and nucleic acid sequencing, and a computational tool for resolving multiple SARS-CoV-2 lineages in short-read sequence data from a mixed sample (lineage deconvolution). Because places of communal living, such as university campuses, are considered key sites for virus spread and represent well-controlled and relatively isolated environments, they are ideal for comparing the relative utility of clinical and wastewater genomic surveillance[14]. Accordingly, we conducted a high-resolution, longitudinal wastewater genomic surveillance effort at the University of California San Diego (UCSD) campus, in parallel with clinical genomic surveillance from nasal swabs in the local community, from November 2020 to September 2021: ten months that effectively capture the surges in the region caused by the three main VOCs, Epsilon, Alpha and Delta[1].

Our wastewater genomic surveillance approach identified VOCs up to 2 weeks prior to detection through clinical genomic surveillance, even though a large proportion of clinical SARS-CoV-2 samples are sequenced in San Diego relative to other cities in the United States. In addition to providing a detailed history of community virus spread, wastewater genomic surveillance also identified multiple instances of cryptic community transmission not observed through clinical genomic surveillance. Matching wastewater and clinical genome sequences provided epidemiological information identifying specific transmission events. Our results demonstrate the viability of wastewater genomic surveillance at scale, enabling early detection and tracking of virus lineages and guiding clinical genomic surveillance efforts.

Results

To directly compare wastewater genomic surveillance to clinical surveillance, we conducted a large-scale SARS-CoV-2 genome sequencing study from wastewater samples collected daily from 131 wastewater samplers covering 360 campus buildings, in many cases reaching single building-level resolution. To identify epidemiological transmission links and monitor lineages in the population, we sequenced all SARS-CoV-2 positive clinical and wastewater samples from campus using a miniaturized tiled-amplicon sequencing approach. During this period of this study, we collected and analyzed 21,383 wastewater samples: 19,944 wastewater samples from the UCSD campus, and, for comparison, 1,439 wastewater samples from the greater San Diego area, including the Point Loma wastewater treatment plant (the primary wastewater treatment plant for the county with a catchment size of 2.3 million people) and 17 public schools spanning four San Diego school districts[15]. We compared sequencing of 600 campus wastewater samples to 759 genomes obtained from campus clinical swabs (46.2% of all positive tests on campus), all processed by the CALM and EXCITE CLIA labs at UCSD. In addition, we compared 31,149 genomes obtained from clinical genomic surveillance of the greater San Diego community to sequencing of 801 wastewater samples collected from San Diego county during the same period.

High-resolution spatial sampling reveals micro-scale community spread

We implemented a GIS (geographic information system)-enabled building-level wastewater surveillance system to cover 360 buildings on the UCSD campus (**Figure 1A**). During the period of daily wastewater sampling, approximately 10,000 students lived on campus and 25,000 individuals were on campus on a daily basis. We found that wastewater test positivity correlated strongly with clinical test positivity at the same site (**Figure 1B**), showing that wastewater effectively captures the community infection dynamics based on total viral load. This is also consistent with our past studies that showed SARS-CoV-2 RNA can be detected ~85% of the time downstream from buildings containing individuals known to be infected[8].

Unlike qPCR-based mutant surveillance, genomic surveillance using full-length virus genomes can detect which strains of SARS-CoV-2 are circulating in the population, and can identify potential transmission links between infected individuals[16,17]. To test the utility of wastewater genomic surveillance for studying virus spread in the community, we obtained near complete virus genomes for wastewater samples with cycle threshold (Ct) values as high as 38 (median genome coverage: 96.49% [75.67% - 100.00%], **Extended Data Figure 1**). However, using two common metrics of virus diversity, Shannon entropy (a measure of the uncertainty associated with randomly sampling an allele) and richness (the number of single nucleotide variant, or SNV, sites)[18], we found that SARS-CoV-2 genetic diversity is significantly greater in wastewater samples than clinical samples (**Figure 1C**, Mann-Whitney U test, p<0.001 for each). This suggests that multiple virus lineages, likely shed from different infected individuals, are often present in wastewater samples.

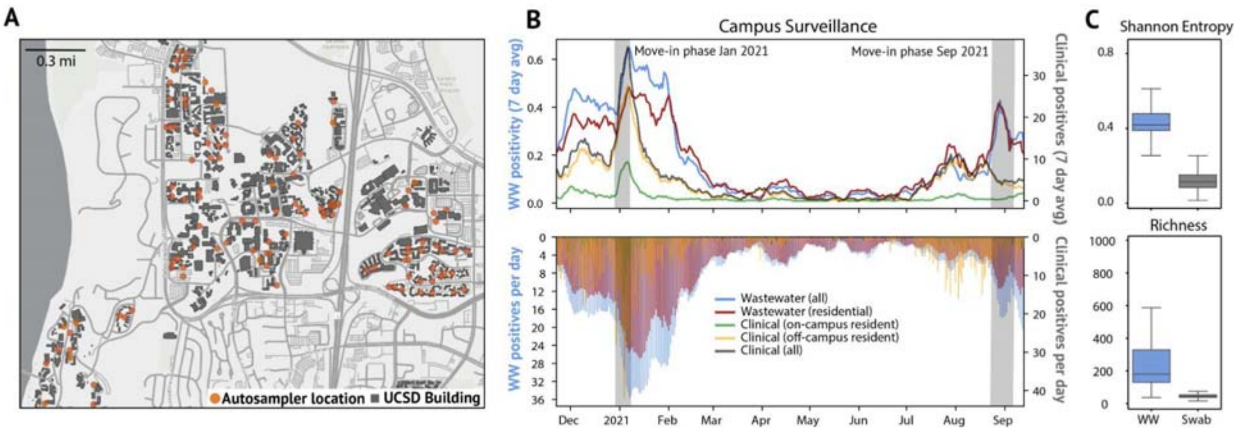


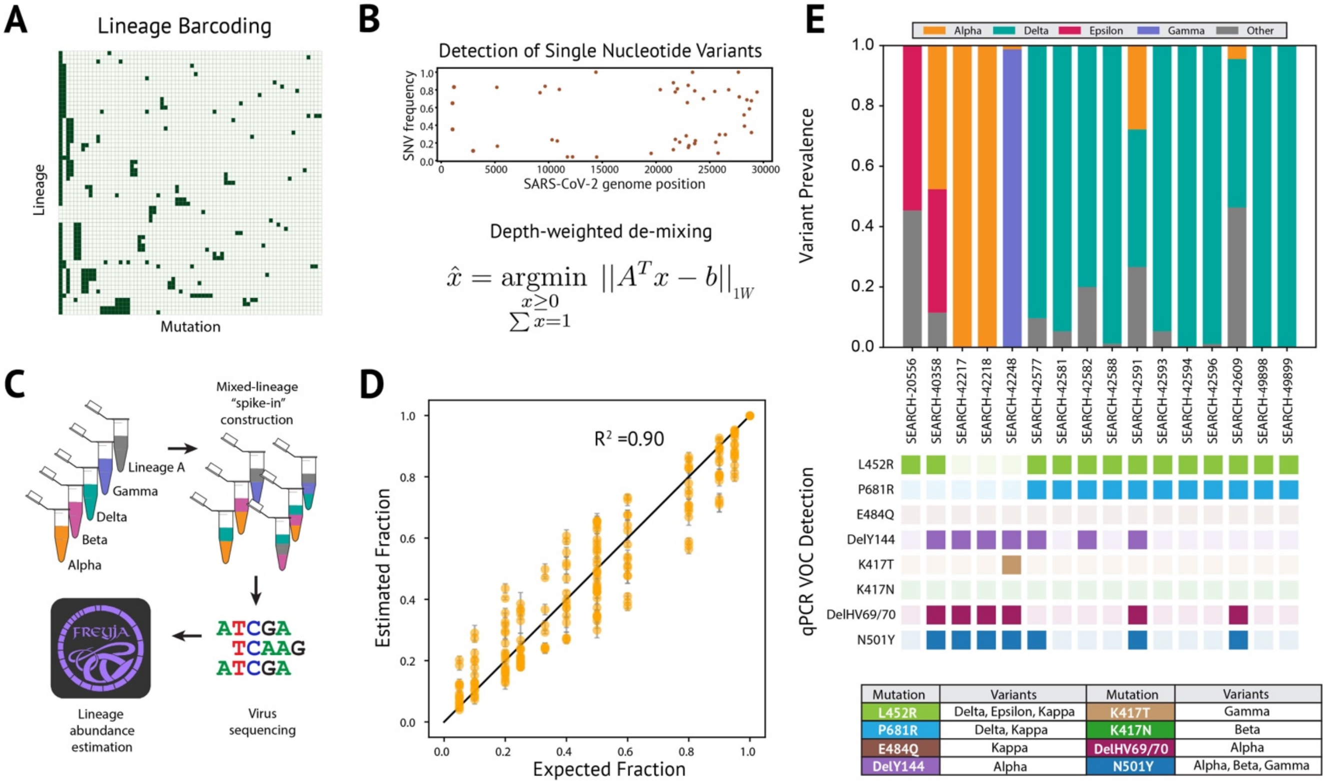
Figure 1: Campus sampling locations and SARS-CoV-2 testing statistics. A. Geospatial distribution of the 131 actively deployed wastewater autosamplers and the corresponding 360 university buildings on the campus sewer network. Building-specific data have been de-identified in accordance with university reporting policies. B. Campus wastewater and diagnostic testing statistics over the 295 day sampling period (WW = wastewater, positivity is the fraction of WW samplers with a positive qPCR signal). C.Virus diversity in wastewater and clinical samples: Boxplots of Shannon entropy (top) and richness (bottom) for each sample type.

Sample deconvolution robustly recovers the abundance of SARS-CoV-2 lineages in mixed samples

Wastewater systems aggregate stool, urine, and other biological waste products carrying viruses from multiple infected individuals in the community in a single location, allowing for sampling of virus mixtures that are representative of local lineage prevalence. However, existing methods for determining virus lineage from sequencing are intended for non-mixed clinical samples and can only be used to identify a single (dominant) lineage per sample.

To fully capture the virus diversity in community biospecimens, we developed Freyja, a tool to estimate the relative abundance of virus lineages in a mixed sample. Freyja uses a “barcode” library of lineage-defining mutations to represent each SARS-CoV-2 lineage in the global phylogeny[19](**Figure 2A**). To encode each sample, Freyja stores the SNV frequencies (proportion of reads at a site that contain the SNV) for each of the lineage-defining mutations (**Figure 2B, top**). Since SNV frequencies at positions with greater sequencing depth more accurately estimate the true mutation frequency, Freyja recovers relative lineage abundance by solving a depth-weighted least absolute deviation regression problem, a mixed sample analog of minimizing the edit distance between sequences and a reference (**Figure 2B, bottom**). To ensure results are meaningful, Freyja constrains the solution space such that each lineage abundance value is non-negative, and overall lineage abundance sums to one.

To validate Freyja, we sequenced “spike-in” synthetic mixtures from five key SARS-CoV-2 lineages (Lineage A, Beta, Delta, Epsilon, and Gamma) at proportions ranging from 5% to 100% in each sample, with between 1 and 5 different lineages per mixture (**Figure 2C**, and see **Table 1**). We found that Freyja robustly recovered the expected lineage abundances for all mixtures, even for lineages at 5% abundance (**Figure 2D**, and see **Extended Data Figure 2** for lineage specific predictions).To further validate Freyja, we used wastewater samples from the UCSD isolation dorms as well as Point Loma wastewater treatment plant, collection sites likely to contain mixed-lineage samples, to compare Freyja-detected lineages with qPCR testing for 8 mutations associated with different variants of concern (N501Y, DelHV69/70, DelY144, K417N, K417T, E484Q, P681R and L452R, **Figure 2E**). We found that Freyja consistently identified the same lineages as qPCR testing, but, as expected, also identified additional lineages with SNVs not included in our qPCR panel that were known to be circulating in San Diego at the time of collection. Combined, these results show that Freyja robustly estimates viral lineage abundance from samples containing a mixture of lineages, including synthetic virus mixtures and field wastewater collections.



Mutation	Variants	Mutation	Variants
L452R	Delta, Epsilon, Kappa	K417T	Gamma
P681R	Delta, Kappa	K417N	Beta
E484Q	Kappa	DelHV69/70	Alpha
DelY144	Alpha	N501Y	Alpha, Beta, Gamma

Figure 2: Sample deconvolution robustly recovers relative virus abundance. A. Subset of lineage defining mutation "barcode" matrix. Each row represents one lineage (out of >1000 lineages included in the UShER global phylogenetic tree), and individual nucleotide mutations are represented as columns. B. Single nucleotide variant frequencies obtained from iVar used for recovering relative abundance of each lineage. C. Schematic of the spike-in validation experiment. D. Depth-weighted de-mixing estimates of the virus abundance versus expected/known abundance. Details on lineage specific predictions are provided in **Supplemental Figure 2**. E. Comparison of wastewater sample deconvolution with VOC qPCR panel, with lookup table (bottom) showing amino acid mutations corresponding to each variant.

Detection of early and cryptic community transmission in wastewater

SARS-CoV-2 RNA concentrations in wastewater have been shown to be an early indicator of rising COVID-19 community incidence[8,20] (and see **Extended Data Figure 3A**), but whether wastewater can be used to detect emerging variants, including VOCs and VOIs, prior to their observation in clinical surveillance is unknown. To test if wastewater can enable early detection of emerging lineages, we applied Freyja to our wastewater sequencing data and compared the collection date of VOC positive samples from wastewater with the collection dates of samples from clinical genomic surveillance (**Figure 3A**). With only 2.6% as many sequenced wastewater samples as sequenced clinical samples, we detected the Alpha and Delta VOC lineages in wastewater genomic surveillance up to 14 days prior to their first detection in genomic clinical surveillance (Epsilon was circulating at the start of wastewater collection, and thus could not be detected early). Since emerging VOC lineages may evade immune responses or lessen the effectiveness of public health interventions[16], this early detection provides additional time to make necessary adjustments to existing countermeasures.

To test if wastewater genomic surveillance can identify changes in the abundance of circulating lineages, we compared VOC detection rates in clinical and wastewater sequencing over time. We found that both wastewater and clinical genomic surveillance tracked changes in lineage abundance, but increases in lineage detection frequency were generally observed first in wastewater surveillance. For example, for the Epsilon variant, which was first detected in San Diego in September of 2020, we observed increases in detection frequency in wastewater approximately 5 days prior to the corresponding increase in clinical genomic surveillance data (**Figure 3A,** see **Methods**). To study the effectiveness of wastewater genomic surveillance at a smaller community scale, we restricted our analysis to samples from the UCSD campus. We found that wastewater genomic surveillance consistently identified the three major VOCs (Epsilon, Alpha, and Delta) throughout their period of occurrence, despite detection gaps of one month or longer in clinical surveillance that included regular asymptomatic testing (**Figure 3B**). From mid-December to late-March, the Alpha variant was detected more than once per week on average in wastewater but was not detected by clinical surveillance. Similarly, wastewater surveillance detected continued Delta transmission from mid-April to mid-June, but no cases were identified by clinical surveillance.

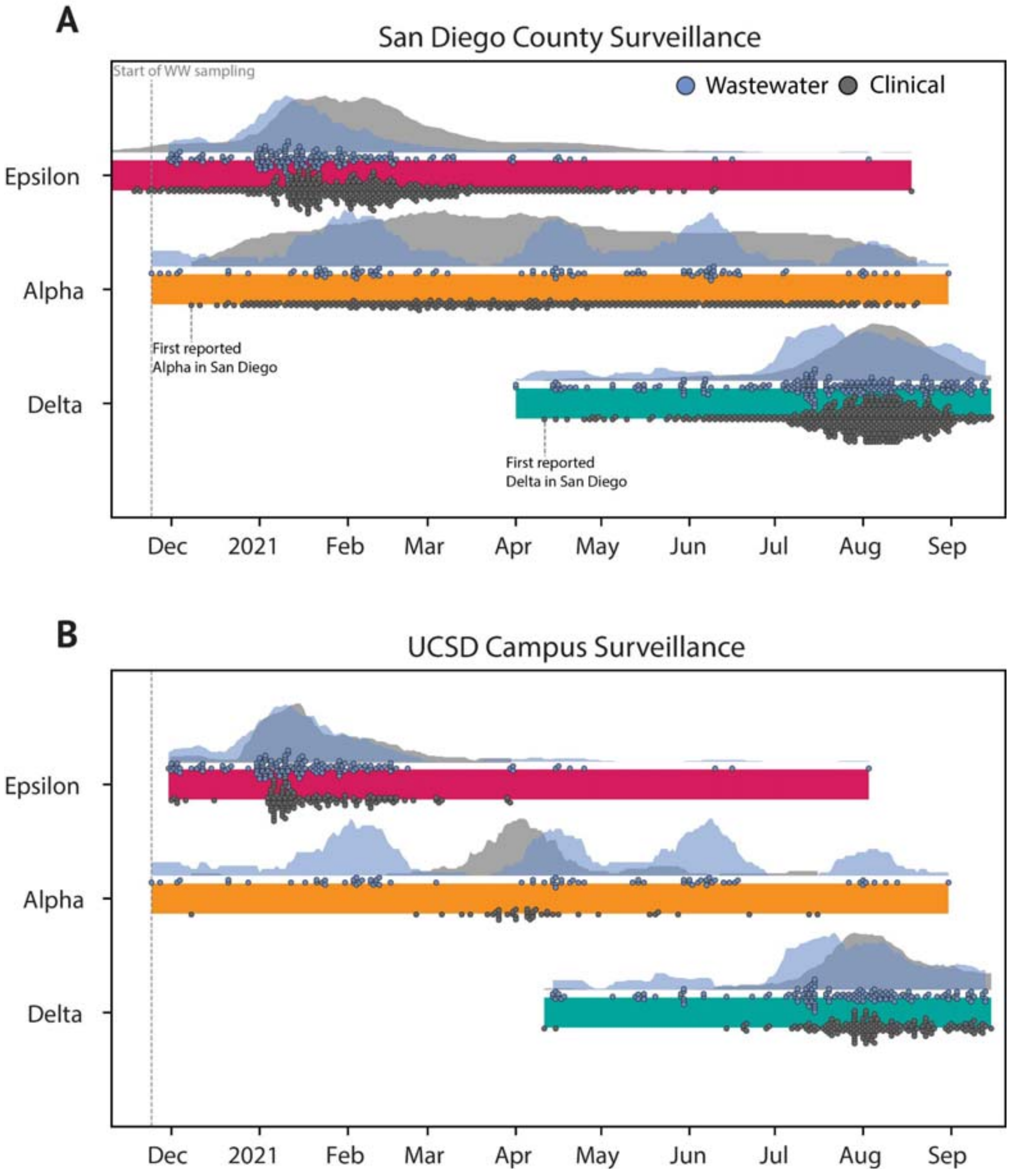


Figure 3: Freyja recovers early and cryptic transmission of SARS-CoV-2 variants of concern A. Timeline and normalized epidemiological curves for VOC detection in both wastewater and clinical sequences from San Diego County for the 3 major VOCs in circulation during the sampling period. Both Alpha and Delta are detected first in wastewater before clinical samples. Markers for clinical detections correspond to the ceiling of the detection count divided by 30, while wastewater markers correspond to a single detection. B. Timeline and epidemiological curves for VOC detection in the campus samples. Markers correspond to a single detection event for both clinical and wastewater surveillance. All wastewater detections correspond to an estimated VOC prevalence of at least 10%.

To study the effectiveness of wastewater surveillance in detecting and tracking other emerging variants, we aggregated all wastewater sequencing data to estimate the temporal profile of community lineage prevalence. We found that estimates of lineage abundance using wastewater enable early identification of other VOCs/VOIs, even for lineages that are rarely observed in clinical surveillance (**Figure 4**). For example, we detected the Mu (B.1.621) variant via wastewater genomic surveillance on July 27th, nearly four weeks prior to its first detection through clinical genomic surveillance on campus, on August 23rd (**Figure 4A,C**). However, despite persistent Mu detection in campus wastewater throughout July and early August, we did not detect the Mu variant in clinical or wastewater genomic surveillance on campus in September, suggesting that local community transmission did not continue. In more recent data, we identified the Omicron variant (B.1.1.529 and descendants) at an abundance of near 1% on November 27th, more than 1 week prior to the first clinical detection in San Diego on December 8th (**Extended Data Table 2**). To confirm these findings, we applied our VOC qPCR panel to the same samples and consistently detected two mutations associated with the Omicron variant (DelHV69/70 and N501Y) in samples detected after November 27th, while neither was detected in samples from earlier in November.

To test if Freyja continues to provide representative estimates of lineage prevalence for mixtures containing closely related lineages, we analyzed the rise of the Delta variant (B.1.617.2) and its sublineages (AY.*) in San Diego, from June-September 2021 (**Extended Data Figure 3B,C**). At both the UCSD campus and the Point Loma wastewater treatment plant, we identified the rapid emergence of B.1.617.2 and its sublineages (AY.*), along with low but persistent levels of the P.1 (Gamma) variant. The relative abundances of each of the variants were within 2-fold of prevalence estimates observed in clinical nasal swab data, suggesting that Freyja effectively identifies prevalence even for closely related lineages, both at the university and county-scale.

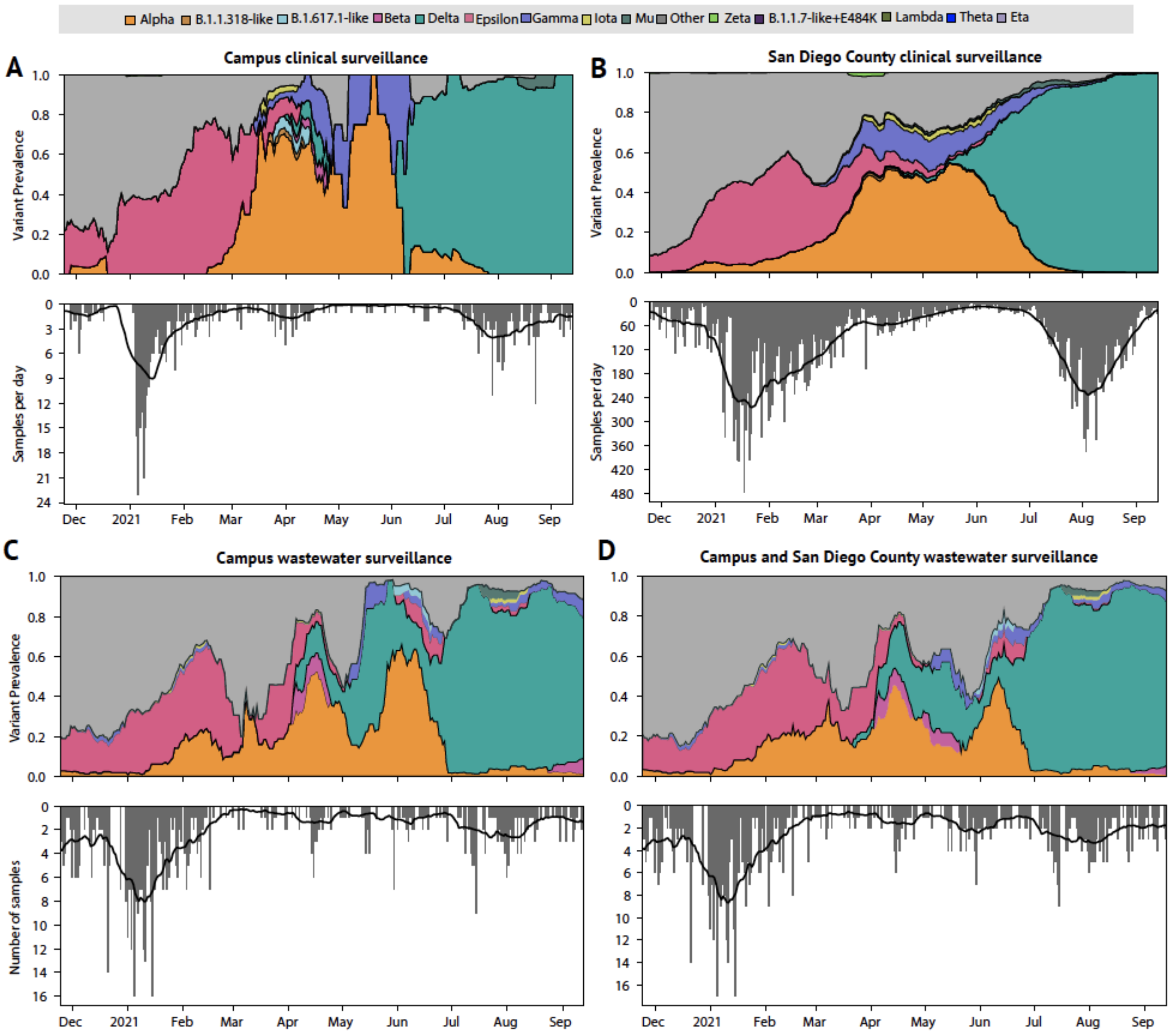


Figure 4: Deconvolution recovers a fine-grained estimate of virus population dynamics. A. Prevalence of SARS-CoV-2 variants in UCSD clinical surveillance, and B. Variant prevalence in all clinical samples collected in San Diego County. C,D. Variant prevalence in wastewater at UCSD as well as the greater San Diego County (includes wastewater samples collected from Point Loma wastewater treatment plant as well as public schools in the San Diego districts). Further analysis of Point Loma wastewater samples is shown in **Extended Data Figure 3**. All curves show rolling average, window ±10 days. “Other” contains all lineages not designated as VOCs. Bottom panels show number of sequenced samples per day.

Wastewater identifies both known and unknown history of campus infections

Phylogenetic analysis of virus genomes can be used to identify fine-scale spatial and temporal transmission networks, but it is unknown if wastewater can be used to further refine possible sites of transmission, elucidate transmission networks (“who-infected-whom”), or identify specific infected individuals[17]. To investigate the scale, structure, and timing of SARS-CoV-2 spread on campus, we reconstructed a maximum likelihood phylogenetic tree for each of the major VOCs using all high-quality genomes (see **Methods** for details) obtained from the UCSD campus, as well as reference sequences for each lineage obtained elsewhere in the United States (**Figure 5A-C**). In each tree, we identified many independent introductions, some of which led to extended transmission on campus. The resulting virus diversity among the VOCs present on campus enables ruling out of most transmission links and suggests campus virus spread consisted of many separate, small outbreaks.

To analyze the spatial structure of virus spread, we identified collection sites for wastewater sequences connected to transmission chains on campus, with building-specific resolution (**Figure 5 A-C,** *building specific transmission data available upon request*). We observed multiple small, linked outbreaks clustered in nearby buildings. Campus isolation protocol required students in congregate living to relocate to an isolation dorm and linkages in the isolation dorm wastewater samples reflected this co-location. We also found multiple instances of successive exactly matching sequences from wastewater collected from a single building, suggesting continued viral shedding from the same infected individuals, possibly due to extended shedding in stool[21,22].

To study the temporal delay between clinical and wastewater lineage detection, we compared collection times of sequences from campus wastewater that match sequences from campus clinical surveillance (including non-VOC lineages). We found 20 exact sequence matches and 103 near-matches (SNP distance of 3 or less) but did not observe any overall bias towards earlier or later detection in wastewater (**Figure 5D**), suggesting that on average, wastewater and clinical genomic surveillance identify a similar timing of individual detection events. However, in cases of delayed or missed detection by clinical surveillance, detections occur first in wastewater, further suggesting that wastewater genomic surveillance can reveal the presence of specific genome sequences prior to clinical surveillance.

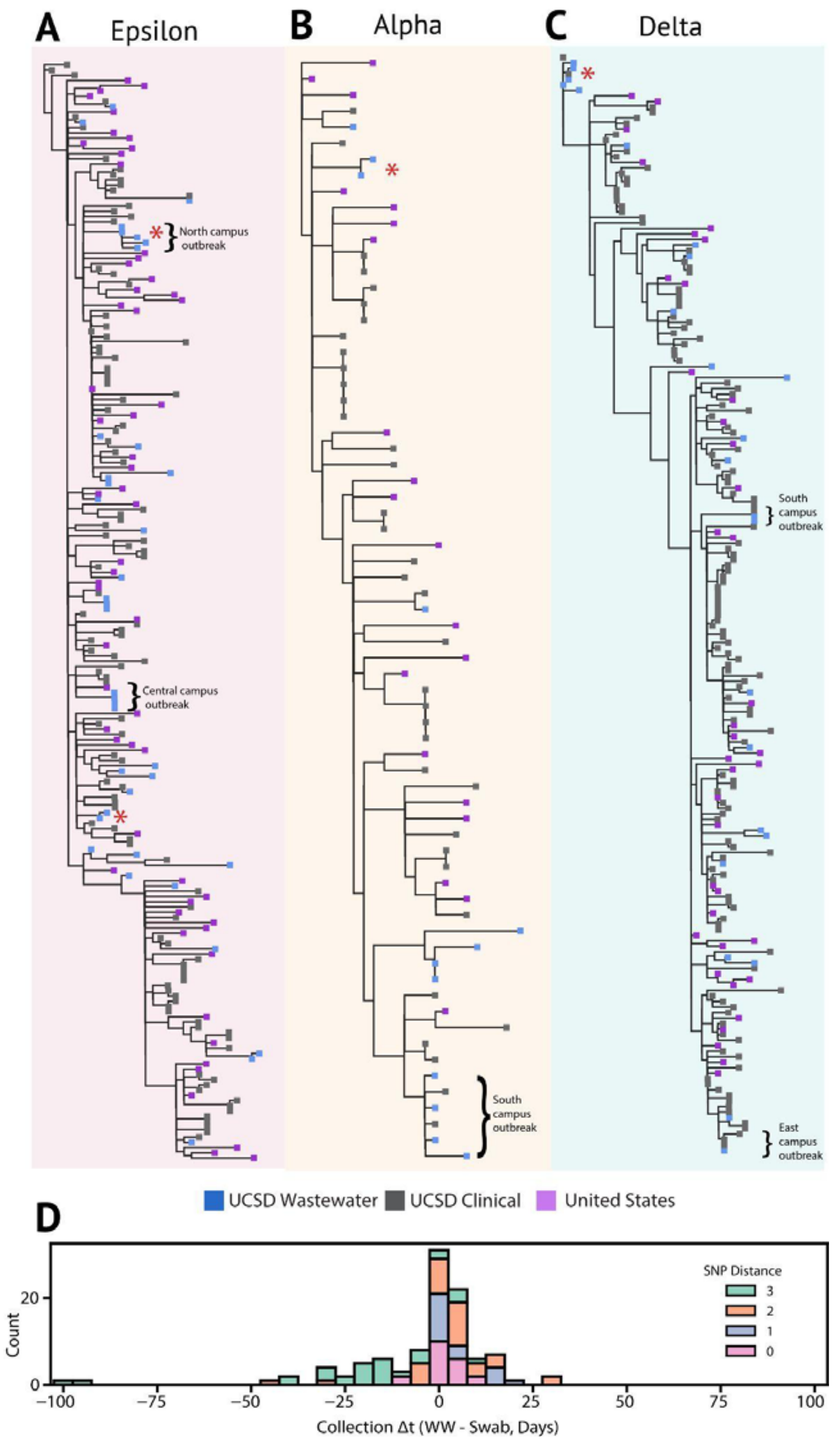


Figure 5: Wastewater identifies clinically known and unknown virus transmission. A-C. Maximum likelihood phylogenetic trees for each of the dominant variants of concern using high quality samples obtained at UCSD, as well as a representative set of sequences from the entire United States. Wastewater sequences from the same sampler that differ by 1 or fewer SNPs are denoted with a red asterisk. Location information is provided for select outbreaks. D. Pairwise comparison of collection date for matching and near-matching wastewater and nasal swab samples obtained at UCSD. Positive values indicate earlier collection in nasal swabs, and negative values indicate earlier detection in wastewater.

Discussion

We show that improved virus concentration from wastewater, coupled with a method for resolving multiple lineages from mixed samples, captures community virus lineage prevalence and enables early detection of emerging variants, often before observation in clinical surveillance. By sequencing both clinical and wastewater samples from the UCSD campus, we detect VOCs persistently in wastewater even when their appearance in clinical samples is intermittent. However, we also found occasions when rarer lineages, like B.1.1.318, were detected in clinical samples but not in wastewater. This is not unexpected on campus since many students living off-campus did not contribute to campus wastewater but were still clinically tested as part of testing mandates and policies. In the larger San Diego community context, this suggests that we may not be able to identify lineages circulating at low prevalence using a single wastewater collection site. In addition, we note that clinical sequences identified from the community may not be observable in the contributing catchment, as precise geolocation of all clinical samples was not possible. On the other hand, we also observed rare lineages in wastewater not seen in clinical samples from campus or the community. Since campus testing mandates are unable to capture all cases (e.g. fully vaccinated individuals were not required to test and not all community samples were sequenced), rare lineages can be missed.

The considerable benefits of wastewater surveillance may stem from biases in clinical testing, including population testing availability and compliance, university quarantine policies, and asymptomatic transmission, which may distort estimates of virus lineage prevalence from clinical samples. Wastewater offers less biased and more consistent viral lineage prevalence estimates, especially in areas with limited access and/or higher testing hesitancy rates. Since it requires considerably fewer samples, it is also more cost-effective than clinical testing, and could serve as a long-term passive surveillance tool. This is particularly important for developing public health interventions in low-resource and underserved communities, where widespread clinical genomic surveillance for SARS-CoV-2 remains limited.

Wastewater is an information-dense resource for estimating the prevalence of specific viral lineages, providing a community wide-snapshot not only of overall infection dynamics but of the rise and fall of specific VOCs. Our method, Freyja, deconvolutes these information-rich mixtures of virus lineages. For a large catchment area, such as San Diego's Point Loma wastewater treatment plant, which covers over 2 million residents, even limited sampling may accurately estimate lineage prevalence in the population and provide an early warning indicator of the rise of new VOCs. In addition, wastewater genomic surveillance with building-level resolution provides a detailed description of the structure and dynamics of community virus transmission, and can be used to better direct public health interventions.

As SARS-CoV-2 continues to evolve, the risk of new VOCs remains high and there is a growing need to identify these viruses ahead of their proliferation in the community. Accordingly, development of technologies that are cost-effective, reduce biases, and provide leading rather than trailing indicators of infection are essential to removing "blind spots" in our understanding of local virus dynamics. Although technical issues have made wastewater sequencing difficult to perform at scale, our key advances in virus concentration and sample deconvolution provide evidence that this approach is now viable. Continued improvements to sequencing turnaround speeds, lineage barcoding, and haplotype recovery from mixed samples will further accelerate efforts to achieve earlier identification of emerging variants and improve the precision and effectiveness of interventions.

Methods

Wastewater sampling

High-resolution spatial sampling at the campus level

131 wastewater autosamplers collecting 24h time-weighted composites were deployed across manholes or sewer cleanouts of 360 campus buildings. GIS (geographic information systems) informed analyses as well as agent-based network modeling of SARS-CoV-2 transmission on the UCSD campus enabled identification of most optimal locations for wastewater sampling. During the pilot phase (November 23-Dec 29th 2020), 68 samplers were prioritized to cover 239 residential buildings identified as the highest risk areas for large outbreaks on campus as a part of an observational study of wastewater monitoring in high-density buildings [23]. This was based on preliminary dynamic modeling which showed the largest potential outbreaks to occur within the largest residential buildings [8]. In addition to the observational study of wastewater monitoring in these high-density buildings, a cluster randomized study was also performed concurrently. This included a randomized modified version of a stepped wedge crossover design, in which there was random assignment of manholes for wastewater sampling. Clusters of manholes associated with residential buildings were randomized to receive wastewater monitors at one of two-time steps to evaluate the impact of wastewater monitoring on outbreak size in the associated buildings. During the same time period, all students in these residences were mandated to undergo weekly diagnostic testing which was used to validate the utility of building-level wastewater monitoring. Furthermore, on-campus residences were initially focused due to the relatively static nature of the population which enabled a more robust cross-validation of the sensitivity and efficacy of the wastewater surveillance. The coverage of wastewater surveillance was then increased to cover the rest of the campus buildings (including non-residential buildings on campus) from January 2021. Four of the deployed wastewater samplers covered the designated isolation and quarantine buildings on campus.

Wastewater composites were collected from the 131 samplers every day for the on-campus residence buildings and Monday through Friday for the nonresidential campus buildings. 19,944 wastewater samples were collected and analyzed for the presence of SARS-CoV-2 RNA via RT-qPCR between November 23rd 2020 and September 20th 2021. During this time, 9700 students lived in campus residences and 25,000 worked on campus on a daily basis. Between October 2020 to January 1st 2021, all on-campus residents were mandated to test on a bi-weekly basis and on a weekly basis from January 2nd 2021 (start of the Winter term). However, fully vaccinated individuals were not mandated to test on a regular basis. Automated, localized wastewater-triggered notifications were sent to the residents/employees of buildings associated with a positive wastewater signal which further led to a surge in testing uptake rates by 2 to 40-fold in the associated buildings.

Wastewater sampling at the county level

24h flow-weighted composites were collected thrice a week from the main pump station for the Point Loma wastewater treatment plant, the primary treatment plant serving the greater San Diego county with a catchment size of approximately 2.3 million. 96 wastewater samples were collected between February 24th 2021 to October 20th 2021.

Wastewater sample processing and viral genome sequencing

Sample processing

SARS-CoV-2 RNA was concentrated from 10ml of raw sewage and processed as described elsewhere[6]. In brief, the viral RNA was concentrated using an automated affinity capture magnetic hydrogel particle (Ceres Nanosciences Inc., USA) based concentration method after which the nucleic acid was extracted and sample eluted in 50uL of elution buffer. The extracted RNA was then screened for SARS-CoV-2 RNA via RT-qPCR for 3 gene targets (N1, N2 and E-gene). PMMoV (pepper mild mottle virus) was also screened to adjust for changes in load. To cross-validate the ability of the deconvolution tool in reliably resolving mixtures of strains in wastewater, the wastewater samples from the county as well as the ones from the isolation dorms on campus (where multiple infected individuals were isolating) were also run through a PCR panel targeting 8 mutations associated with the strains designated as VOCs. The mutations screened for in wastewater using RT-qPCR included N501Y, DelHV69/70, DelY144, K417N, K417T, E484Q, P681R and L452R (Promega Corp. Cat# CS3174B02).

Miniaturized wastewater SARS-CoV-2 amplicon sequencing

The Swift Normalase® Amplicon Panels (SNAP) kit (PN: SN-5X296 (core) COVG1V2-96 (amplicon primers), Integrated DNA Technologies, Coralville, IA) was used on RNA from wastewater samples that were positive for SARS-CoV-2 RNA to prepare the multiplex NGS amplicon libraries and indexed using the SN91384 series of dual indexing oligos, yielding up to 1536 index pairs per pool. A miniaturized version of the protocol was used with the following modifications: the Superscript IV VILO (Thermo Fisher, Carlsbad, CA) cDNA synthesis reaction was scaled down to ~1/12 the normal reaction volume with 0.333uL of enzyme mix and 1.333uL of RNA being used. The multiplex amplicon amplification and Ampure XP bead purification steps were scaled down ~1/6 the normal reaction volume. The Index adapter PCR reaction and Ampure XP bead purification steps were scaled down to ~2/13 the normal reaction volume. The final library resuspension volume was 29uL. 1uL of each library was pooled for an initial shallow NGS run on a MiSeq (Illumina, San Diego, CA) using a Nano flow cell. This equal volume pool was used to estimate the differential volumes required for similar read depths across samples using a NovaSeq SP or S4 flow cell (Illumina, San Diego, CA). Between 5uL and 0.2uL of library material, depending on the data provided from the MiSeq Nano run, was pipetted into a single pool for the NovaSeq run. Transfer volumes were capped at 5uL to reduce pipetting time and because these types of “high volume” samples typically contained a higher proportion of likely adapter dimers that inhibit flow cell performance for all samples. A Dragonfly Discovery (SPT Labtech, UK) was used to dispense reaction master mixes or water depending on the step. A BlueWasher (BlueCatBio, MA) was used for high throughput centrifugal 384-well plate washing during the AmpureXP bead reaction cleanup steps. An IKA MS3 Control linear plate mixer (IKA Works Inc, Wilmington, NC) set to 2600 RPM for 5’ was used to resuspend the AmpureXP beads during the rehydration steps. A Mosquito Genomics HV 16 channel robotic liquid handler (SPT Labtech, UK) was used to dispense the RNA, the reaction master mixes, and prepare the equal volume pools for the initial MiSeq Nano (Illumina, San Diego, CA) balancing runs. A Mosquito X1 single channel "hit picker" robotic liquid handler (SPT Labtech, UK) was used for the final library balancing for the NovaSeq (Illumina, San Diego, CA) NGS lanes.

Sequencing data were analyzed using the C-VIEW (COVID-19 VIral Epidemiology Workflow) platform for initial QC and SARS-CoV-2 lineage assignment and phylogenetics. In brief, sequencing reads are aligned with minimap2[24], and primer sequences trimming and quality filtering is applied using the iVar trim method[18]. Sequencing depth and single nucleotide variant (SNV) calls are obtained using samtools mpileup[25] and the iVar variants method[18].

Virus diversity

As reported previously[18], virus SNVs were used to characterize the populations derived from wastewater and clinical samples. Richness was defined as the total number of SNV sites, and mean Shannon entropy was defined as

$$H(p) = \frac{1}{N}\sum_{i=1}^{N} -p_i \log_2(p_i) - (1-p_i)\log_2(1-p_i).$$

Wastewater sample deconvolution

To infer relative abundance within a wastewater sample, we use a "barcode" matrix containing the lineage defining mutations for each known virus lineage,

$$A = \begin{bmatrix} a_{1,1} & \cdots & a_{1,N} \\ \vdots & \ddots & \vdots \\ a_{M,1} & \cdots & a_{M,N} \end{bmatrix}$$

where $a_{i,j}$ denotes the i-th lineage, at mutation j. Lineage defining mutations are obtained from the UShER global phylogenetic tree using the matUtils package[13]. Similarly, we let b and d encode the frequency of each mutation and the corresponding sequencing depth (using the log-transform $d_i = log_2(\text{depth}_i + 1)$ to adjust for large differences in depth across amplicons) ,

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}, d = \begin{bmatrix} d_1 \\ \vdots \\ d_N \end{bmatrix}.$$

We can then write this as a constrained (weighted) least absolute deviations problem

$$\hat{x} = \underset{\substack{x \geq 0 \\ \sum x = 1}}{\text{argmin}} ||A^T x - b||_{1W}, \quad \text{where} \quad ||\mu||_{1W} = \sum_{i=1}^{N} d_i |\mu_i|$$

which yields the "demixing" vector $\hat{x} = [\hat{x}_1 \ldots \hat{x}_M]$ that specifies the relative abundances of each of the known haplotypes. Analysis was only performed on samples with greater than 70% coverage, with the exception of March samples from UCSD for which all samples with greater than 50% coverage were used. Constrained minimization is performed in Python using the cvxpy convex optimization package[26,27]. Mapping of lineages to variant WHO lineages (VOCs, VUMs, etc.) is performed using curated lineage data from outbreak.info[1].

Spike-in mixture experiment

RNA was isolated from supernatants of a mammalian cell culture infected with one of five strains of SARS-CoV-2. (A, B.1.1.7, B.1.351, P.1, or B.1.617.2).

RNA concentration standardization

Virus concentration was quantified by the UCSD EXCITE COVID testing laboratory using the Thermo COVID-19 Test kit (PN:A47814, Thermo Scientific Corporation, Carlsbad, CA). The median Cq values (N-gene, Orf1ab, & S-gene (where applicable)) was calculated and used to determine how much the RNA needed to be diluted with water to reach a Cq value of 23. A post dilution RT-qPCR reaction was performed and used to calculate the final dilution of the more concentrated samples to the new target value of Cq 23.296. The number of freeze thaw cycles between RNA samples was kept the same.

Virus Mixing

RNA standardized in the prior section was used to make a volumetric mixing array (final volume 10uL) using a Mosquito X1 HV robotic liquid handler (SPT Labtech, UK). Pairwise mixes of 5:95, 10:90, 20:80, 60:40, and 50:50 were made for each virus strain and in both directions. Equal mixes (20%) for each of the five test strains were made. 25% mixes and 33% mixes were made for a subset of possible combinations and controls of 100:0 were prepared. See Table 1 for complete array.

Estimation of delay in detection frequency

Estimation of the lag time between epidemiological curves for wastewater and clinical surveillance of the Epsilon variant in San Diego was performed by identifying the shift with maximal cross-correlation. All time points leading up to the time of initial peak in detection frequency were included for both wastewater and clinical data.

Phylogenetic analyses

Reconstruction of maximum likelihood trees was performed on all SARS-CoV-2 VOC genomes with 10x genome coverage >95% and quality score >20 obtained from UCSD campus sampling, using IQtree[28]. This analysis included 150 (112 clinical, 38 wastewater) Epsilon, 49 (37 clinical, 12 wastewater) Alpha, and 160 (136 clinical, 24 wastewater) Delta lineage genomes from UCSD, in addition to 60 Epsilon, 20 Alpha, and 39 Delta randomly selected genomes from elsewhere in the United States. We also masked known homoplasic sites prior to tree reconstruction[29]. Analysis of temporal comparison was performed on 608 samples (443 clinical, 165 wastewater, all lineages were included) with 10x genome coverage >95% and quality score >20 from UCSD. Sample collection SNP distances were calculated without considering ambiguous bases and gaps.

Code availability

Freyja is hosted publicly on github (https://github.com/andersen-lab/Freyja) and is available under a BSD-2-Clause License. Freyja is accessible as a package via bioconda (https://bioconda.github.io/recipes/freyja/README.html) in container form via dockerhub (https://hub.docker.com/r/andersenlabapps/freyja). COVID-19 VIral Epidemiology Workflow (C-VIEW) is available at https://github.com/ucsd-ccbb/C-VIEW as an open-source, end-to-end workflow for viral epidemiology focused on SARS-CoV-2 lineage assignment and phylogenetics.

Data Availability

Consensus sequences from clinical and wastewater surveillance are all available on GISAID. Spike-in sequencing data is available via google cloud (https://console.cloud.google.com/storage/browser/search-reference_data).

Acknowledgements

We thank Dr. Alexander T. Yu of the California Department of Public Health (CDPH) for his valuable suggestions and comments on the manuscript; Laralyn Asato and the Microbiology Lab at the SD Public Utilities Department for providing us with county wastewater samples. We thank UC San Diego's Return to Learn (RTL) program for funding the campus-wide wastewater surveillance efforts. We also thank the rest of the RTL leadership team and Bradley Sollenberger of the Operational Strategic Initiatives (OSI) team for ensuring the smooth functioning of the program. We also thank Jason Kayne, Rich Cota, Jesus Ortiz, and the Facilities management team (FM) at UCSD, Joseph Mayer from the Center for Aerosol Impacts on Chemistry of the Environment (CAICE) and Luke Arnold of the Campus Research Machine Shop (CRMS) for assistance with the installation and operation of the autosamplers; Robbie Jacobs, Shawn Knepple and their team at UCSD Logistics for assisting with our daily sampling efforts; Christopher Longhurst and the UCSD Health Information Services team; Brett Pollak and the UCSD Information Technology Services team for assisting with the daily notifications; Office of Academic Affairs for contact tracing and targeted campus messaging assistance; Jana Severson, Patrick Hochstein, Hemlata Jhaveri, the UCSD HDH team, and the UCSD Environmental Health and Safety (EHS) personnel; Jack Gilbert and the Microbiome Sample Processing Core at UCSD for access to qPCR equipment. We also thank the CDC SPHERES consortium, SEARCH (San Diego Epidemiology and Research for COVID Health) Alliance and members of the Andersen lab for discussion and help with logistics. We thank the healthcare workers, frontline workers, and patients who made the collection of this SARS-CoV-2 dataset possible and all those who made genomic data available for analysis via GISAID.

Funding

This work has been funded by CDC BAA contracts 75D30121P10258 (Helix) and 75D30120C09795 (G.W.Y., R.K., L.C.L., and K.G.A.), NIH NIAID 3U19AI135995-03S2 (K.G.A.), U19AI135995 (K.G.A.), U01AI151812 (K.G.A.), NIH NCATS UL1TR002550 (K.G.A.), the Conrad Prebys Foundation (K.G.A.), NIH 5T32AI007244-38 (J.I.L.), NIH Pioneer Grant 1DP1AT010885 (R.K), NSF RAPID 2029069 (R.K.), San Diego County Health and

Human Services Agency (R.F.M), NIH S10OD026929 (K.J.). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention.

Ethics declarations

The University of California San Diego Institutional Review Boards (IRB) provided human subject protection oversight of the of the data obtained by the EXCITE lab for the campus clinical samples (IRB approval # 210699). All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived, and any sample identifiers included were de-identified. The wastewater component of this project was discussed with our Institutional Review Board, and was not deemed to be human subject research, as it did not record personally identifiable information.

Author Contributions

Conceptualization: RK, KGA
Methodology: SK, JIL, NKM, PDH, AK, SS, KMF, AB, LCL, GWY, KGA, RK
Software: JIL, DM, NM, KMF, AB, BH, SS, KG, NLM, KSR, CMA, EH
Formal Analysis: SK, JIL, PDH, GH
Investigation: SK, JIL, NM, SF, HMT, TV, CET, RT, NAB, TB, MC, WC, ESC, ERE, AH, GH, ALL, EL, TTN, TO, AP, RAS, PS, PBF, EWS, SA, PDH, CAM, LCL, GWY, CA, EK, MAS, SAP, JL, EP, MZ, ES, RFK, TG, RG, KGA, RK
Resources: CA, NKM, RMN, RS, EHS, AMS, SFK, DPD, CAH, AM, SS, BA, SS, NG, JDM, EM, IAM, AH, OB, AM, AB, KMSB, ETC, NLW, WL, MI, DB, LN, SW, MZ, RRS, RFM, TG, RG
Data curation: SK, JIL, PDH, GH, SF, HMT, CET, RT, TV, PDH, AB, NM, KMF
Writing – Original Draft: SK, JIL, KGA, RK
Writing – Review and editing: all authors
Visualization: SK, JIL, PDH
Supervision: RMN, NKM, RS, ALS, EHS, AMS, PDH, LCL, GWY, KGA, RK
Project administration: RMN, NKM, RS, ALS, EHS, AMS, PDH, LCL, GWY, KGA, RK
Funding acquisition: RK, KGA

References

1. Julia L. Mullen, Ginger Tsueng, Alaa Abdel Latif, Manar Alkuzweny, Marco Cano, Emily Haag, Jerry Zhou, Mark Zeller, Emory Hufbauer, Nate Matteson, Kristian G. Andersen, Chunlei Wu, Andrew I. Su, Karthik Gangavarapu, Laura D. Hughes, and the Center for Viral Systems Biology. outbreak.info. *outbreak.info* https://outbreak.info/ (2021).

2. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* **19**, 409–424 (2021).

3. Reitsma, M. B. *et al.* Racial/Ethnic Disparities In COVID-19 Exposure Risk, Testing, And Cases At The Subcounty Level In California. *Health Aff.* **40**, 870–878 (2021).

4. Lieberman-Cribbin, W., Tuminello, S., Flores, R. M. & Taioli, E. Disparities in COVID-19 Testing and Positivity in New York City. *Am. J. Prev. Med.* **59**, 326–332 (2020).

5. Hata, A., Hara-Yamamura, H., Meuchi, Y., Imai, S. & Honda, R. Detection of SARS-CoV-2 in wastewater in Japan during a COVID-19 outbreak. *Sci. Total Environ.* **758**, 143578 (2021).

6. Karthikeyan, S. *et al.* High-Throughput Wastewater SARS-CoV-2 Detection Enables Forecasting of Community Infection Dynamics in San Diego County. *mSystems* **6**, (2021).

7. Randazzo, W. *et al.* SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water Res.* **181**, 115942 (2020).

8. Karthikeyan, S. *et al.* Rapid, Large-Scale Wastewater Surveillance and Automated Reporting System Enable Early Detection of Nearly 85% of COVID-19 Cases on a University Campus. *mSystems* **6**, e0079321 (2021).

9. Mercer, T. R. & Salit, M. Testing at scale during the COVID-19 pandemic. *Nat. Rev. Genet.* **22**, 415–426 (2021).

10. Crits-Christoph, A. *et al.* Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *MBio* **12**, (2021).

11. Baaijens, J. A. *et al.* Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. *medRxiv* (2021) doi:10.1101/2021.08.31.21262938.

12. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* **5**, 1403–1407 (2020).

13. Turakhia, Y. *et al.* Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).

14. Walke, H. T., Honein, M. A. & Redfield, R. R. Preventing and Responding to COVID-19 on

College Campuses. *JAMA* **324**, 1727–1728 (2020).

15. Fielding-Miller, R. K. *et al.* Wastewater and surface monitoring to detect COVID-19 in elementary school settings: The Safer at School Early Alert project. (2021) doi:10.1101/2021.10.19.21265226.

16. Ladner, J. T., Grubaugh, N. D., Pybus, O. G. & Andersen, K. G. Precision epidemiology for infectious disease control. *Nat. Med.* **25**, 206–211 (2019).

17. Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat Microbiol* **4**, 10–19 (2019).

18. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8 (2019).

19. McBroome, J. *et al.* A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol. Biol. Evol.* (2021) doi:10.1093/molbev/msab264.

20. Peccia, J. *et al.* Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nat. Biotechnol.* **38**, 1164–1167 (2020).

21. Xu, Y. *et al.* Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nat. Med.* **26**, 502–505 (2020).

22. Wu, Y. *et al.* Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *Lancet Gastroenterol Hepatol* **5**, 434–435 (2020).

23. Goyal, R., Hotchkiss, J., Schooley, R. T., De Gruttola, V. & Martin, N. K. Evaluation of SARS-CoV-2 transmission mitigation strategies on a university campus using an agent-based network model. *Clin. Infect. Dis.* (2021) doi:10.1093/cid/ciab037.

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

25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

26. Diamond, S. & Boyd, S. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *J. Mach. Learn. Res.* **17**, (2016).

27. Agrawal, A., Verschueren, R., Diamond, S. & Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision* **5**, 42–60 (2018).

28. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

29. Issues with SARS-CoV-2 sequencing data. https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 (2020).

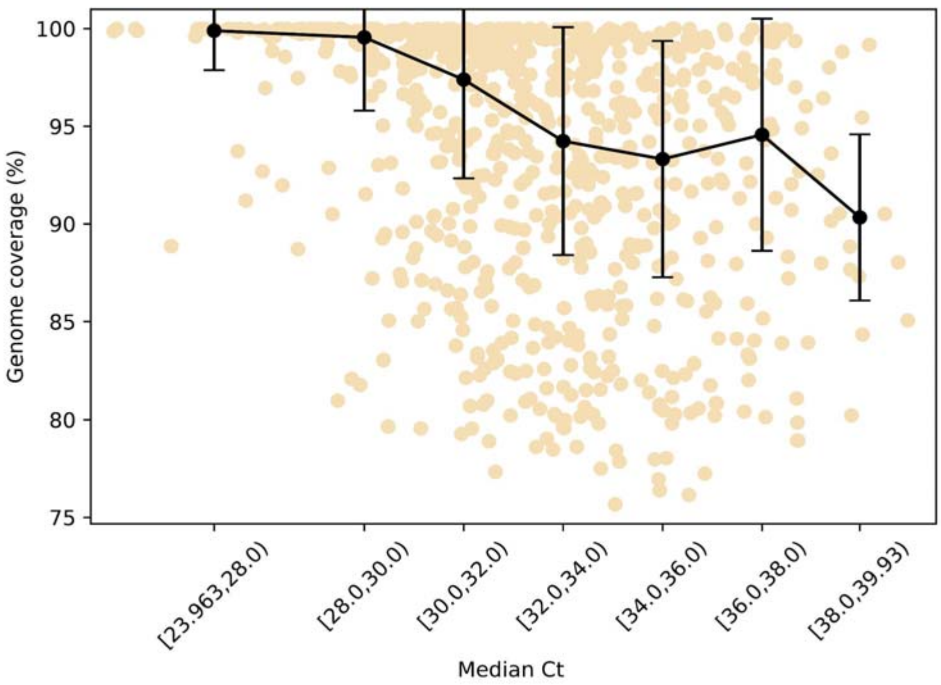
Extended Data:

Extended Data Table 1: Platemap of spike-in mixtures used for method validation

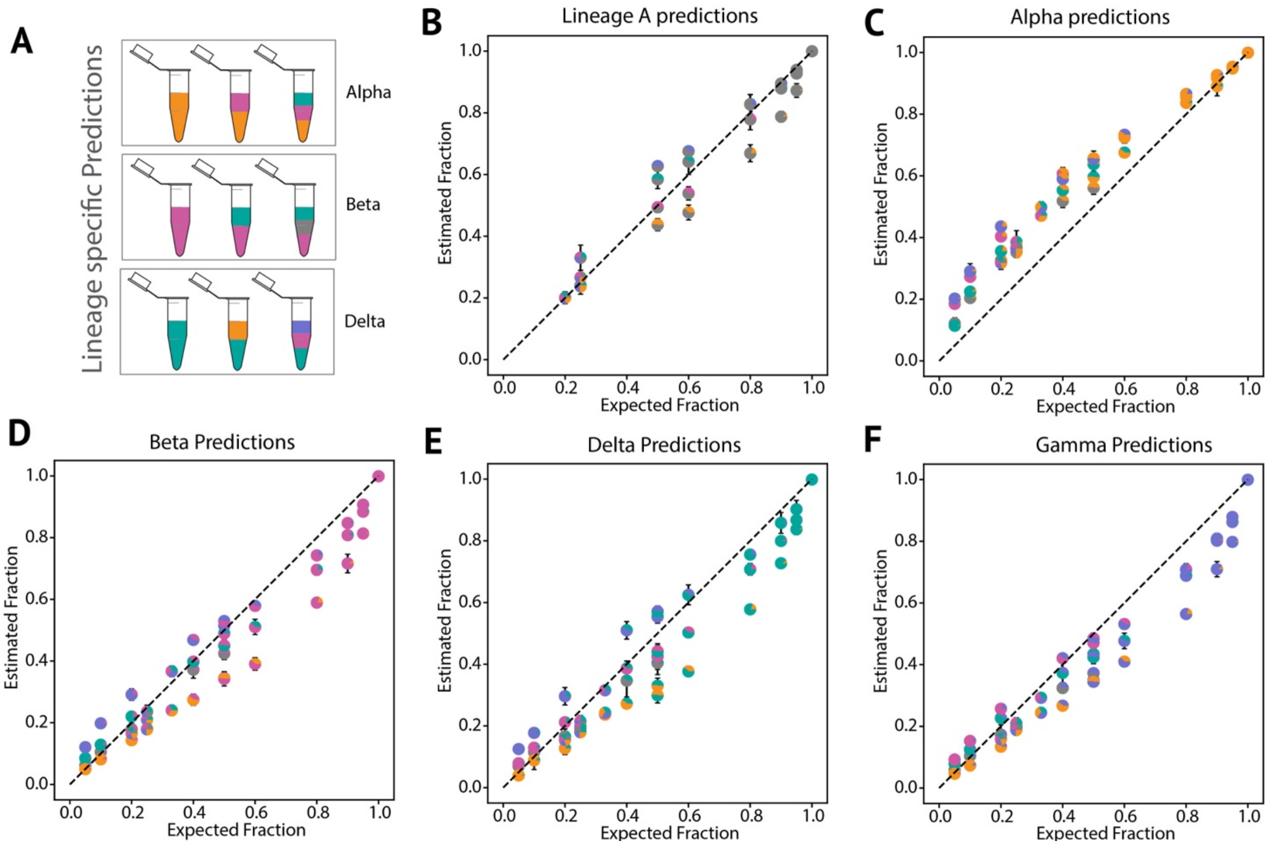
	1	2	3	4	5	6
A	5% Delta: 95% A	10% Delta: 90% A	20% Delta: 80% A	40% Delta: 60% A	50% Delta: 50% A	100% A
B	5% Delta: 95% Beta	10% Delta: 90% Beta	20% Delta: 80% Beta	40% Delta: 60% Beta	50% Delta: 50% Beta	100% Delta
C	5% Delta: 95% Gamma	10% Delta: 90% Gamma	20% Delta: 80% Gamma	40% Delta: 60% Gamma	50% Delta: 50% Gamma	100% Beta
D	5% Delta: 95% Alpha	10% Delta: 90% Alpha	20% Delta: 80% Alpha	40% Delta: 60% Alpha	50% Delta: 50% Alpha	100% Gamma
E	5% Beta: 95% A	10% Beta: 90% A	20% Beta: 80% A	40% Beta: 60% A	50% Beta: 50% A	100% Alpha
F	5% Beta: 95% Delta	10% Beta: 90% Delta	20% Beta: 80% Delta	40% Beta: 60% Delta	50% Beta: 50% Delta	20% A: 20% Delta: 20% Beta: 20% Gamma: 20% Alpha
G	5% Beta: 95% Gamma	10% Beta: 90% Gamma	20% Beta: 80% Gamma	40% Beta: 60% Gamma	50% Beta: 50% Gamma	25% Delta: 25% Beta : 25% Gamma: 25% Alpha
H	5% Beta: 95% Alpha	10% Beta: 90% Alpha	20% Beta: 80% Alpha	40% Beta: 60% Alpha	50% Beta: 50% Alpha	25% Delta: 25% Beta: 25% Gamma: 25% A
I	5% Gamma: 95% A	10% Gamma: 90% A	20% Gamma: 80% A	40% Gamma: 60% A	50% Gamma: 50% A	25% Delta: 25% Beta: 25% A: 25% Alpha
J	5% Gamma: 95% Delta	10% Gamma: 90% Delta	20% Gamma: 80% Delta	40% Gamma: 60% Delta	50% Gamma: 50% Delta	25% Delta: 25% A: 25% Gamma: 25% Alpha
K	5% Gamma: 95% Beta	10% Gamma: 90% Beta	20% Gamma: 80% Beta	40% Gamma: 60% Beta	50% Gamma: 50% Beta	25% A: 25% Beta: 25% Gamma: 25% Alpha
L	5% Gamma: 95% Alpha	10% Gamma: 90% Alpha	20% Gamma: 80% Alpha	40% Gamma: 60% Alpha	50% Gamma: 50% Alpha	33% Delta: 33% Beta: 33% Gamma
M	5% Alpha: 95% A	10% Alpha: 90% A	20% Alpha: 80% A	40% Alpha: 60% A	50% Alpha: 50% A	33% Delta: 33% Beta: 33% Alpha
N	5% Alpha: 95% Delta	10% Alpha: 90% Delta	20% Alpha: 80% Delta	40% Alpha: 60% Delta	50% Alpha: 50% Delta	33% Delta: 33% Alpha: 33% Gamma
O	5% Alpha: 95% Beta	10% Alpha: 90% Beta	20% Alpha: 80% Beta	40% Alpha: 60% Beta	50% Alpha: 50% Beta	33% Alpha: 33% Beta: 33% Gamma
P	5% Alpha: 95% Gamma	10% Alpha: 90% Gamma	20% Alpha: 80% Gamma	40% Alpha: 60% Gamma	50% Alpha: 50% Gamma	Neg

Extended Data Table 2: Omicron surveillance at Point Loma Wastewater Treatment Plant

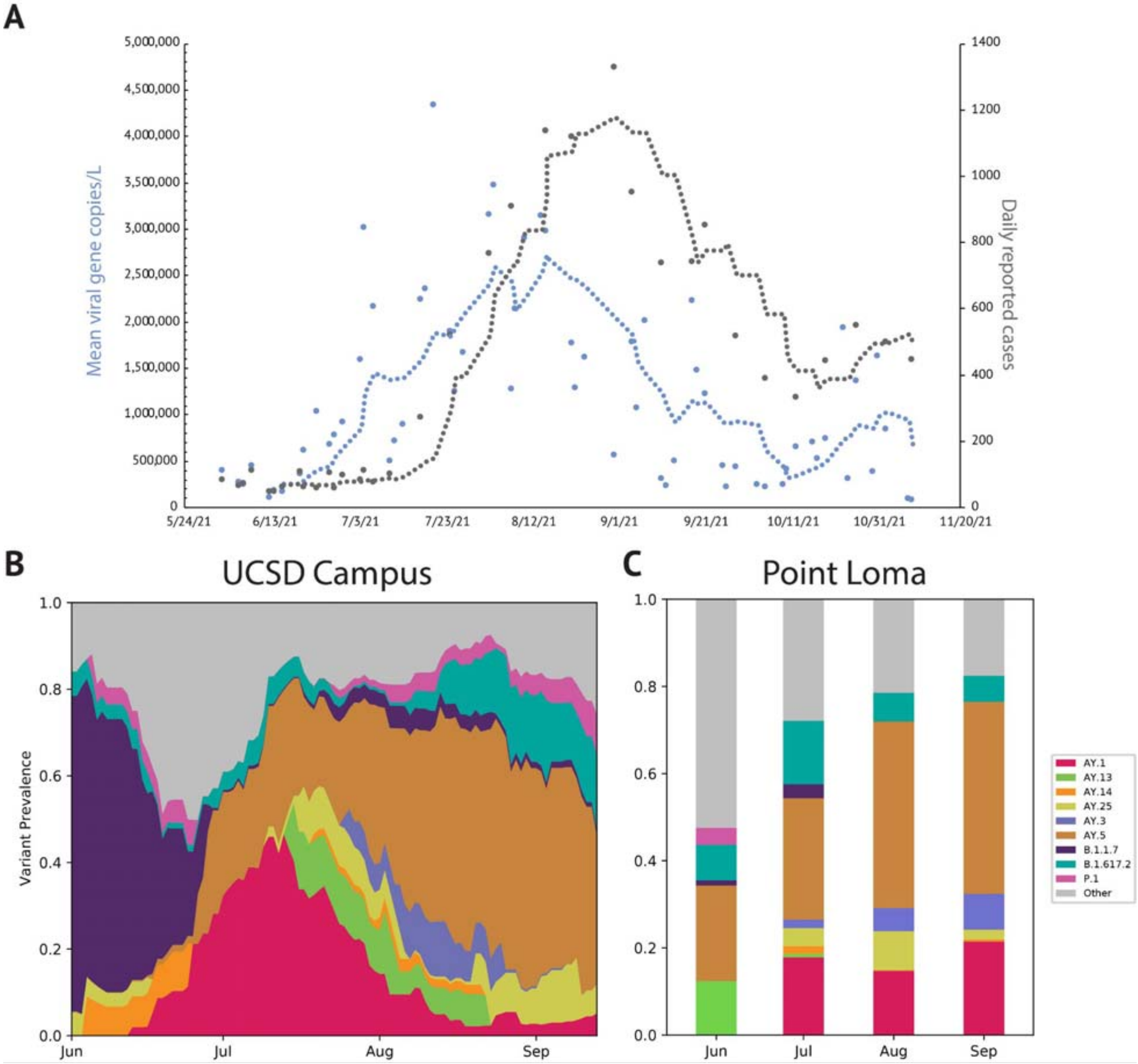
Collection Date	Estimated Omicron Abundance	qPCR Detection		
		DelHV69/70	N501Y	P681R
11/8/21	NS			x
11/13/21	NS			x
11/14/21	NS			x
11/16/21	NS			x
11/20/21	NS			x
11/21/21	0			x
11/27/21	0.6%	x	x	x
11/28/21	0 (Low Coverage)	x	x	x
12/1/21	0.8%	x	x	x



Extended Data Figure 1: Relationship between genome coverage and cycle threshold. 10X genome coverage remains high, even for Ct values of nearly 38. Points indicate median value in each bin, while error bars indicate the median absolute deviation.



Extended Data Figure 2: Lineage-specific prediction of variant abundance in spike-in validation samples. A. Schematic of "spike-in" sample design. B-F. Lineage specific prediction. Proportions of each lineage in the sample are shown as a pie chart marker (Grey = Lineage A, Orange = Alpha, Pink = Beta, Turquoise = Delta, and Purple = Gamma) with error bars indicating the standard deviation from the mean, across four replicates.



Extended Data Figure 3: The rise of the Delta variant during Summer 2021. A. Mean SARS-CoV-2 viral gene copies/L of raw sewage (blue) collected from the Point Loma Wastewater Treatment Plant and caseload (gray) reported by the county during the same period. SARS-CoV-2 concentrations were normalized by PMMoV (pepper mild mottle virus) concentration to adjust for load changes. B. Lineage distribution in UCSD campus wastewater. C. Monthly lineage averages for wastewater collected at Point Loma Wastewater Treatment Plant during the Delta surge (N= 5, 20, 25, 7)