Crykey: Rapid Identification of SARS-CoV-2 Cryptic Mutations in Wastewater

Yunxi Liu[1], Nicolae Sapoval[1], Pilar Gallego-García[2,3], Laura Tomás[2,3], David Posada[2,3,4], Todd J. Treangen[1*], Lauren B. Stadler[5*]

[1]Department of Computer Science, Rice University, Houston, TX, 77005, USA
[2]CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[3]Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO.
[4]Department of Biochemistry, Genetics, and Immunology, Universidade de Vigo, 36310 Vigo, Spain
[5]Department of Civil and Environmental Engineering, Rice University, Houston, TX, 77005, USA
*Corresponding authors: lauren.stadler@rice.edu, treangen@rice.edu

Abstract

We present Crykey, a computational tool for rapidly identifying cryptic mutations of SARS-CoV-2. Specifically, we identify co-occurring single nucleotide mutations on the same sequencing read, called linked-read mutations, that are rare or entirely missing in existing databases, and have the potential to represent novel cryptic lineages found in wastewater. While previous approaches exist for identifying cryptic linked-read mutations from specific regions of the SARS-CoV-2 genome, there is a need for computational tools capable of efficiently tracking cryptic mutations across the entire genome and for tens of thousands of samples and with increased scrutiny, given their potential to represent either artifacts or hidden SARS-CoV-2 lineages. Crykey fills this gap by identifying rare linked-read mutations that pass stringent computational filters to limit the potential for artifacts. We evaluate the utility of Crykey on >3,000 wastewater and >22,000 clinical samples; our findings are three-fold: i) we identify hundreds of cryptic mutations that cover the entire SARS-CoV-2 genome, ii) we track the presence of these cryptic mutations across multiple wastewater treatment plants and over a three years of sampling in Houston, and iii) we find a handful of cryptic mutations in wastewater mirror cryptic mutations in clinical samples and investigate their potential to represent real cryptic lineages. In summary, Crykey enables large-scale detection of cryptic mutations representing potential cryptic lineages in wastewater.

Introduction

Wastewater monitoring is a vital tool complementing clinical testing for COVID-19 surveillance[1–9], and can fill in the surveillance gap when clinical testing is unavailable or halted. Multiple studies have demonstrated that SARS-CoV-2 variants of concern (VOCs) can be detected in wastewater samples[10–15], preceding clinical testing by up to two weeks[9]. Furthermore, wastewater samples contain information on the genomic diversity of the circulating variants in the entire community, avoiding the sampling bias inherent to the clinical surveillance, which focuses on symptomatic patients[16–18]. Importantly, wastewater monitoring can also be used to detect novel and rare SARS-CoV-2 lineages not represented in GISAID's EpiCoV database[19], termed cryptic lineages[20]. A few methods have been proposed for the detection of cryptic lineages from wastewater samples, but they often require a combination of ultra-deep sequencing of specific genomic regions as well as a mixture of short-read, long-read, and proximity ligation sequencing technologies and thus are not compatible with most wastewater sequencing protocols used for routine monitoring due to time and cost limitations[16,21]. Moreover, non-uniform sequencing coverage caused by amplicon efficiency heterogeneity and environmental RNA degradation creates a challenge for detecting cryptic lineages from wastewater samples[5,22,23]. Furthermore, the origin of these cryptic lineages in wastewater is still an open question[24]. It has been proposed that they could be rare intra-host lineages that are not represented in the consensus genomes available in public databases, rare lineages with low prevalence in the population, lineages from non-human hosts (like rats), or technical artifacts[20,21,25].

In this manuscript, we introduce Crykey, a novel computational method for detecting rare linked-read mutations from wastewater samples that exploits the co-occurrence of point mutations on the same sequencing read or read-pair (from now on, linked-read or LR mutations). The rationale is that LR mutations found in wastewater samples but nonexistent or at a very low prevalence (e.g. < 0.0001) in public databases represent potential cryptic lineages (from now potential cryptic lineage will be denoted as CR); i.e. rare linked-read mutations supported by 5 or more reads which we claim are indicative of one of the following: real cryptic lineages, SARS-CoV-2 transcription variation due to subgenomic mRNAs[26], or systematic artifacts. We used Crykey to analyze 3,175 wastewater samples collected in Houston, Texas, USA, from February 2021 to November 2022. Our results are threefold: (i) We discover numerous cryptic mutations spanning the whole SARS-CoV-2 genome, (ii) we monitor the occurrence of these cryptic mutations across numerous wastewater treatment plants (WWTPs), observing them over a period of three years in Houston, and (iii) we identify a cryptic mutations in wastewater samples that reflect those in clinical samples, and explore the possibility of these representing actual hidden lineages.

Results

To evaluate the utility and efficiency of Crykey, we applied it to SARS-CoV-2 amplicon sequencing data from 3,175 wastewater samples collected from 39 wastewater treatment plants (WWTPs) in Houston between February 2021 and November 2022, as well as in 5,060 short-read clinical samples collected within the Greater Houston between December 2021 and January 2022 (Supplementary Figure 1), and nearly 9,000 short-read clinical samples collected outside of Texas over the same 8-week time period (between 2021-12-06 and 2022-01-31; Supplementary Figure 2). In addition, we examined over 7,000 long-read clinical samples on a specific CR. We will now delve into the specific results from these data.

Overview of Crykey and Computational Performance

Crykey is a computational tool designed to search for cryptic mutations from samples by performing fast variant queries to determine how rare a set of LR mutations is among millions of publicly available genomes. Genomes of the same lineage during a short period of time have more mutations in common. Therefore, we indexed the Crykey database by partitioning the genomes into bins by sample collection date and lineage (Figure 1a). By pre-computing the prevalence rate of each mutation in each bin, Crykey is able to quickly reduce the search space from the entire database to a few hundreds of genomes for exact matching (Figure 1b, Figure 1c, and Figure 1d). We identified a total of 6,744 CR candidates in wastewater samples from Houston. More than 67.8% of the candidate CRs were found in 50 or less sequences in the GISAID EpiCoV database. Fully novel CRs (zero prevalence, or meaning it has not been previously observed) constitute more than 32.8% of the data (Figure 2a). We benchmarked the processing time of exact searches on a Linux machine with Intel Xeon Gold 6138 2.00GHz CPU. The process time increases as the rarity of the CR candidates decreases, as the result of the expansion of the search space (Figure 2b).

Genomic distribution of CRs in wastewater samples

After quality control, we identified 705 CRs in the wastewater samples. Figure 3a shows the location of the CRs along the reference SARS-CoV-2 genome, their mean allele frequency (AF) across wastewater samples, prevalence in GISAID, and the number of weeks (not necessarily consecutive) detected in wastewater. 74.8% of the CRs had a mean AF less than 0.2, while 7.8% of the CRs were detected at consensus level AF (mean AF >= 0.5) (Figure 3a). The occurrence of the CRs varied significantly, ranging from 1 to 33 weeks (size of the dot in Figure 3a). Almost half of them were located in the S (20.1%) and N (28.4%) genes. Most of the genome regions are dominated by CRs that contain only non-synonymous mutations, except for gene N (Figure 3b).

Emergence of CRs co-occurs with the spread of new variants

The emergence of the CRs coincided with the spread of new VOCs. For example, the number of CRs and viral load in wastewater increased significantly around July 2021 (Figure 4a), corresponding to the Delta wave in Houston (Figure 4b). Similar patterns were observed during the emergence of B.1.1.529 (Omicron) in December 2021, BA.2 (Omicron) in May 2022, and BA.5 (Omicron) in July 2022. Most CRs could be associated with one (77%) or more (17%) known PANGO lineages circulating at the time (Supplementary section 1). We observed fewer CRs between April and August 2022, when the sequencing breadth and depth of coverage dropped due to primer dropouts in lower quality sequencing runs (thinner bars in Figure 4a). It is hard to untangle whether this effect was due to specific genomic features of the BA.2 and BA.5 variants or whether it was a consequence of the lower quality of the sequencing data during this period.

Houston CRs display distinct patterns

More than 400 CRs found in Houston wastewater were completely novel, or can only be found in less than 50 genomes in GISAID (Figure 5). The vast majority of the CRs did not persist for long, with over 85% found in less than four consecutive weeks. Short-duration CRs (less than 4 weeks) were generally found in only a few WWTPs and at low AFs (Figure 5). Interestingly, some CRs were detected in multiple WWTP across the city and persisted for 4 months or longer. The most persistent CR observed, which we named CR12, was detected in the wastewater for 33 weeks. CR12 contains mutations A29039T and G29049A, which cause K256* (stop codon) and R259Q amino acid changes, respectively, on the N gene. The mean AF of CR12 across WWTPs was generally low (<0.1), with a few exceptions. CR12s presence ramped up slowly in 1-7 WWTPs, peaked in late November 2021 when observed in 16 WWTPs (Figure 6), and phased out in late February 2022 and

remained undetected for two months (but notice that the sequencing coverage also dropped during this period), re-appearing for a short time around May 2022.

Many CRs exhibited perplexing patterns of allele frequency, duration, and clinical sample prevalence. The first occurrence of CR12 in Houston wastewater can be traced back to Aug 2, 2021, when Delta was dominant in the community. As Omicron emerged in December 2021, CR12 continued to be present in the wastewater and clinical samples (Figure 4 and Figure 6). To evaluate the possibility of CR12 being a technical artifact, we first explored whether this could be due to a read mapping error by using a different read mapper, Bowtie2[27]. The results were consistent with those obtained previously with BWA MEM[28]. To investigate whether the mutations comprising CR12 were due to systematic sequencing errors, for example primer dependent errors, we further examined 7,113 clinical samples sequenced with the PacBio HiFi system (Sequel II), including 2,458 samples collected from Texas, and 4,655 samples collected from other US regions between Nov 06, 2021 and Mar 21, 2022. Forty-five of those samples included reads that supported the presence of CR12, and 28 of them from Texas (Supplementary Table 1). While observing the rare LR mutations comprising CR12 in PacBio data does not fully rule out the possibility of technical artifact, it provides a basis for ruling out platform specific errors.

Investigating CRs in clinical samples

As Omicron became dominant in the community, several CRs specific to the VOC emerged and became more prevalent among hosts (Figure 7). We explored the occurrence of 20 CRs with short-term or long-lasting patterns in 5,060 clinical samples collected within the Greater Houston (between 2021-12-06 and 2022-01-31; Supplementary Figure 2). 12 out of 20 CRs detected in the wastewater were seen in the clinical samples (Figure 7), including CR3, CR5, CR8, and CR12 (Figure 8, Supplementary Figure 3-5). Remarkably, for these CRs, the mean AF within the clinical samples was very low (< 0.05), except for CR5 (Supplementary Figure 4). CR2 was associated with Delta, while the remaining eleven were mainly associated with Omicron. Likewise, the consensus genomes for most of the Houston clinical samples carrying CRs (all but CR2) were identified as Omicron, mainly BA.1.1 and BA.1.15 (Figure 7). The clinical prevalence of the Omicron CRs increased as the Omicron variant spread in the city, as reflected by both the viral load in the wastewater (Figure 4a) and the number of sequences from Texas in GISAID (Figure 4b). In contrast, CR2 was detected in the wastewater only during the first two weeks of the sampling period, while also detected with very low prevalence in the clinical samples during weeks 1-6 (Figure7). We also queried the number of sequences with each amino acid change associated with CR1-CR12 all over the world using outbreak.info[29,30]. As expected, the consensus level mutations are often found in millions of SARS-CoV-2 sequences, and the mutations with low AF are found in much lower number of sequences, ranging from a handful to thousands (Supplementary Table 2). The compendium of evidence for CR1-CR12 provides a mixed picture of factors driving the rare LR mutations comprising these CRs. Low allele frequency, high prevalence, and geographic discordance casts doubt on these CRs representing legitimate cryptic lineages.

However, CR8 exhibited a different pattern. We first detected CR8 in clinical samples in the 1st week at a low prevalence rate. As the prevalence rate in clinical samples increases, we were able to detect CR8 in wastewater on the 3rd week (Figure 8), from samples sequenced using distinct protocols. CR8 consists of two mutations, C10449A and T10459C. C10449A is a consensus level mutation for Omicron strains, and it has an individual mean AF close to 1 in both wastewater and clinical samples (Figure 8a and Figure 8e). The prevalence rate of C10449A alone gradually increases in the first 3 weeks of detections starting from the week of 2021-12-06, until the prevalence rate reaches 1, and the pattern is consistent in both wastewater and clinical

samples; on the other hand, mutation T10459C is present as a low frequency mutation with individual mean AF close to 0.02. The prevalence rate of T10459C alone in clinical samples increased at the first half of the sampling period, reaching peak at week 4, and then decreased in the second half of the sampling period (Figure 8a and Figure 8e). Since CR8 contains both a consensus level mutation C10449A and a low frequency mutation T10459C, both mean AF and prevalence rate of the co-occurring mutations follows the pattern of the T10459C (Figure 8b and Figure 8f), and as the prevalence rate of CR8 in clinical samples increases, we start to detect it in wastewater on week 3 as well. The average number of reads that span CR8 regions are shown in Figure 8c and Figure 8g as coverage and Crykey is sensitive enough to detect CR8 in wastewater given that the coverage of wastewater samples is much lower than clinical samples (Figure 8d).

To assess whether there were geographic patterns at a national level associated with these CRs, we processed nearly 9,000 clinical samples collected outside of Texas over the same 8-week time period (between 2021-12-06 and 2022-01-31; Supplementary Figure 2). CR5 was detected across clinical samples from Maryland (very high prevalence) and Massachusetts (low prevalence) (Supplementary Figure 6a). CR8 was detected in Maryland again at a very high prevalence (Supplementary Figure 6b). In addition, we identified five additional CRs shared across clinical samples from Houston and Maryland (CR3, CR4, CR7, CR9, CR11). Note that the distribution of the PANGO assignments for samples containing CR5 and CR8 differed between states. Although both CRs are associated with Omicron, Houston was dominated by BA.1.15, while Maryland and Massachusetts had a much higher proportion of BA.1.1 and BA.1.17, and Maryland had much higher proportion of BA.1.18 and BA.1.20 as well.

Discussion

Wastewater monitoring for SARS-CoV-2 has been widely used for complementing clinical genomic surveillance during the COVID-19 pandemic[14,31] . A recent study claims to have identified cryptic SARS-CoV-2 lineages in the wastewater that went undetected in the clinic, leaving it an open question as to the origin of these CRs[20]. Our contribution centers on a novel detection tool, Crykey, designed to identify rare linked-read mutations in wastewater using sequencing data. Specifically, Crykey leverages an optimized database lookup for the co-occurrence of mutations that are present on the same reads or read pairs and to detect the presence of CRs. Our method is fully compatible with standard mutation calling pipelines for SARS-CoV-2, and considers CRs defined by mutations that may occur across the entire SARS-CoV-2 genome. To demonstrate the efficacy of our new computational tool, we applied CryKey to >3,000 wastewater samples from Houston.

By examining three years of wastewater sequencing data and eight weeks of local clinical surveillance data, our goal was to demonstrate the potential of Crykey to provide a finer-grain view of the emergence of potential cryptic lineages within Houston. We also provide results that show twelve CRs were found in wastewater and clinical samples from the same time period. Future work is required to validate CRs as they emerge and to discern between potential systematic biases and legitimate CRs. In particular, cases where the CR is highly prevalent in clinical samples but at a low frequency within each individual and cases when all the mutations comprising a CR exhibit similar allele frequencies (which could represent read mapping or alignment error caused by indels, or strong evidence for a novel cryptic lineage given the very low likelihood of multiple errors being introduced on a single read from a high-fidelity sequencing platform), and combinations of consensus level established mutations observed with single low-frequency mutations (which could represent the emergence of a novel SARS-CoV-2 lineage or also represent an error co-located with a characteristic mutation from a PANGO lineage) warrant further investigation.

Interrogating outlier CRs with Crykey

While 85% of the CRs lasted for less than 4 weeks, we also observed some CRs that persisted for more than 10 weeks (Figure 5). Notably, CR12 was detected across multiple WWTPs in Houston for 33 weeks (Figure 6). CR12 contains two LR mutations, A29039T and G29049A, which cause K256* and R259Q amino acid changes on the N gene. The combination of these mutations is rare; there are only three entries in GISAID that contain both mutations, and none of these originated from the United States. Previous work has shown that N:K256 is one of the eight lysine residues in the protein N of SARS-CoV-2 that is likely to be directly involved in RNA binding[32]. A29039T generates a stop codon that may affect the linker region, suppressing the immunogenic domain of the nucleocapsid protein which might help vaccine escape[32,33]. N:R259 belongs to one of the identified guanosine triphosphate binding pockets, and is well-conserved in multiple human coronaviruses, including NL63, 229E, HKU1, OC43, as well as MERS and SARS-CoV-1[34], the N:R259Q mutation has been reported multiple times at low prevalence rates in several SARS-CoV-2 lineages, likely representing a hotspot mutation mostly belonging to the Delta variant[35]. A previous study suggested that the nucleocapsid protein of SARS-CoV-2 is flexible and dynamic, and CR12 happens to be located on one of the predicted disordered regions of the N gene[36].

Potential origins of cryptic mutations: do they represent cryptic lineages, poorly understood biology signals, or systematic noise?

The precise origin of the cryptic mutations we found in Houston wastewater remains an open question. One could think of five possible scenarios: 1) they represent rare circulating SARS-CoV-2 lineages that went un-sampled or under-sampled in the clinical samples, 2) they exist as intra-host mutations from the population that have high enough prevalence to be detected in wastewater, 3) they represent signal from SARS-CoV-2 transcription such as subgenomic mRNAs, 4) they are spillover from an unidentified animal reservoir, or 5) they are technical artifacts from environmental degradation, sample preparation or sequencing.

A first possible explanation behind CRs in the wastewater not being captured by clinical surveillance is low community prevalence rates[20,25,21]. As only a small portion of the SARS-CoV-2 infections are sequenced, transient cryptic lineages are likely to be missed by clinical surveillance. Clinical data also suffers from sampling bias, where people with severe symptoms and access to healthcare resources are more likely to be represented in the databases. Figure 5 shows that most of the cryptic mutations detected in Houston wastewater were only found over 1 to 3 weeks, and these short-duration cryptic mutations may represent those not captured by clinical testing. In support of this hypothesis, we found that a subset of the cryptic mutations that were also supported by reads from clinical sequencing.

However, we observed several cryptic mutations with a high prevalence and low intra-host AF in the clinical samples. Indeed, it is common to only report consensus-level mutations (i.e., mutations with allele frequencies greater or equal to 0.5), or consensus genomes/assemblies to the public databases such as GISAID. As a result, although CRs might be sampled, they will remain unreported. A recent study has shown that molnupiravir treatment can induce de novo mutations in multiple individuals, but it is not clear whether the cryptic mutations found in clinical samples are tied to therapy-related lineages[37]. We also observed cases where a CR persisted for multiple weeks in wastewater samples but had little to no trace in clinical samples. Why these CRs were not captured by clinical surveillance remains unknown. As a possible explanation for this second scenario, previous studies have suggested that cryptic lineages may be carried by non-human hosts[20],

especially for those that persist for very long periods[38]. Given we lack representative genomes from non-human hosts during the time frame of our results, we are unable to investigate the plausibility of this scenario.

Temporally linked CRs (especially those that appear and disappear within a few weeks) provides compatible evidence for legitimate novel cryptic lineages. Additionally, CRs that contain multiple low-frequency mutations in a single read (and all supported by 5 or more reads) contrast themselves with CRs that contain a mutation shared with a circulating PANGO lineage in addition to a companion low-frequency mutation (or mutations). On the surface, low mean allele frequencies combined with high prevalence rates in clinical samples raises some concern regarding their validity, especially given the lack of plausible explanation for transmission/community spread of low frequency intrahost mutations (Supplementary Figure 5).

To investigate the potential origin of CRs, we evaluated a dataset of 5,060 clinical samples collected within Greater Houston from 2021-12-06 to 2022-01-31. Our results suggest that CRs detected in wastewater could be related to intra-host low frequency co-occurring mutations in clinical samples (Figures 7 and 8). The unusual long life span of CR12, together with its high prevalence at low intra-host AF in clinical samples, suggests that it could be a previously unreported artifact. However, examining over 7,000 clinical samples sequenced with the PacBio HiFi system, and tested multiple read aligners, we can likely rule out that CR12 is primarily related to primer artifacts, sequencing technology-dependent artifacts, and sequence alignment errors.

Finally, cryptic mutations could represent some type of systematic bias. Even though we have taken extreme care to filter out known sources of artifacts and have observed them across different sample types, amplicon panels, read mappers, and sequencers, we cannot rule out unknown systematic artifacts or biases leading to CR detection. Given there is no single mutational pattern observed (they can be comprised by multiple low allele frequency mutations that are short duration or very long duration, as well as contain mutations that pair an established consensus level mutation from a VOC with a transient low-frequency mutation), explanations for each of these patterns and their variability over time requires further investigation. Indeed, the goal in developing Crykey was to provide an efficient and sensitive tool for interrogating cryptic mutation patterns over time and geography, with the hope of shedding light on their origin and facilitating the identification of artifacts.

Open challenges in tracking potential cryptic lineages mutations in wastewater

One of the key challenges in reliably detecting CRs in wastewater is the quality of the samples[39,40]. As shown in Figure 4, the number of newly emerging CRs follows the same pattern as the viral load until June 2022, where the samples collected afterward had worse quality regarding breadth of coverage. The performance of Crykey is limited because the samples did not have enough sequencing depth across most of the regions of SARS-CoV-2 genome during those weeks. Due to the inherent limitations of short-read sequencing platforms that generate 100-200 bp reads, protocols used for sequencing, and the fragmented state of the viral RNA in wastewater, there is a natural limit on the genomic span of the cryptic mutations we can use. Indeed, degradation of genetic material in wastewater impacts the sequencing quality of the sample, at the same time introducing noise for rare mutation detection[41,42]. Furthermore, genomic regions corresponding to sequencing primers or adapters create coverage gaps (regions without read support) along the genome and pose a challenge for identifying CRs that span longer regions. However, these limitations could be addressed using

long-read sequencing if sample manipulation and extraction procedures allow intact longer RNA fragments to be recoverable from wastewater samples.

Concluding remarks

Crykey represents an efficient and easy-to-use tool specifically designed to rapidly and comprehensively find cryptic mutations across thousands of wastewater samples. We applied Crykey to detect numerous CRs in Houston, some persisting for months. The concept of searching for rare LR mutations inside of a viral genome that have never or rarely been reported is generalizable, and Crykey is not limited to SARS-CoV-2. Crykey can be expanded and applied to multiple pathogens, such as influenza A virus, as long as the pathogen has an established database of genomic sequences[43,44]. We are hopeful that our findings will help promote community-wide discussion on best practices for cryptic lineage tracking in wastewater.

Methods

Crykey is a computational method for the identification of cryptic mutations (linked read mutations supported by 5 or more reads and occurring less than 0.01% of the total GISAID EpiCoV samples, which represents 1298 genomes or less and as few as zero genomes) representing potential cryptic lineages (CRs) in wastewater samples on a full-genome scale. The workflow of the Crykey pipeline can be divided into 3 steps, including database construction, sample processing to find CR candidates, and rarity calculations for each candidate found in the previous step. Crykey first builds mutation look-up tables and a genome-to-mutation database using the full GISAID's EpiCoV database (Figure 1a) and then searches for CRs (Figure 1b). Specifically, Crykey first extracts LR mutations from the alignment and searches for CR candidates by querying the mutation look-up table (Figure 1c). Then, Each CR candidate is queried against a pre-built database to check if it is novel or rare in terms of prevalence to create candidate CRs (Figure 1d). Candidate CRs are then passed through a rigorous set of filters to nominate a subset as detected CRs. Due to the optimized database structure that partitions the mutation prevalence information according to the associated PANGO lineage/lineages for a given time period, Crykey is highly efficient and can easily scale to thousands of samples. We will now provide specific details regarding the filtering steps and analysis methods used in this manuscript.

Candidate CR Lineage Determination

The database used in Crykey is built based on the multiple sequence alignment (MSA) generated by the GISAID EpiCoV database. We extracted the mutations for each SARS-CoV-2 genome in the MSA using vdb with the command *vdbCreate -N input.msa*[45]. We then trimmed the list of mutations associated with each genome sequence with the vdb *trim* command. Combining the lineage assignment of each genome sequence in the metadata, we calculate the prevalence rate of each mutation in each of the known lineages of SARS-CoV-2, as well as building a mutation database containing mutation information for each individual SARS-CoV-2 genome. The results shown in this manuscript are based on the GISAID EpiCoV database dated at 2023-01-10.

To identify candidate cryptic lineages, Crykey first builds a default mutation lookup table where each mutation in GISAID is associated with a set of lineages and specific weeks (based on sample collection date) of occurrence in GISAID, regardless of prevalence rate. A second mutation lookup table is built at the same time where only mutations with a prevalence rate greater than 0.5 are stored, which allows us to perform a fast query on whether a set of SNPs belongs to any of the SARS-CoV-2 genome in a given time period.

quality reads[49]. Then the filtered reads were aligned to the reference genome of SARS-CoV-2 (NCBI Reference Sequence: NC_045512.2) with bwa mem v0.7.17-r1188 with default parameters[28,49]. The alignment files were sorted and indexed with samtools v1.14[50]. Mutation calling was done using lofreq v2.1.5 with command “lofreq call --no-default-filter --call-indels”, and then filtered with command “lofreq filter --cov-min 20 --af-min 0.02 -b fdr -c 0.001”[51]. Consensus genomes were generated and PANGO calling was done using pangolin v4.2 with default parameters[52].

After read mapping, the BAM files and the VCF files are collected for searching for CRs detected in wastewater samples. 20 wastewater CRs detected during the 8 week sampling period were selected for testing. 10 of the 20 wastewater CRs occurred 2 of the 8 weeks, representing CRs with a short burst pattern; when cross-referencing with the clinical sample data, we selected short lived CRs that were detected in the most wastewater treatment plants. The rest of the 10 wastewater CRs we selected for the query are those that had the longest duration of the clinical sampling period, ranging from 4 to 8 weeks of occurrence, representing CRs with a long lasting detection pattern in both wastewater and clinical samples.

By using the alignments in the clinical samples, we counted the total number of reads spanning the regions that the CRs contained, and counted the number of reads supporting all mutations from the CRs at the same time. 5 bases towards both ends of the reads were ignored to avoid noise caused by sequencing errors. The allele frequency of a potential cryptic lineage was calculated as the number of CR supporting reads over the number of total reads covering those positions.

During the analysis, we further filtered the results, and samples with CRs with less than 5 supported reads or with AF less than 0.02, or any of the mutations within the CR missing from the variant calling are considered as CR absent. We counted forward and reverse read fragments that do and do not fully support all cryptic mutations, and calculated both the p-value of the Fisher’s exact test and strand bias scores described in the previous studies[53,54]. Samples with reads containing strain bias scores greater or equal to 1 and p-value of the Fisher’s exact test less than 0.05 are also considered as CR absent.

All wastewater samples and clinical samples with insufficient coverage for a CR (number of reads that cover all mutation positions of a CR is less than 10) are excluded from the calculation of AF, prevalence rate, and coverage in Figure 7 and Supplementary Figure 3-5. The AFs of individual mutations from the CR are extracted from the variant calling results. The AF of CR is calculated as the number of reads that contain all mutations of a CR over the number of reads that cover all mutation positions of a CR. The prevalence rate is calculated as the count of CR/mutation detected samples over the count of samples with sufficient coverage.

Validating CR12 (A29039T-G29049A) CR with PacBio clinical samples

The samples were downloaded from the NCBI SRA database under the BioProject PRJNA716984. We subsampled 7,113 SRA runs with sample collection dates between 2021-11-06 and 2022-03-21, including 2,458 samples collected in Texas and 4,655 samples from 50 other regions (49 US states and Puerto Rico). Samples with missing metadata (location or sample collection date) were excluded. The reads were aligned to the reference genome of SARS-CoV-2 (NCBI Reference Sequence: NC_045512.2) with minimap2 using map-pb preset[55]. The alignment files were sorted and indexed with samtools v1.14. The number of supporting reads and the depth of coverage is calculated using the same method described in the previous section.

Data availability

Source data is provided with this paper and has been deposited in the OSF database with DOI: 10.17605/OSF.IO/3SPRZ. All sequencing data supporting the findings of this study is publicly available. Houston wastewater datasets are available for download via NCBI BioProject PRJNA796340. Houston short read clinical datasets are available for download via NCBI BioProject PRJNA764181. Non-Texas short read clinical datasets are available for download via NCBI BioProject PRJNA686984. The PacBio SARS-CoV-2 clinical datasets are available for download via NCBI BioProject PRJNA718231.

Code availability

The source code for Crykey is publicly available at https://github.com/treangenlab/crykey. The code used for analysis and figure generation used in this study can be found in https://github.com/treangenlab/crykey_analysis_scripts.

Acknowledgements

The authors thank all of the GISAID contributors who provided the SARS-CoV-2 genomic data. We also thank Dr. Loren Hopkins, Dr. Kathy Ensor, Kaavya Domakonda, Rebecca Schneider, and Anthony Mulenga for their leadership and contributions to the Houston Wastewater Epidemiology system. We also thank Dr. Adolfo Lara, Roger Sealy, Pamela Brown, Ryker Penn, and Yanlai Lai (Houston Health Department), as well as Dr. Esther Lou, Lauren Bauhs, Robert Campos, Russell Carlson-Stadler, Madeline Wolken, Kyle Palmer, Whitney Rich (Rice University). This work was supported in part by the Houston Health Department. Y.L., N.S., and T.J.T. were supported in part by the C3.ai DTI, Centers for Disease Control (CDC) contract 75D30121C11180, and P01-AI152999 NIH award. T.J.T. was also supported by National Science Foundation (NSF) grants EF-2126387, IIS-2239114, and CNS-1338099. N.S. was also supported by the Ken Kennedy Institute Andrew Ladd Memorial Excellence in Computer Science Fellowship. L.B.S. was supported in part by the National Science Foundation (CBET 2029025), seed funds from Rice University, the City of Houston, and CDC contract 75D30122C14709.

Figures

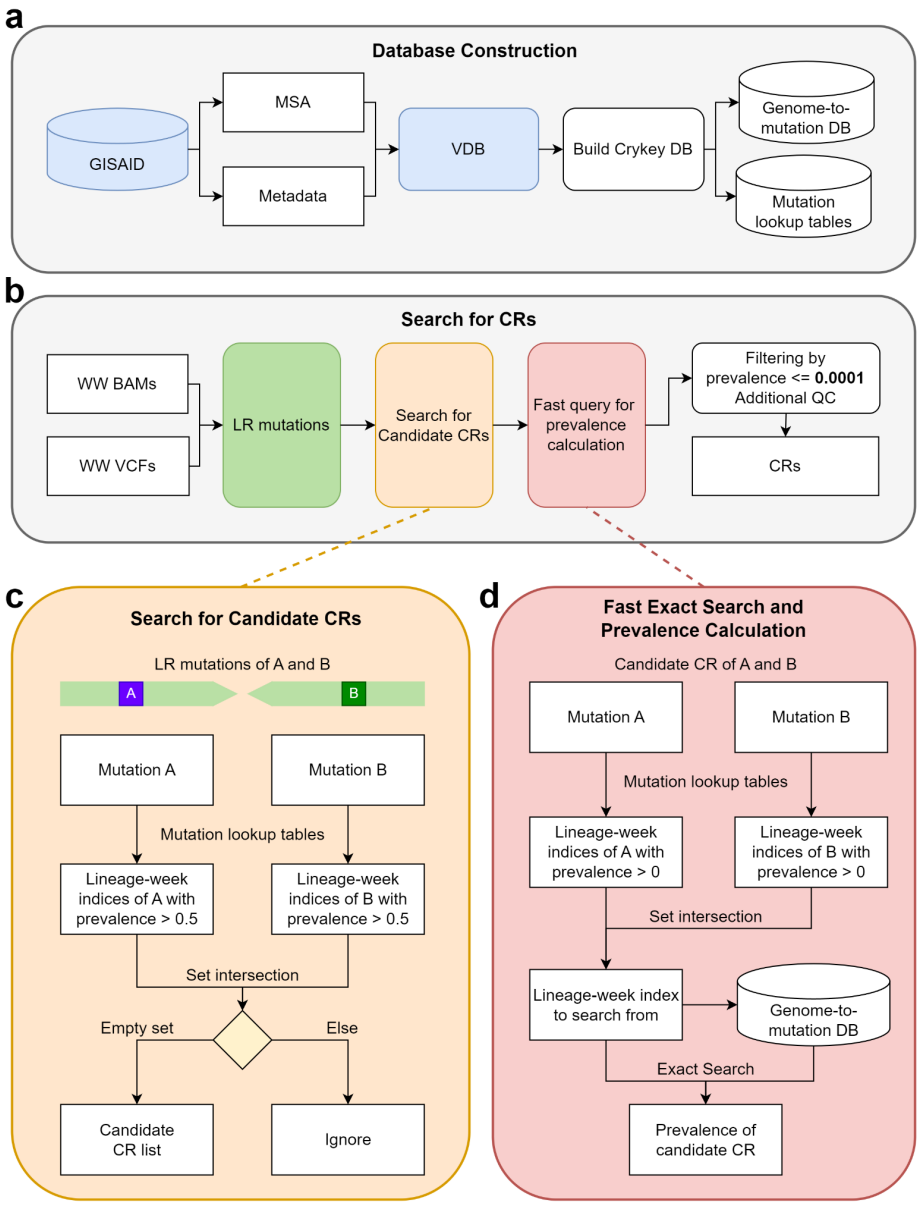

Figure 1. Workflow and algorithms of Crykey. a) Crykey constructs a genome-to-mutation database and a set of mutation lookup tables using GISAID sequences and metadata. **b)** Crykey searches for two or more mutations located on the same read or read-pair and uses the mutation lookup tables to identify whether the linked read mutations represent a candidate CR. Then, each candidate CR is queried against the genome-to-mutation database to calculate its prevalence rate; if they meet the indicated thresholds they are then considered a CR. **c)** Algorithm to search candidate CRs, with an example of a read-pair containing mutations A and B. **d)** Algorithm for the fast exact search for prevalence calculation, with an example of a candidate CR containing mutations A and B.

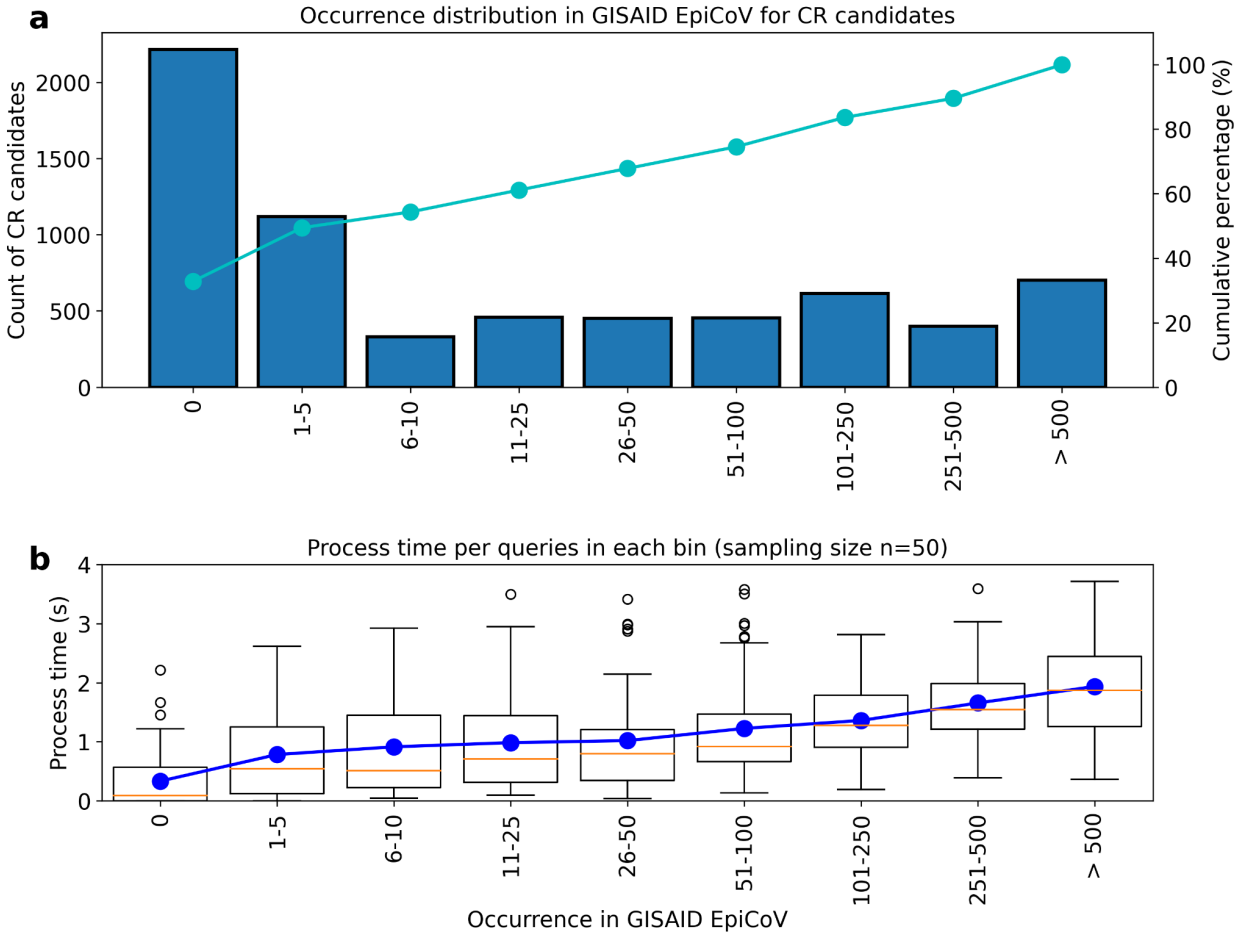

Figure 2. Occurrence distribution and query time of candidate CRs found in Houston wastewater samples. The candidate CRs identified in the samples are partitioned into bins based on their prevalence in the GISAID database. a) y-axis shows the number of candidate CRs in each bin (n=50). Cumulative percentages are plotted with a solid line on the second y-axis. b) shows the process time of each bins in the box plot. The box plot includes both median lines (solid), and the box bounds the interquartile range (IQR). The Tukey-style whiskers extend from the box by at most 1.5 × IQR. The outliers are shown. The average process time of each bin is shown as a solid blue line.

Figure 3. Distribution of CRs found in Houston wastewater. In both a) and b), the locations of CRs on the SARS-CoV-2 reference genome found in Houston wastewater samples are shown on the x-axis, with SARS-CoV-2 ORFs shown above the figure panels. In panel a), each CR is represented by a colored dot, the y-axis indicates its mean AF in the wastewater sample, and the color indicates its rarity, defined as -log10((n+1)/total number of sequences in GISAID), where n is the number of genomes supporting the CR in the GISAID EpiCoV database; the larger the number the more rare the mutation in GISAID. The darker color suggests that the CR is rare or unreported. The size of the dot shows the number of weeks the CR was detected. Larger dots indicate the CR persisted longer in the community. Panel b) is a histogram showing the count of CRs found in different 400 bp regions of the reference genome. CRs containing exclusively non-synonymous mutations are marked in orange, and the CRs containing at least one synonymous mutation are marked in gray. Higher bars indicate that more CRs were found in the associated region.

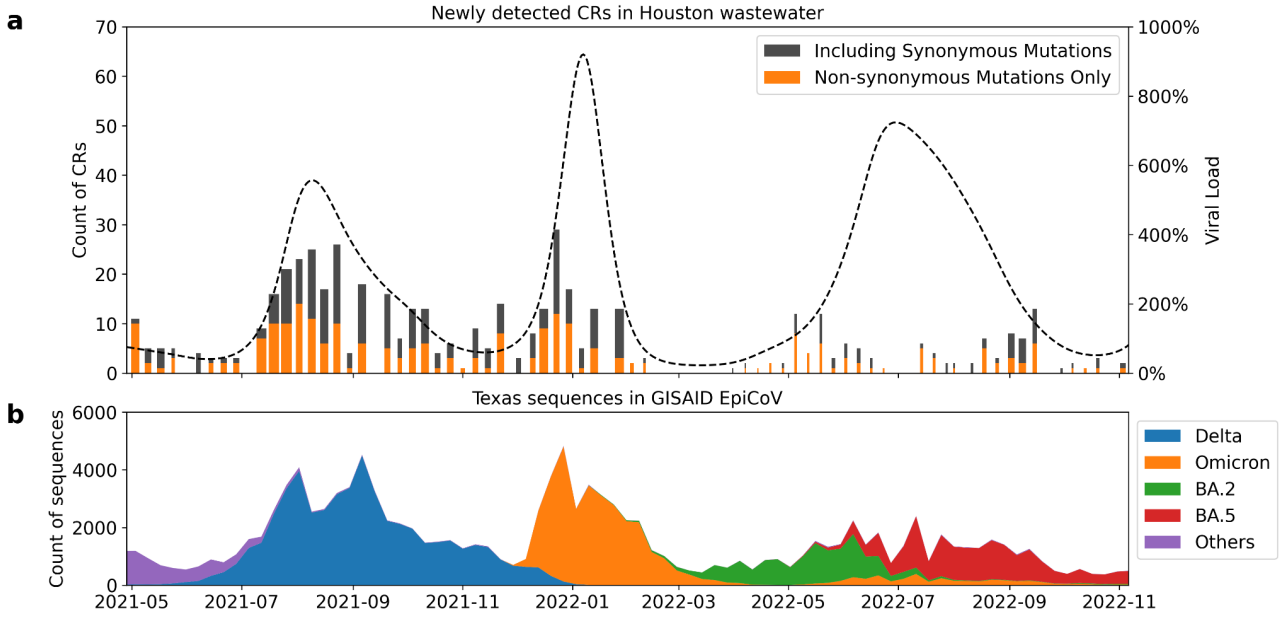

Figure 4. CRs and viral load in Houston wastewater. In both panels a) and b), the x-axis shows the dates from May, 2021 to November 2022. Panel a) shows the number of CRs (left y-axis) newly detected in Houston wastewater per week as bars. The proportion of CRs containing only non-synonymous mutations is indicated orange, while the remainder is in gray. The width of the bar indicates the average breadth of genome coverage across all WWTP, ranging from 0.02 to 0.74. The normalized viral load in wastewater (right y-axis) (based on the viral load from samples collected on July 6, 2020 in Houston) is shown as a dotted line. Panel b) shows the number of SARS-CoV-2 sequences in the GISAID EpiCoV database from Texas, USA per sampling week. Color corresponds to their PANGO lineage assignments. Omicron lineages other than BA.2, BA.5, and their descendants are combined and denoted as “Omicron”.

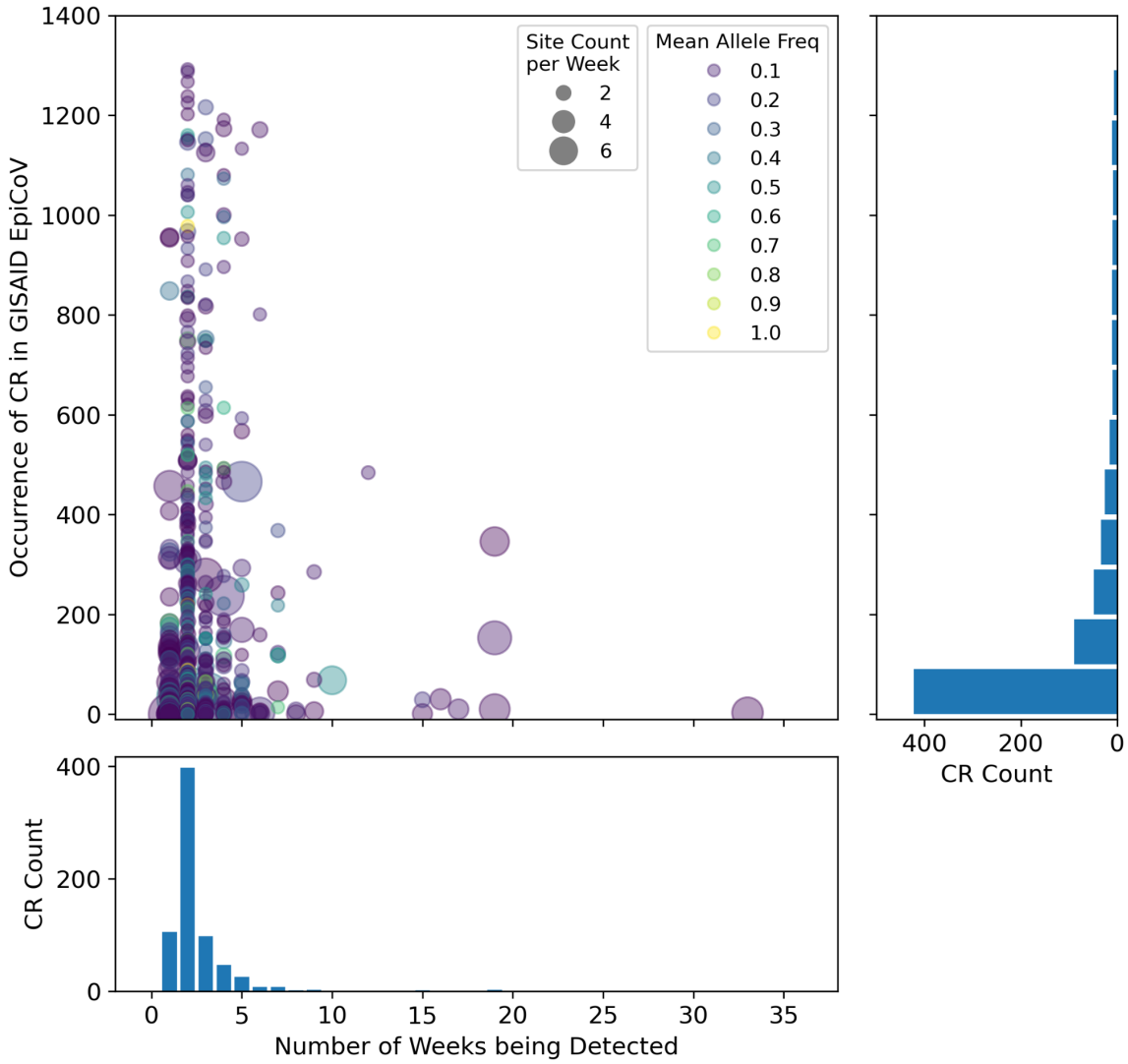

Figure 5. Persistence and Occurrence of CRs found in Houston wastewater. Each CR is represented by a dot, with the size of the dot indicating the mean count of the wastewater treatment plants the CR was detected at each week, and the color of the dot indicating mean allele frequency. The histogram on the bottom shows the number of weeks that the unique CRs have been detected and their associated counts. The histogram on the right shows the rarity of unique CRs in terms of occurrence in GISAID and their associated counts.

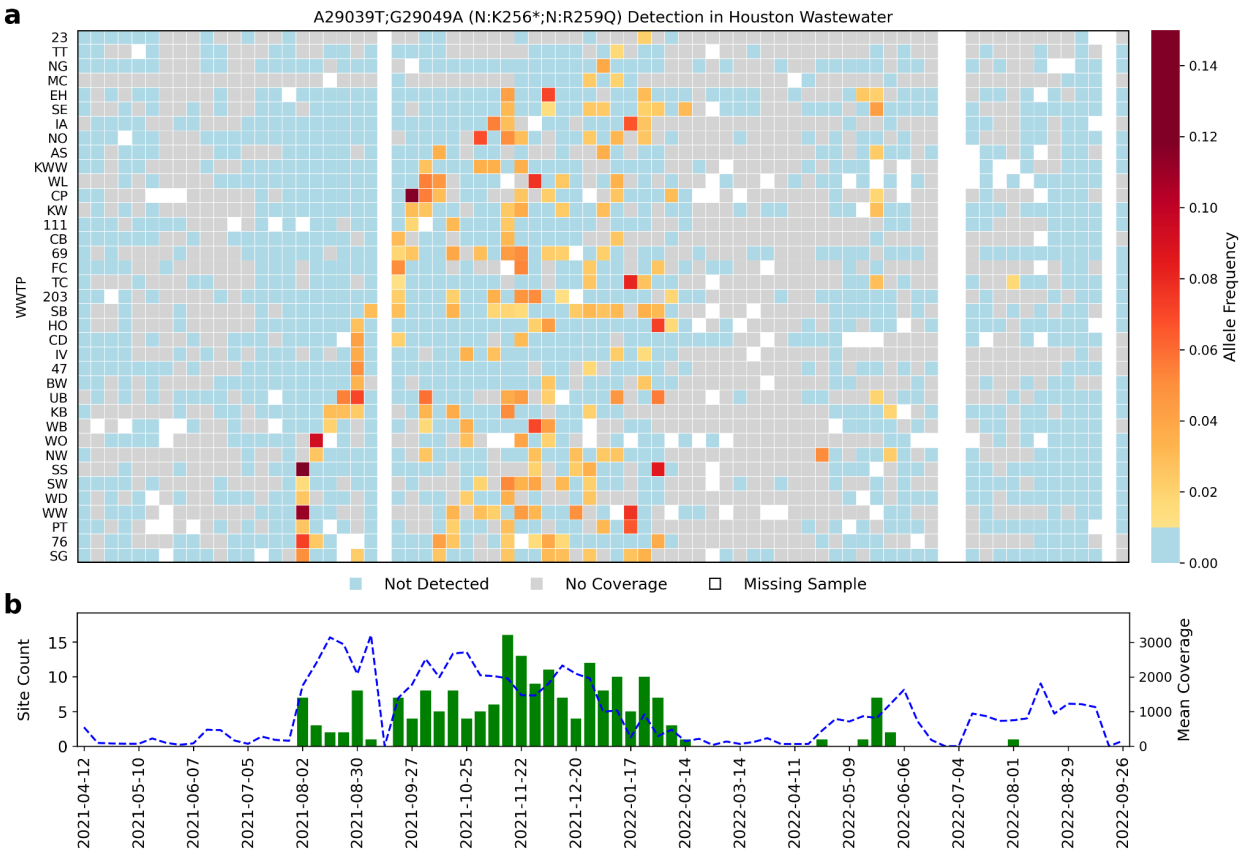

Figure 6. Persistence of CR12 (A29039T-G29049A) in Houston wastewater. For both panels, the x-axis shows time. In panel a), rows correspond to wastewater treatment plants (WWTPs) sampled, with the cell color indicating the mean allele frequency (MAF) of the mutation set (cells with MAFs below 0.01 are colored as blue and labeled as “Not Detected”. Samples with coverage below 10x are marked in gray and denoted as “No Coverage”. Missing samples are marked in white. In panel b) the bars indicate the number of WWTPs in which CR12 was detected, per week (left y-axis). The dotted blue line indicates the mean coverage of the wastewater samples with coverage above 10x, per week (right y-axis).

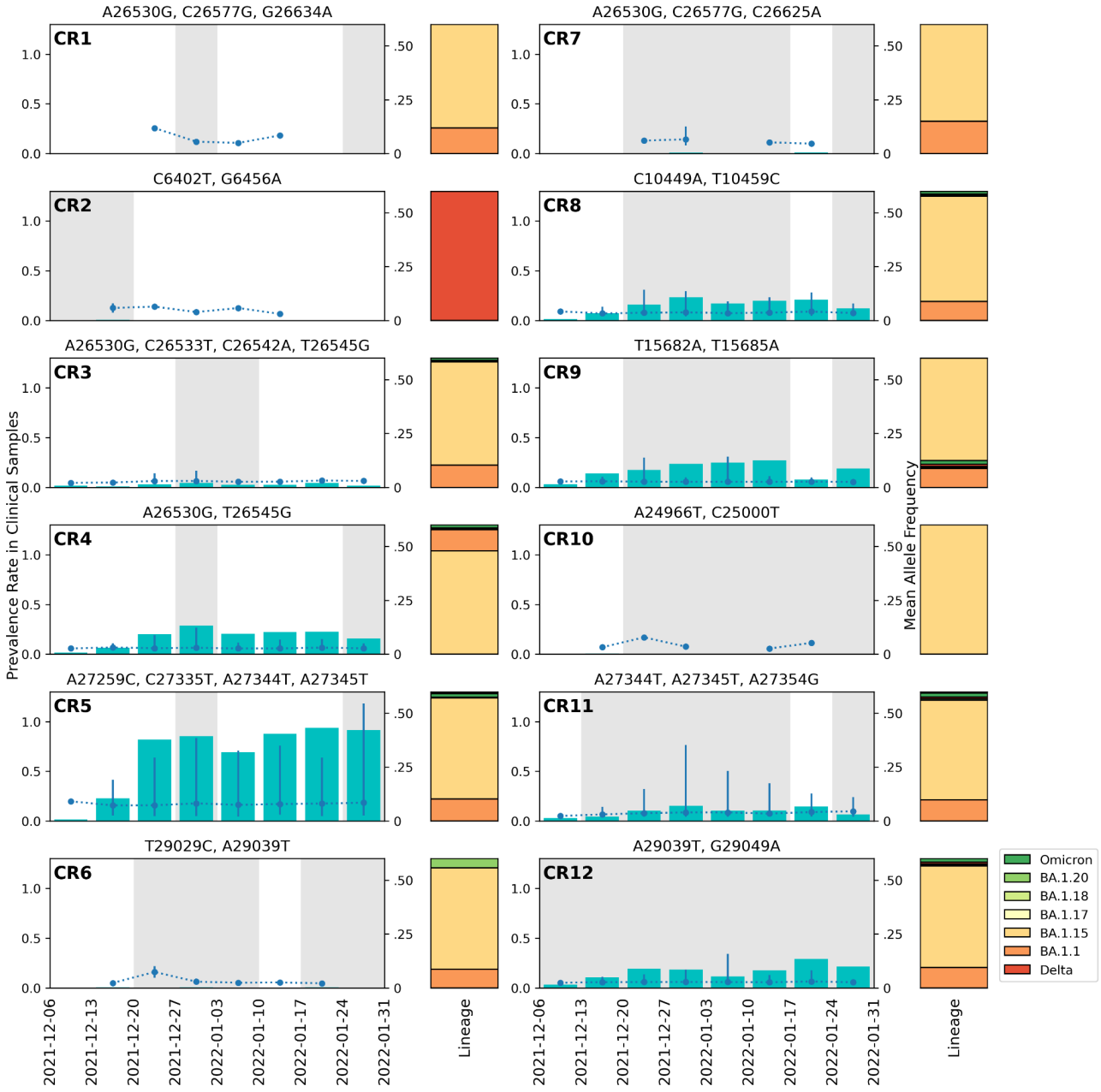

Figure 7. CRs detected in clinical samples from Greater Houston. The mutation combinations for each wastewater CR are shown at the top of each panel. Cyan bars indicate the prevalence (left y-axis) and dotted blue lines (right y-axis) the mean AF (right y-axis) of the CR in the clinical samples, with the error bars showing the minimum and maximum of the observed AF. The areas shaded in gray indicate the periods during which the CR was detected in the wastewater. The stacked bars to the right of the panels show the distribution of the PANGO lineages of the consensus genomes of clinical samples with CRs. All Delta genomes are combined. All Omicron genomes other than BA.1.1, BA.1.15, BA.1.17, BA.1.18, BA.1.20 are combined and denoted as Omicron.

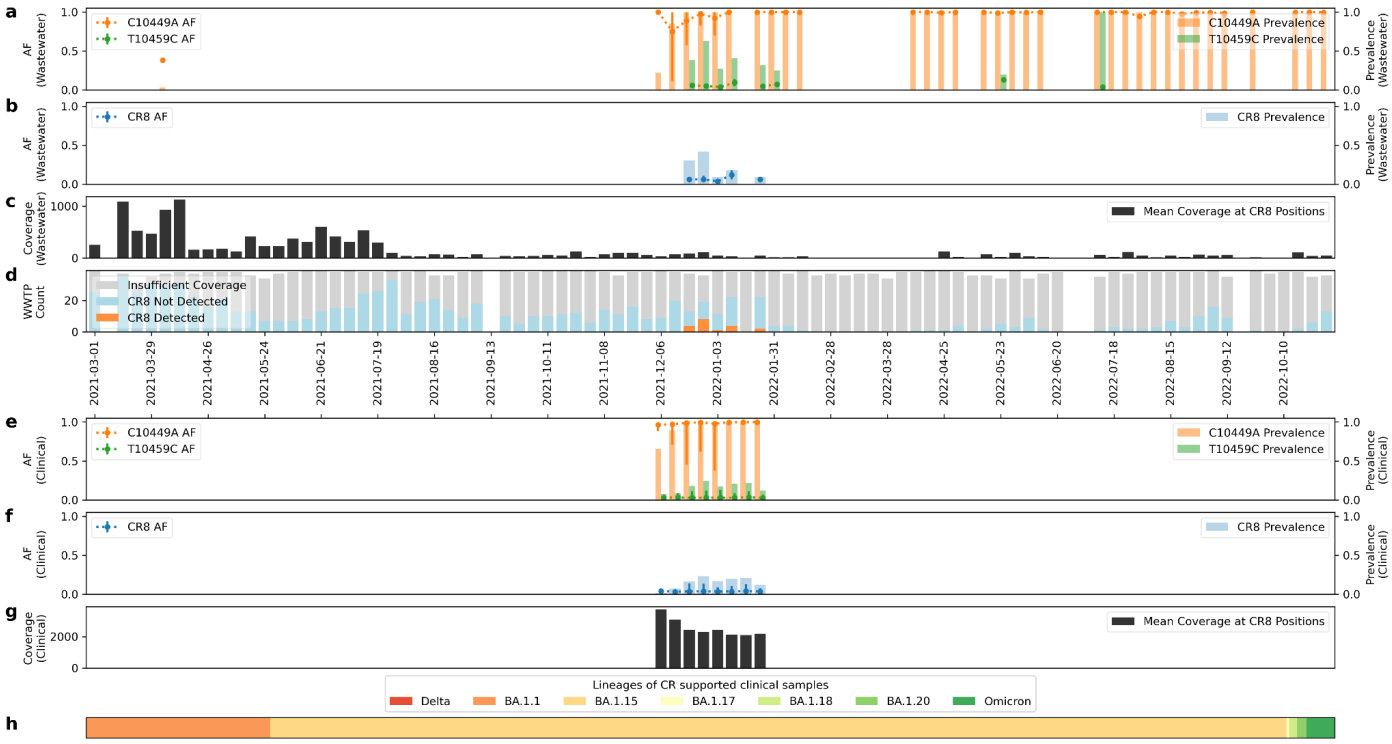

Figure 8. CR8 detected in wastewater and clinical samples from Greater Houston. Figure a-d are information of CR in wastewater each week. a) shows the individual AF (with mean AF shown in dotted line, minimum/maximum of AF observed shown as error bars, the same below) and prevalence rate (shown as bars, the same below) of mutations within CR. b) shows the AF and prevalence rate of CR. c) shows the mean coverage at CR5 locations. d) shows the sample qualities and Crykey detections, with samples of insufficient coverage colored in gray, samples of CR absent colored in blue, and samples of CR detected colored in orange. Figure e-h are information of CR in clinical samples of Houston for 8 weeks of sampling period. e) shows the individual AF and prevalence rate of mutations within CR. f) shows AF and prevalence rate of CR. g) shows the mean coverage at CR locations. h) shows the distribution of the PANGO lineages of the consensus genomes of the clinical samples with CR. Delta genomes are not found in any of the samples. All Omicron genomes other than BA.1.1, BA.1.15, BA.1.17, BA.1.18, BA.1.20 are combined and denoted as Omicron. For figure a-c, and e-h, wastewater and clinical samples with insufficient coverage (<10x at CR location) are excluded from the analysis.

References

1. Herold, M. *et al.* Genome Sequencing of SARS-CoV-2 Allows Monitoring of Variants of Concern through Wastewater. *Water* vol. 13 3018 Preprint at https://doi.org/10.3390/w13213018 (2021).
2. Fontenele, R. S. *et al.* High-throughput sequencing of SARS-CoV-2 in wastewater provides insights into circulating variants. *Water Res.* **205**, 117710 (2021).
3. Randazzo, W. *et al.* SARS-CoV-2 RNA titers in wastewater anticipated COVID-19 occurrence in a low prevalence area. Preprint at https://doi.org/10.1101/2020.04.22.20075200.
4. McClary-Gutierrez, J. *et al.* Sars-Cov-2 Wastewater Surveillance for Public Health Action: Connecting Perspectives From Wastewater Researchers and Public Health Officials During a Global Pandemic. Preprint at https://doi.org/10.20944/preprints202104.0167.v1.
5. Polo, D. *et al.* Making waves: Wastewater-based epidemiology for COVID-19 - approaches and challenges for surveillance and prediction. *Water Res.* **186**, 116404 (2020).
6. Wu, F. *et al.* SARS-CoV-2 Titers in Wastewater Are Higher than Expected from Clinically Confirmed Cases. *mSystems* **5**, (2020).
7. Peccia, J. *et al.* Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nat. Biotechnol.* **38**, 1164–1167 (2020).
8. Kitajima, M. *et al.* SARS-CoV-2 in wastewater: State of the knowledge and research needs. *Sci. Total Environ.* **739**, 139076 (2020).
9. Sapoval, N. *et al.* Enabling accurate and early detection of recently emerged SARS-CoV-2 variants of concern in wastewater. *Nat. Commun.* **14**, 2834 (2023).
10. Karthikeyan, S. *et al.* Wastewater sequencing uncovers early, cryptic SARS-CoV-2 variant transmission. *medRxiv* (2022) doi:10.1101/2021.12.21.21268143.
11. Kirby, A. E. *et al.* Notes from the Field: Early Evidence of the SARS-CoV-2 B.1.1.529 (Omicron) Variant in Community Wastewater - United States, November-December 2021. *MMWR Morb. Mortal. Wkly. Rep.* **71**, 103–105 (2022).
12. Ellmen, I. *et al.* Alcov: Estimating Variant of Concern Abundance from SARS-CoV-2 Wastewater Sequencing Data. Preprint at https://doi.org/10.1101/2021.06.03.21258306.
13. Crits-Christoph, A. *et al.* Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *MBio*

12, (2021).

14. Amman, F. *et al.* Viral variant-resolved wastewater surveillance of SARS-CoV-2 at national scale. *Nat. Biotechnol.* **40**, 1814–1822 (2022).
15. Jahn, K. *et al.* Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. *Nat Microbiol* **7**, 1151–1160 (2022).
16. Wolfe, M. *et al.* Detection of SARS-CoV-2 Variants Mu, Beta, Gamma, Lambda, Delta, Alpha, and Omicron in Wastewater Settled Solids Using Mutation-Specific Assays Is Associated with Regional Detection of Variants in Clinical Samples. *Appl. Environ. Microbiol.* (2022) doi:10.1128/aem.00045-22.
17. Baaijens, J. A. *et al.* Lineage abundance estimation for SARS-CoV-2 in wastewater using transcriptome quantification techniques. *Genome Biol.* **23**, 1–20 (2022).
18. Brunner, F. S. *et al.* City-wide wastewater genomic surveillance through the successive emergence of SARS-CoV-2 Alpha and Delta variants. *Water Res.* **226**, (2022).
19. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
20. Smyth, D. S. *et al.* Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat. Commun.* **13**, 1–9 (2022).
21. Gregory, D. A. *et al.* Genetic diversity and evolutionary convergence of cryptic SARS- CoV-2 lineages detected via wastewater sequencing. *PLoS Pathog.* **18**, (2022).
22. McCall, C. *et al.* Modeling SARS-CoV-2 RNA Degradation in Small and Large Sewersheds. Preprint at https://doi.org/10.1101/2021.09.17.21263708.
23. Wu, F. *et al.* SARS-CoV-2 RNA concentrations in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases. *Sci. Total Environ.* **805**, 150121 (2022).
24. Shafer, M. M. *et al.* Tracing the origin of SARS-CoV-2 Omicron-like spike sequences detected in wastewater. *medRxiv* 2022.10.28.22281553 (2023) doi:10.1101/2022.10.28.22281553.
25. Karthikeyan, S. *et al.* Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* **609**, 101–108 (2022).
26. Alexandersen, S., Chamings, A. & Bhatta, T. R. SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication. *Nat. Commun.* **11**, 1–13 (2020).

27. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
28. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
29. Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Weekly* **3**, 1049 (2021).
30. Gangavarapu, K. *et al.* Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat. Methods* **20**, 512–522 (2023).
31. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* 1–19 (2023).
32. Lu, S. *et al.* The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat. Commun.* **12**, 1–15 (2021).
33. Armero, A., Berthet, N. & Avarre, J.-C. Intra-Host Diversity of SARS-Cov-2 Should Not Be Neglected: Case of the State of Victoria, Australia. *Viruses* **13**, (2021).
34. Rafael Ciges-Tomas, J., Franco, M. L. & Vilar, M. Identification of a guanine-specific pocket in the protein N of SARS-CoV-2. *Communications Biology* **5**, 1–9 (2022).
35. Hughes, L. *et al.* Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Res Sq* (2022) doi:10.21203/rs.3.rs-1723829/v1.
36. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat. Commun.* **12**, 1–17 (2021).
37. Sanderson, T. *et al.* A molnupiravir-associated mutational signature in global SARS-CoV-2 genomes. *Nature* 1–3 (2023).
38. Yousif, M. *et al.* SARS-CoV-2 genomic surveillance in wastewater as a model for monitoring evolution of endemic viruses. *Nat. Commun.* **14**, 1–9 (2023).
39. Factors influencing SARS-CoV-2 RNA concentrations in wastewater up to the sampling stage: A systematic review. *Sci. Total Environ.* **820**, 153290 (2022).
40. Kantor, R. S., Nelson, K. L., Greenwald, H. D. & Kennedy, L. C. Challenges in Measuring the Recovery of SARS-CoV-2 from Wastewater. *Environ. Sci. Technol.* (2021) doi:10.1021/acs.est.0c08210.
41. Defining biological and biophysical properties of SARS-CoV-2 genetic material in wastewater. *Sci. Total Environ.* **807**, 150786 (2022).
42. Next generation sequencing approaches to evaluate water and wastewater quality. *Water Res.* **194**, 116907 (2021).

43. Wolken, M. *et al.* Wastewater surveillance of SARS-CoV-2 and influenza in preK-12 schools shows school, community, and citywide infections. *Water Res.* **231**, 119648 (2023).
44. Wolfe, M. K. *et al.* Wastewater-Based Detection of Two Influenza Outbreaks. *Environmental Science & Technology Letters* (2022) doi:10.1021/acs.estlett.2c00350.
45. West, A. P. *et al.* Detection and characterization of the SARS-CoV-2 lineage B.1.526 in New York. *Nat. Commun.* **12**, 1–10 (2021).
46. Sapoval, N. *et al.* SARS-CoV-2 genomic diversity and the implications for qRT-PCR diagnostics and transmission. *Genome Res.* **31**, 635–644 (2021).
47. De Maio, N. *et al.* Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
48. Leinonen, R., Sugawara, H. & Shumway, M. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19 (2011).
49. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
50. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189 (2012).
52. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* **7**, veab064 (2021).
53. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**, 1–11 (2012).
54. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 1–19 (2019).
55. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
56. Omicron variant (B.1.1.529) and its sublineages: What do we know so far amid the emergence of recombinant variants of SARS-CoV-2? *Biomed. Pharmacother.* **154**, 113522 (2022).
57. Chakraborty, C., Bhattacharya, M., Sharma, A. R. & Dhama, K. Recombinant SARS-CoV-2 variants XD, XE, and XF: The emergence of recombinant variants requires an urgent call for research – Correspondence. *Int. J. Surg.* **102**, 106670 (2022).