Enhanced Detection of Recently Emerged SARS-CoV-2 Variants of Concern in Wastewater

Nicolae Sapoval[1], Esther Lou[2], Loren Hopkins[3,4], Katherine B Ensor[4], Rebecca Schneider[3], Todd J Treangen[1#*], Lauren B Stadler[2#*]

[1]Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005, USA
[2]Department of Civil and Environmental Engineering, Rice University, 6100 Main Street, Houston, TX 77005, USA
[3]Houston Health Department, 8000 N. Stadium Dr., Houston, TX 77054
[4]Department of Statistics, Rice University, 6100 Main Street, Houston, TX 77005, USA
*Corresponding authors
#These authors share senior authorship

Abstract. SARS-CoV-2 RNA shedding in stool enabled wastewater surveillance for the genetic material of the virus. With the emergence of novel variants of concern and interest it becomes increasingly important to track arrival and spread of these variants. However, most current approaches rely on the manually curated lists of mutations phenotypically associated with the variants of concern. The resulting data has many overlaps between distinct variants leading to less specific characterization of complex sample mixtures that result from wastewater monitoring. In our work we propose a simple and specific method for characterization of wastewater samples by introducing the concept of quasi-unique mutations. Our approach is data driven and results in earlier detection and higher resolution of variants of concern emergence patterns in wastewater data.

Importance. Wastewater-based epidemiology has emerged as a powerful tool for public health response to the SARS-CoV-2 pandemic. As wastewater is a pooled, community sample of all persons contributing to the waste stream, there are several challenges in using sequencing information from wastewater samples to detect variants. Wastewater typically will consist of fragmented genomes from multiple, circulating variants. While it is straightforward to call the

mutations present in a wastewater sample, it is more challenging to call the presence of variants that are defined by a set of characteristic mutations, particularly when mutations are shared among many circulating variants. Hence, we present a novel approach for screening for variants of concern in wastewater. Our computational approach introduces the concept of a "quasi-unique mutation" corresponding to a given PANGO lineage. We show that our method enables detection of the emergence of variants of concern in communities, providing a new approach for wastewater-based epidemiology of SARS-CoV-2.

Observation

The prevalence of SARS-CoV-2 RNA in stool has opened the door to several SARS-CoV-2 wastewater surveillance efforts across the globe (1, 2). The main goal of these efforts is tracking levels of SARS-CoV-2 in the wastewater, and screening for the presence of the variants of concern (VoCs) and variants of interest (VoIs) in the samples (3, 4). Several sequencing and RT-qPCR techniques have been employed in the process. They can be summarized by three major categories: (a) direct metagenomic sequencing of the wastewater samples, (b) targeted amplification and sequencing of the SARS-CoV-2 genetic material in the samples, e.g., using ARTIC protocol (5), and (c) direct RT-qPCR detection of specific regions of the SARS-CoV-2 genome. Each of these approaches has its benefits and drawbacks (6), but the commonly adopted strategy currently is the targeted amplification approach. SARS-CoV-2 wastewater variant surveillance is complicated by potential partial degradation of the viral single stranded RNA, as well as the low relative abundance of the genetic material in the environmental samples. Additionally, the majority of the sequencing data and protocols are aimed to produce short amplicons (ARTIC v3 protocol produces amplicons that are 400bp long), and as the result short reads (i.e., even if the amplicons are sequenced on a long-read sequencer, the length of the read

cannot exceed the length of an amplicon). Thus, the task of complete haplotype phasing becomes nearly impossible in this setting, warranting development of other computational methods that can aid in the wastewater monitoring process.

Characterizing variants of concern by leveraging clinical data

Currently GISAID (7) is the largest collection of publicly available SARS-CoV-2 genome assemblies, which are predominantly obtained from the clinical data. Genomes deposited into the GISAID database are automatically assigned a PANGO lineage (8) via Pangolin software (9). There are several obstacles in translating lineage assignment to sequencing data obtained from wastewater samples.

First, Pangolin requires inputs to be assemblies rather than just sequencing reads. In the case of the clinical samples, one would often perform a reference guided assembly. The result of this process is a single genome, which in case of the wastewater data will be an inadequate representation of the genetic diversity within the sample and will not allow examination of the sample for potential presence of multiple VoCs/VoIs. An alternative approach would involve *de novo* assembly which can result in multiple contigs. However, due to high nucleotide similarity between SARS-CoV-2 variants and the need for simplifying heuristics in *de novo* assembly from short read data, these approaches will also fail to completely resolve full variant diversity of a sample. Thus, while we are able to identify individual mutations within a sample and their relative abundances, also referred to as allele frequencies (AF), we are not able to reliably reconstruct the genomic mixture that gives rise to a sample within reasonable computational time. Therefore, we are unable to directly use Pangolin software for lineage assignment.

Second, one can perform direct phylogenetic placement of sequencing reads onto the SARS-CoV-2 global phylogenetic tree. However, there are several limiting factors for this

approach. The global SARS-CoV-2 phylogeny contains more than 1.5 million SARS-CoV-2 sequences, implying that in order for wastewater monitoring to be computationally efficient some sub-sampling technique should be used. Such sub-sampling should ideally be minimally biased and should reflect the general features of the global phylogeny. This presents an issue, as we note that for example Nextstrain (10) sub-sampled version of global SARS-CoV-2 phylogeny contains clades conflicting with the PANGO lineage designation (Figure 1).

A third approach is to define a rule-based system by leveraging annotated data available from GISAID and extracting corresponding mutations from the multiple sequence alignment of the SARS-CoV-2 genomes. It is natural to ask whether each lineage can be characterized by a set of unique mutations that occur exclusively within that lineage and not in any other. We have evaluated the GISAID data (downloaded on May 6th, 2021) by extracting all nucleotide mutations using vdb (11) and analyzing mutational signatures corresponding to each present lineage. We found that for the VoCs/VoIs there are no mutations that would uniquely determine corresponding lineages among all observed ones.

To this extent we introduce the concept of quasi-unique mutations corresponding to a given PANGO lineage. This approach is motivated by using wastewater surveillance for SARS-CoV-2 VoCs/VoIs for early screening for *potential* emergence of the variants. We define a quasi-unique mutation for a lineage A as a mutation that is found in more than 50% of all available SARS-CoV-2 genomes that belong to the lineage A and found in less than 50% of genomes belonging to any other lineage B (additional details in SI). This combination of rules allows us to extract mutational signatures for each of the lineages from the clinical data, and therefore screen for potential presence/absence of the VoCs/VoIs in the environmental samples. We compare our screening process to using the manually curated list of characteristic mutations

for the VoCs/VoIs maintained by CDC (12). The latter list is what the majority of the current wastewater screening approaches rely on (3, 4). Finally, we integrate coverage information for the given quasi-unique positions to distinguish between the cases of "no detection" and "no coverage", as the latter can be indicative of the sample degradation or amplification failure rather than absence of the variant in the sample.

We applied this approach to track the emergence of the Delta (B.1.617.2) variant across 39 wastewater treatment plants in Houston, Texas (Figure 2). We observe that in the presence of a strong signal for VoC (weeks of 06/21 and 06/28) information from both quasi-unique mutations and characteristic mutations agrees. However, we also note that as we aim to track the early emergence, co-occurrence of certain characteristic mutations within other lineages can confound the picture, while quasi-unique mutations indicate a clear trend (Figure 2, bottom subplots, weeks 05/10 to 06/14).

Discussion

We observed that usage of quasi-unique mutations as markers for the potential emergence of VoC/VoI SARS-CoV-2 lineages is an effective strategy that can improve our ability to screen for variants in a community as they are emerging. The uptick in our quasi-unique signal matches the increase in Delta variant genomes being deposited into GenBank for the state of Texas (Supplemental Figure 2). Compared with the commonly taken approach of only using a curated list of characteristic mutations for the VoCs/VoIs, which shows a presence signal consistently throughout the sampling dates, our approach provides an alternative for detection of the emerging VoC/VoIs that matches the trend in cases supported by clinical data.

Quasi-unique mutational signatures can shift over time with the arrival of additional data into the GISAID database, a characteristic common to all genomic signature based approaches.

Therefore, this approach requires periodic updates to reflect the current state of lineage designation, as well as the corresponding quasi-unique mutational signatures. We note that the similar problem of periodic updates is common to all of the potential methods except ones that rely on assembly and external tools (e.g. Pangolin) for lineage assignment on the assembled genomes.

We also note that it is possible to forgo the manual threshold selection by implying a statistical approach in which we view this problem as a maximum likelihood estimation. However, due to the bias in the GISAID data, such an approach would require careful corrections for imbalanced classes (13).

Conclusion

We proposed a simple method for screening wastewater derived SARS-CoV-2 sequencing samples for emergence of VoC/VoI lineages. Our method relies on detection of quasi-unique mutations for target lineages and provides a more specific view than the lists of characteristic mutations for the corresponding lineages. While future improvements can be made to the method, we consider it important to apply these ideas in SARS-CoV-2 wastewater screening early on.

Data availability

Scripts used to process data have been uploaded to Box alongside a copy of the vdb database used in this manuscript. Files are accessible at the following URL: https://rice.box.com/v/VoC-Detection-in-WW.

Acknowledgements

The authors thank the GISAID contributors who provided the SARS-CoV-2 assemblies. We thank Lauren Bauhs, Madeline Wolken, Kyle Palmer, Whitney Rich, and Russell Carlson-

Stadler for their assistance in sample collection, processing, and analysis. We thank Ryker Penn and Lilian Mojica, from the Houston Health Department for their assistance in sample collection and sequencing. This work was supported by the Houston Health Department. E.L. and L.B.S. were supported in part by the National Science Foundation (CBET 2029025), and seed funds from Rice University. T.T. and N.S. were supported in part by C3.ai DTI and P01-AI152999 NIH awards. K.B.E. was supported in part by National Institute of Environmental Health Sciences, R01ES028819.

Author contributions: N.S. performed software and methodology development, formal analysis, metadata acquisition, and visualization of the results. E.L. performed experimental data acquisition, curation, and analysis. E.L, K.B.E., R.S, T.J.T, and L.B.S. provided critical feedback specific to the methodology development. T.J.T and L.B.S contributed to conceptualization and supervision of the study. L.H., L.B.S. provided supervision and resources for the project, and have administrated the study. N.S., L.B.S, and T.J.T. contributed to writing of the original draft. All authors contributed to reviewing and editing of the final manuscript.

References

1. Peccia J, Zulli A, Brackney DE, Grubaugh ND, Kaplan EH, Casanovas-Massana A, Ko AI, Malik AA, Wang D, Wang M, Warren JL, Weinberger DM, Arnold W, Omer SB. 2020. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nat Biotechnol 38:1164–1167.

2. Agrawal S, Orschler L, Lackner S. 2021. Long-term monitoring of SARS-CoV-2 RNA in wastewater of the Frankfurt metropolitan area in Southern Germany. Sci Rep 11:5372.

3. Agrawal S, Orschler L, Schubert S, Zachmann K, Heijnen L, Tavazzi S, Gawlik BM, Graaf M de, Medema G, Lackner S. 2021. A pan-European study of SARS-CoV-2 variants in wastewater under the EU Sewage Sentinel System. medRxiv 2021.06.11.21258756.

4. Bar-Or I, Weil M, Indenbaum V, Bucris E, Bar-Ilan D, Elul M, Levi N, Aguvaev I, Cohen Z, Shirazi R, Erster O, Sela-Brown A, Sofer D, Mor O, Mendelson E, Zuckerman NS. 2021. Detection of SARS-CoV-2 variants by genomic analysis of wastewater samples in Israel. Science of The Total Environment 789:148002.

5. Quick J. 2020. nCoV-2019 sequencing protocol v3 (LoCost).

6. Nasir JA, Kozak RA, Aftanas P, Raphenya AR, Smith KM, Maguire F, Maan H, Alruwaili M, Banerjee A, Mbareche H, Alcock BP, Knox NC, Mossman K, Wang B, Hiscox JA, McArthur AG, Mubareka S. 2020. A Comparison of Whole Genome Sequencing of SARS-CoV-2 Using Amplicon-Based Sequencing, Random Hexamers, and Bait Capture. 8. Viruses 12:895.

7. Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Global Challenges 1:33–46.

8. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, du Plessis L, Pybus OG. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. 11. Nat Microbiol 5:1403–1407.

9. 2021. cov-lineages/pangolin. Python, CoV-lineages.

10. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 34:4121–4123.

11. West AP, Wertheim JO, Wang JC, Vasylyeva TI, Havens JL, Chowdhury MA, Gonzalez E, Fang CE, Lonardo SSD, Hughes S, Rakeman JL, Lee HH, Barnes CO, Gnanapragasam PNP, Yang Z, Gaebler C, Caskey M, Nussenzweig MC, Keeffe JR, Bjorkman PJ. 2021. Detection and characterization of the SARS-CoV-2 lineage B.1.526 in New York. bioRxiv 2021.02.14.431043.

12. CDC. 2020. Coronavirus Disease 2019 (COVID-19). Centers for Disease Control and Prevention.

13. Oommen T, Baise LG, Vogel RM. 2011. Sampling Bias and Class Imbalance in Maximum-likelihood Logistic Regression. Math Geosci 43:99–120.

14. LaTurner ZW, Zong DM, Kalvapalle P, Gamas KR, Terwilliger A, Crosby T, Ali P, Avadhanula V, Santos HH, Weesner K, Hopkins L, Piedra PA, Maresso AW, Stadler LB. 2021. Evaluating recovery, cost, and throughput of different concentration methods for SARS-CoV-2 wastewater-based epidemiology. Water Research 197:117043.

15. BBMap. SourceForge.

16. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio].

17. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

18. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM, Brackney DE, Grewal S, Gurfield N, Van Rompay KKA, Isern S, Michael SF, Coffey LL, Loman NJ, Andersen KG. 2019. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biology 20:8.

19. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res 40:11189–11201.

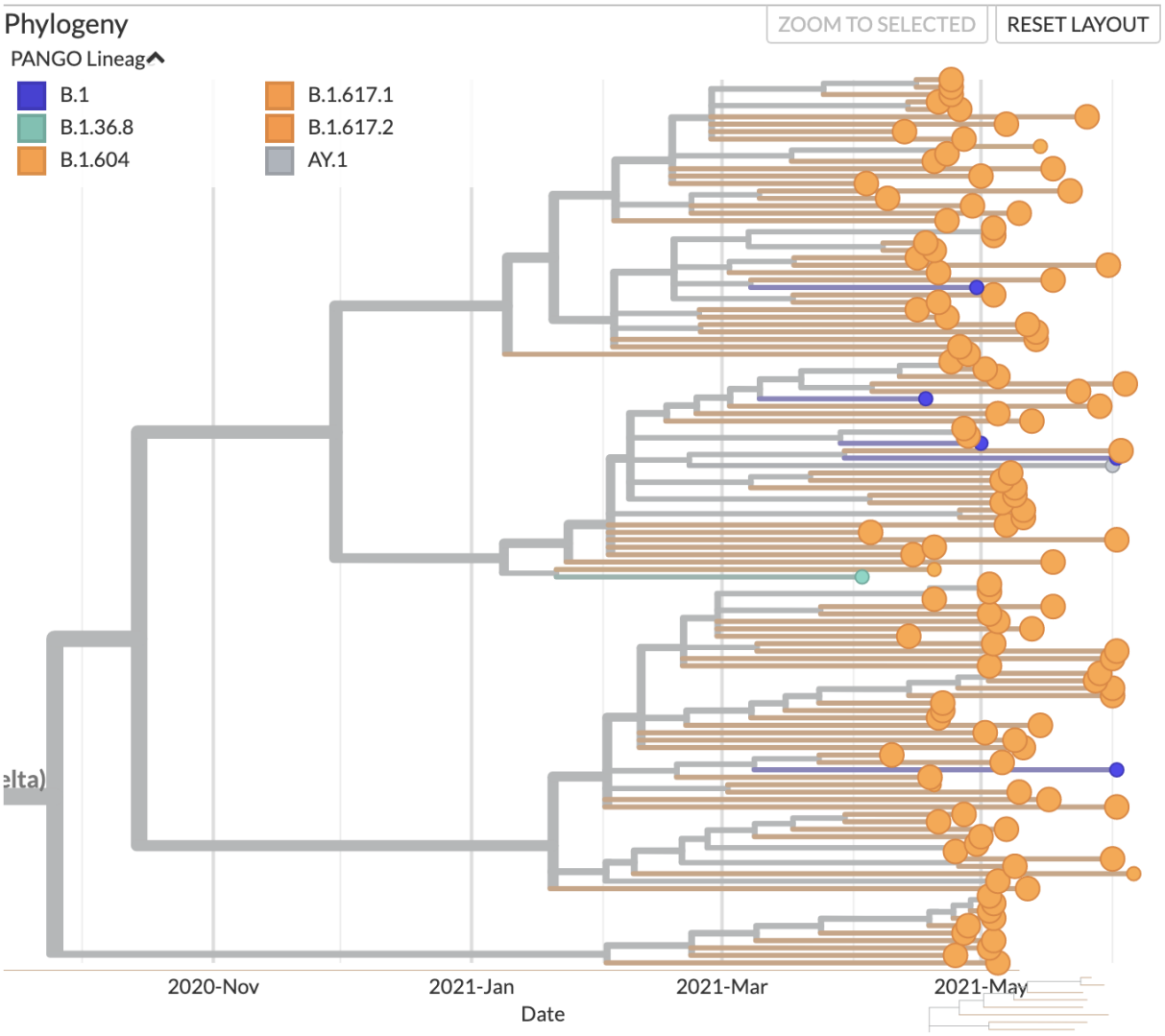


Figure 1. Nextstrain global phylogenetic tree highlighting PANGO lineages and the limitations of using phylogenetic tree clades to perform direct placement of sequencing reads into the lineage context by mapping to a phylogeny. Purple and cyan colors indicate PANGO lineages misplaced into the same clade as Delta (B.1.617.2).

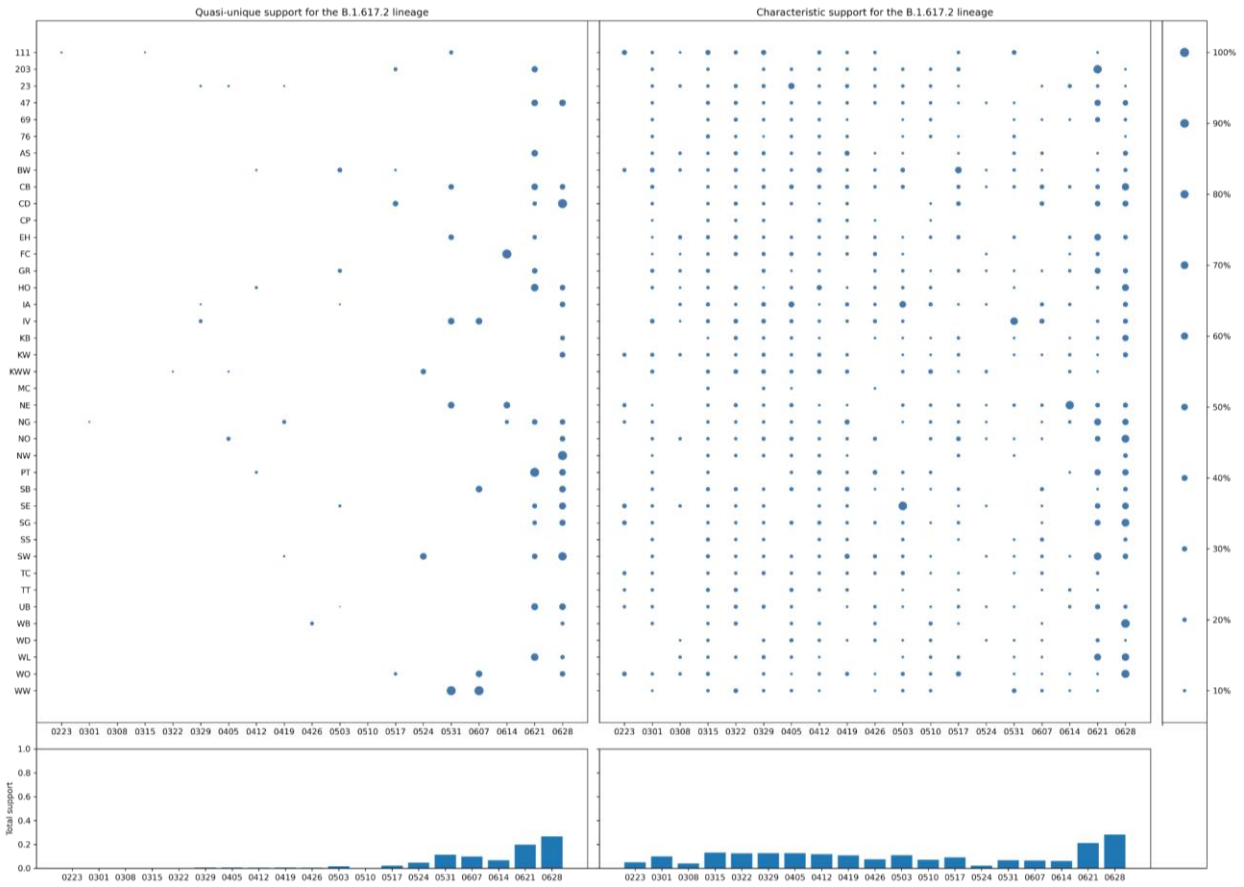


Figure 2. Quasi-unique (left) and CDC characteristic (right) detection plot for multiple sites over 19 weeks for Delta variant of interest (B.1.617.2). Each dot in the left subplots shows total allele frequency of observed quasi-unique mutations out of total possible number of quasi-unique mutations for the lineage (excluding positions which have zero coverage in the sample). In the right subplot, each dot shows the count of observed characteristic mutations out of total possible (without regard for allele frequency and coverage). The bottom plots provide an aggregate view for all sites. (Additional details in SI)

Supplementary material

Wastewater sample collection and SARS-CoV-2 concentration from wastewater

Between February 23rd and June 28th, 2021, weekly wastewater samples were collected from 39 wastewater treatment plants (WWTPs) across Houston covering a total service area of 580 square miles and serving 2.3 million people. Houston Water operators sampled from the influent channels at each WWTP and stored them in ice box or fridge for transportation and short-term storage. Samples were all collected from the influent channels of each WWTP and were 24-hour time-weighted composites. An electronegative filtration method (14) was applied for concentrating SARS-CoV-2 in wastewater. In brief, each wastewater sample (50 mL) was centrifuged for 10 minutes at 4,100 g and 4 ℃ to pellet solids. Supernatant was poured into a 6-head, Multi-Vac 610-MS Manifold (180310-01, Sterlitech), with each head containing an electronegative microbiological analysis HA filter (HAWG047S6, MilliporeSigma). Then, $MgCl_2 \cdot 6H_2O$ solution was added to each sample in the manifold to achieve a final concentration of 25 mM. Subsequently, samples were mixed gently with a pipette tip and left to sit for five minutes before applying a vacuum. A vacuum pump was used to pull the samples through the filters. Each filter was then folded and placed into a bead-beating tube preloaded with 0.1 mm glass beads. Tubes were stored at 4 ℃ prior to nucleic acid extraction, which occurred within 24 hours.

Nucleic acid extraction

RNA extraction was performed using a Chemagic 360 with integrated chemagic dispenser and the viral RNA/DNA purification protocol (PerkinElmer). 1000 μL of Lysis Buffer 1 was added to each bead beating tube containing the filters and glass beads. Bead beating was performed using a Mini-Beadbeater 24 (112011, BioSpec) at the max speed (3500

oscillations/min) for 1 minute. Then, tubes were allowed to rest on ice for 2 minutes, followed by a second round of bead-beating (max speed, 1 minute). Finally, all tubes were taken out from the bead beater and centrifuged for 2 minutes at 17,000×*g* and 4 ℃ to pellet the beads and filter paper. 300 μL supernatant (lysate) was withdrawn from each tube and used for the downstream automated nucleic acid extraction using Chemagic 360 along with the viral RNA/DNA purification protocol (PerkinElmer). Finally, 11 μL of sample extract was used for WGS library preparation.

SARS-CoV-2 whole genome sequencing using ARTIC v3 method

ARTIC v3 Illumina library construction (400 bps) and sequencing protocol for SARS-CoV-2 genome library preparation was performed (5). In brief, 11 μL RNA extract for each sample was processed with reverse transcription and cDNA preparation using the Superscript IV RT kit (ThermoFisher Scientific, 18090050) following the manufacturer's protocol. SARS-CoV-2 genome enrichment was conducted using ARTIC v3 protocol that consists of two separate pools for multiplexing PCR. Each 25 μL reaction contained 12.5 μL Q5 hot-start HiFi 2X mastermix (NEB, M0494L), 3.7 μL pre-mixed ARTIC v3 panels (Pool A or Pool B, 10 μM, IDT) and 2.5 μL synthesized cDNA product from the previous step. The PCR reaction program consisted of: initial denaturation at 98℃ (30 sec), followed by 35 cycles of denaturation at 98℃ (15 sec), annealing and extension at 63℃ (5 min), and 4℃ hold. After generating the reverse transcribed, tiled-PCR enriched DNA, Pool A and Pool B were combined for each sample and cleaned-up using AMPure XP beads (Beckman Coulter). Illumina DNA Prep kit containing the reagent buffer (Cat. No 20015828) and the beads (Cat. No 20015880) as well as the manufacturer's protocol (DNA Flex) were applied for sample tagmentation and flex amplification, followed by a final clean-up using SPB beads, 80% EtOH and RSB. After

purification, each sample library was pooled, denatured and diluted to 6 pM in the final HT1 solution. Finally, sequencing was performed on an Illumina MiSeq instrument using MiSeq Reagent Kit v2 (300-cycles, MS-102-2002) following the 151 + 10 + 10 + 151 recipe.

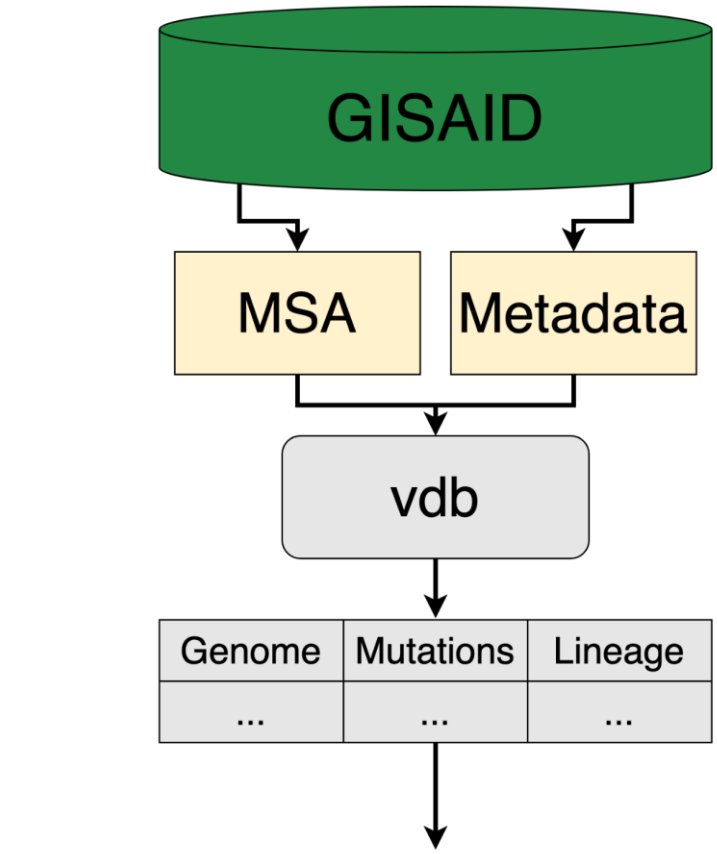
WGS data analysis and quasi-unique score calculation

First, in order to define the quasi-unique mutation sets we download multiple sequence alignment (MSA) and metadata files from GISAID website. We then process both using vdb v2.0 (13) in order to extract nucleotide changes (both SNPs and indels) and group them based on the lineage to which the corresponding genomes belong. In order to define quasi-unique mutations sets for each lineage we first form consensus mutation sets per lineage (i.e., all nucleotide changes that are present in more than 50% of the genomes in the lineage) and then subtract from these sets consensus mutation sets of all other lineages. The resulting mutation sets are defined as quasi-unique mutations for a specific lineage. A summary overview of this process is given in Supplemental Figure 1. We have attempted other values of the cut-offs ranging from 50% to 95% for the inclusion in the lineage A, and from 5% to 50% for the exclusion of the mutation occurring in any other lineage B. However, due to large imbalance in the number of genomes corresponding to different lineages in GISAID data (e.g., 462,747 high quality genomes correspond to the B.1.1.7 lineage, while only 8,188 correspond to P.1, and less than 1,000 to B.1.617.2) stricter threshold choices resulted in very small sets of quasi-unique mutations for some VoCs/VoIs.

In order to process WGS wastewater SARS-CoV-2 sequencing data we first trim the reads and remove adapter sequences using BBDuk (15) after which read mapping with BWA MEM (16) is performed. Resulted read mappings are then sorted using samtools (17) and primer locations are soft-clipped in the resulting mapping using iVar (18). Finally, variant calls are

performed with respect to the Wuhan reference (NC_045512.2) sequence for SARS-CoV-2 using LoFreq (19) and iVar variant callers. The resulting calls were then merged with only calls made by both softwares and having allele frequency of at least 2% in the sample were retained.

Finally, for each wastewater sample and each lineage of concern/interest the sum of allele frequencies of quasi-unique mutations has been computed. The results were reported both per wastewater treatment plant and an aggregate for the city.



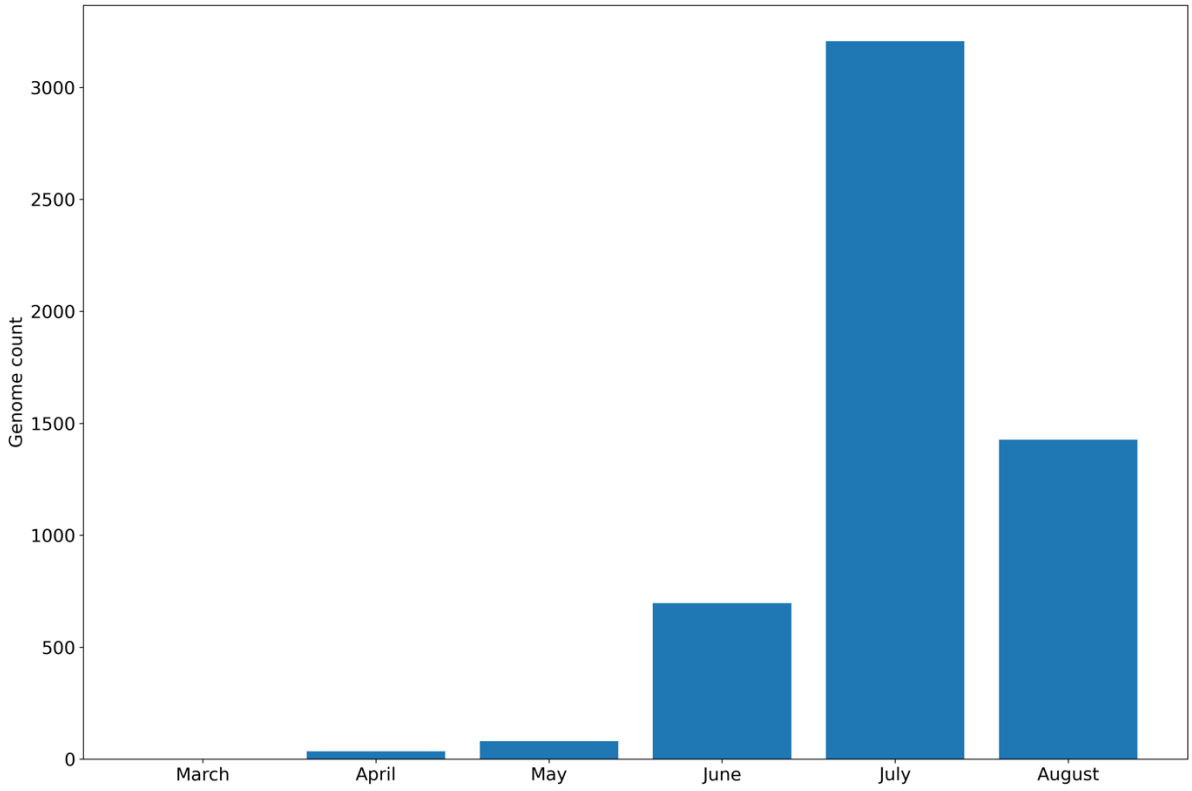

```
for lineage in PANGO_lineages:
        lineage_mutations := mutations ∈ >50% of
all genomes in lineage
        non_lineage_mutations := mutations ∈ >50%
of all genomes for any other_lineage ∈
PANGO_lineages \ {lineage}

        quasi_unique_mutations :=
lineage_mutations - non_lineage_mutations
```

Supplemental Figure 1. Overview of the GISAID data processing performed in order to obtain quasi-unique mutation sets for lineages of interest.

Clinical sequencing data for B.1.617.2 lineage in Texas

We downloaded metadata from NCBI Virus Portal for Texas and filtered it to the genomes that were assigned the B.1.617.2 PANGO lineage. We sorted the data by sample collection date and plotted the resulting counts in the Supplemental Figure 2. The data for August is incomplete as some of the sequencing is still being performed and results analyzed. This highlights one of the issues with using clinical data as a monitoring tool, since the upload date for many clinical samples is up to several months later than the actual date of sequencing.



Supplemental Figure 2. Count of genomes deposited into GenBank from Texas and assigned B.1.617.2 PANGO lineage grouped by date of sample collection. GenBank upload date can be up to several months later than the sample collection date.