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

Yunxi Liu[1], Nicolae Sapoval[1], Todd J. Treangen[1*] , Lauren B. Stadler[2*]

[1]Department of Computer Science, Rice University, Houston, TX, 77005, USA
[2]Department of Civil and Environmental Engineering, Rice University, Houston, TX, 77005, USA
*Corresponding authors: treangen@rice.edu, lauren.stadler@rice.edu

Abstract

We present Crykey, a computational tool for identifying SARS-CoV-2 cryptic mutations from wastewater. While previous exist for identifying cryptic mutations in specific regions of the SARS-CoV-2 genome, there is a need for computational tools capable of tracking cryptic mutations across the entire genome and at scale. Crykey fills this gap and leverages the co-occurrence of single nucleotide variants on the same read combined with variant frequency information. We evaluated Crykey on SARS-CoV-2 sequences from 3175 wastewater samples and more than 14000 clinical samples. Our results are threefold, we show: 1) Crykey can accurately identify cryptic lineages that are rare or missing in existing databases ; 2) the emergence of cryptic lineage can be related to increased transmission rates in the communities, and 3) some cryptic lineages in wastewater mirror intra-host low frequency co-occurring variants in individuals. In summary, Crykey facilitates rapid and comprehensive identification of SARS-CoV-2 cryptic mutations in wastewater samples.

Introduction

Wastewater monitoring serves as an important tool that can supplement clinical testing for COVID surveillance[1–9]. Multiple studies have demonstrated that the signal of SARS-CoV-2 variants of concern (VOC) can be extracted from wastewater samples collected from local regions[10–15]. This allows early detection of community spread of the viral variants that precedes clinical testing by up to two weeks[9]. Further, wastewater provides information on the genomic diversity and lineage abundance estimation of circulating lineages in the community, and overcomes the sampling bias inherent to clinical surveillance[16–18].

Wastewater monitoring for SARS-CoV-2 can also be used to detect novel cryptic lineages[19]; here, we define a cryptic lineage of SARS-CoV-2 as a set of co-occurring mutations that have never been reported or rarely observed (prevalence less than 0.0001) in publically available assembled genomes. However, non-uniform sequencing coverage caused by amplicon efficiency difference and environmental RNA degradation creates a challenge for detecting and phasing cryptic lineages from variant calling results[5,20,21]. Previous methods have reported the detection of cryptic lineages from wastewater monitoring, but they required ultra-deep sequencing on specific targeted regions of SARS-CoV-2, or hybrid sequencing technology, and thus are not compatible with most wastewater sequencing protocols used for routine monitoring[16, 22]. In addition, the origin of the cryptic lineages is still an open question. Previous studies have reported that rare lineages are found in wastewater samples but not in clinical samples, suggesting that undersampling in clinical surveillance might explain the case[23,24]. There is also some speculation that cryptic lineages are associated with human intra-host minor variants, or originate from non-human hosts[19].

In this manuscript, we introduce Crykey, a fast computational method for detecting cryptic lineages of SARS-CoV-2 from wastewater samples that exploits co-occurrence of SNVs on the same sequencing read across the full length of SARS-CoV-2 genome. We used Crykey to analyze SARS-CoV-2 sequencing data from wastewater samples collected in Houston, Texas, USA. Collectively, our results highlight that Crykey can accurately identify cryptic combinations of SNVs from the sequencing data that have not been found or have low prevalence (less than 0.0001) in GISAID's EpiCoV database.

To investigate the derivation of cryptic lineages, we took a closer look at the cryptic lineages found in Houston wastewater, and found that the number of cryptic lineages increases when a new strain of VOC starts to circulate in the community. By analyzing more than 5000 clinical samples collected from the Houston area, and 9000 clinical samples collected across the US from other states, we found longitudinal connections between intra-host minor variants and wastewater cryptic lineages. Our result indicates that a large proportion of the cryptic lineages, especially long lasting cryptic lineages, are derived from human intra-host landscape. We found the cryptic lineages are geographically constrained, and are specific to certain PANGO lineages of the consensus genome of the human hosts. This research shows that wastewater monitoring can detect low frequency intra-host variants, which may be useful for understanding transmission events in communities.

Results

Crykey is a computational method for cryptic lineage identification in wastewater on a full genome scale. Figure 1 shows the workflow and algorithm of Crykey. Crykey first selects candidate cryptic lineages by searching co-occurring mutations supported by the same sequencing read (Figure 1c). Then each candidate is queried against a pre-built database to check if the combination of the mutations is novel/rare in terms of its prevalence

(Figure 1d). The detailed performance benchmark can be found in Supplementary Section 1. We applied Crykey on SARS-CoV-2 data sequenced from 3,175 wastewater samples collected from Houston wastewater treatment plants. The samples were collected from February 2021 to November 2022, and identified a large number of mutation combinations that originated from cryptic lineages of SARS-CoV-2.

Genomic Distribution of Cryptic Mutation Sets

We identified a total of 716 cryptic mutation sets in our samples (see Methods). Figure 2a shows the location of the cryptic mutation sets on the reference genome, their mean allele frequencies, rarity in the GISAID database, and the number of weeks (not necessarily consecutive) detected in wastewater samples. We observed cryptic mutation sets across the entire SARS-CoV-2 genome. Some regions of the genome were enriched in cryptic mutations, such as the S and N genes (see Figure 2b) with 20.3% of the cryptic mutation sets containing mutations located on the S gene, and 27.9% of the cryptic mutation sets containing mutations located on the N gene.

More than 92% of the cryptic mutation sets had mean allele frequency less than 0.5 in the wastewater samples, with a few exceptions. We found that the most cryptic mutation sets that were rare in the GISAID database also have low mean allele frequency in the samples that support them (see Figure 2a). The occurrence of those cryptic mutation sets (counted by week) varied significantly, ranging from 1 week to 33 weeks, which are shown by the size of the dot in Figure 2a. In most regions of the SARS-CoV-2 genome, non-synonymous cryptic mutation sets are dominated in number compared to cryptic mutation sets that include synonymous mutations. The only exception is the N gene (see Figure 2b). We also calculated the dN/dS ratio of unique SNVs in cryptic lineage and found a noticeable dN/dS increase on S gene, indicating positive selection effects (See Figure 4). The detailed results of dN/dS analysis can be found in Supplementary Section 2.

Emergence of cryptic lineages co-occurred with surges in community infections driven by new variants

The emergence of the cryptic mutations co-occurred with increases in SARS-CoV-2 infections across the city, and corresponding increase in the citywide wastewater viral load, which happened when a new VOC started to circulate in the community. For example, the number of new cryptic lineages and viral load both increased significantly around July, 2021 (see Figure 3a), which corresponded to the Delta wave in Houston and (see Figure 3b). Similar patterns were observed during the emergence of B.1.1.529 lineage (Omicron) in December, 2021, and the emergence of BA.2 (Omicron) in May, 2022, and BA.5 (Omicron) in July, 2022. This may be the effect of both new VOC strains introducing associated cryptic mutations, as well as increasing viral load causing more signals to be captured.

The detection of cryptic lineages is also a function of data quality. We observed fewer new cryptic lineages during the BA.2 and BA.5 period (from Apr, 2022 to Aug, 2022) when the data quality dropped and became a limiting factor of the detection of the cryptic lineages. The different bin width in Figure 3a indicates the normalized sample qualities derived from breadth of the genome coverage of the wastewater samples collected from the same week, where wider bins represent better data quality.

Long Lasting Cryptic Lineage

We next asked whether there were any cryptic lineages that persisted over several weeks to months in the Houston community. Most of the cryptic lineages did not persist for a long period of time, with over 85.4%

(612/716) of the cryptic mutations found in 3 or less distinct time points (samples are collected on a weekly frequency). Interestingly, some cryptic mutations were detected in samples from multiple wastewater treatment plants across the city and persisted for over 10 weeks (Supplementary Figure 1). Figure 5 shows the most persistent cryptic lineage observed in Houston wastewater; it contains the co-occurrence of two SNVs, A29039T and G29049A, which causes K256* (stop codon) and R259Q amino acid changes on the N gene. This cryptic lineage was detected over 33 weeks (shown in Figure 5b), mostly between Aug, 2021, and Feb, 2022. Detections peaked in late November, 2021, where the cryptic lineage was observed in 16 wastewater treatment plants concurrently (shown in Figure 5a). The presence of this cryptic lineage phased out in late February 2022 and remained silent for two months. Then it re-appeared for a short period of time around May, 2022. Figure 5a shows that the mean allele frequencies of the combination of the SNVs were generally lower than 0.10, with a few exceptions at some wastewater treatment plants during early weeks. The count of supported reads that contained both SNVs at the same time are shown using different colors.

Mutations in Cryptic Lineages

We next asked whether the individual mutations that make up the cryptic lineages were associated with other known PANGO lineages. To do this, we defined a mutation that is present in more than 50% of the assemblies in GISAID under the same known PANGO lineage as a signature variant for that PANGO lineage[25]. Based on the PANGO associated signature variants that each cryptic lineage contained, we classified the wastewater cryptic lineages into three categories: cryptic lineages that hold no signature variants, cryptic lineages that hold signature variants specific to a single PANGO lineage, cryptic lineages that contain signature variants that are specific to more than one different PANGO lineages.

We found that more than 76.5% (548 out of 716) of cryptic lineages contained signature variants specific to only one PANGO lineage. In less than 6.1% (44 out of 716) of the cases, the cryptic lineages did not contain any signature variants specific to any known PANGO lineage, which means all of the mutations had a prevalence rate less than 0.5 in all PANGO lineages. The remaining 17.3% (124 out of 716) cryptic lineages held signature variants specific to more than one different PANGO lineage, which are likely indicative of recombination events.

Cryptic lineage Detection in Clinical Samples

The sources of cryptic lineages present in wastewater are not well understood. Possible sources include undersampling in clinical surveillance, contribution from intra-host low allele frequency variants, and possible animal hosts. To test whether the cryptic lineages originated from an intra-host environment, we selected 5,060 sequenced clinical samples collected within Greater Houston between 12/06/2021 to 01/31/2022 (8 weeks). We checked whether the co-occurrence of the mutations in 20 cryptic lineages detected in Houston wastewater were also present in the clinical sequencing reads. The details of sample processing and cryptic lineage selection can be found in the methods section and Supplementary Figure 2 shows the distribution of samples over the 8-week period. We calculated the prevalence rate of the cryptic lineage in clinical samples as the number of samples supporting the cryptic lineage over the total number of samples.

We found that more than half of the 20 cryptic lineages investigated were also prevalent in clinical sequencing reads. Figure 6 shows the top 12 cryptic lineages of the highest prevalence rate in the clinical samples, including 5 short-duration cryptic lineages CR1-CR5 (detected only over 2 weeks), and 7 long duration cryptic lineages CR6-CR12 (4-8 week occurrence). The results suggested that both long- and short-duration cryptic lineages detected in wastewater can be found within clinical samples at low allele frequencies, most of the time

below 0.05. Of the 12 wastewater cryptic lineages with the highest clinical prevalence rates, 11 were associated with the Omicron strains, except CR2. The PANGO assignment of the consensus genomes shows most of the cryptic-supported clinical samples were labeled as BA.1.15 and BA.1.1. The Omicron-related cryptic lineages shared a pattern that the prevalence rate in the clinical samples increased as the Omicron variant spread in the city, which is reflected by both the viral load in the wastewater (shown in Figure 3a) and sequence count in the database (shown in Figure 3b). These cryptic lineages appeared as intrahost variants in clinical samples prior to appearing in wastewater samples. In contrast,C6402T-G6456A, a cryptic lineage specific to Delta strain,was detected in the wastewater in the first 2 weeks of the 8-week period, while also being present at trace levels over the first 6 weeks in the clinical samples.

The intra-host allele frequency of a cryptic lineage varies by host. Supplementary Figure 4 shows the intra-host allele frequencies for the cryptic lineage A29039T-G29049A during the clinical sampling period. However, the overall mean AF is relatively stable as time goes by, which suggests the detection of cryptic lineages in wastewater samples is determined by the prevalence rate (vary by specific cryptic lineages) in the population. As a VOC, in this case Omicron, starts to become dominated in the region, new cryptic lineage specific to the VOC emerges and gains its popularity among the hosts. At the same time, old cryptic lineage specific to the previous dominating VOC, in this case Delta, dies out, as fewer and fewer people are infected with its associated strain.

Geographically Constrained Cryptic Lineages

To assess whether there were geographic patterns at a national level associated with cryptic lineages,we processed 8,969 clinical samples collected from states outside of Texas over the same 8-week time period (between 12/06/21 and 01/31/22) as from Texas (see Supplementary Figure 3 for detailed geological distribution). We used the same method to search for the previous most significant 12 cryptic lineages found in wastewater and clinical samples from Houston for those clinical samples outside of Texas.

Two of the cryptic lineages were shared across clinical samples from Houston, Maryland, and Massachusetts. In addition, we identified 5 additional cryptic lineages shared across clinical samples from Houston and Maryland. Five of the cryptic lineages were specific to Houston and were not found in other states. Cryptic lineage A27259C-C27335T-A27344T-A27345T was found in samples from Houston, Maryland, and Massachusetts (see Figure 7a), and cryptic lineage C10449A-T10459C, was found in Houston and Maryland (see Figure 7b). We found although the same cryptic lineage may be shared by specific regions, the prevalence rate can vary between each of them. For A27259C-C27335T-A27344T-A27345T, Houston and Maryland have a much higher proportion of the cryptic samples supporting the cryptic lineage as compared to Massachusetts. In addition, the demographics of PANGO assignments for the cryptic supporting samples were also different between states. Although the two cryptic lineages are all associated with Omicron, Houston was dominated by BA.1.15 while Maryland and Massachusetts had a much higher proportion of BA.1.1, BA,1.18, as well as BA.1.17 for Maryland only.

Discussion

Wastewater monitoring for SARS-CoV-2 has been widely used for genomic surveillance during the COVID pandemic[14]. It reveals information about genetic diversity of the virus within communities that can complement

clinical surveillance[26]. Recent studies have shown that the presence of novel variants may be introduced during the transmission of SARS-CoV-2, and by exploring wastewater data scientists have been able to identify novel strains of SARS-CoV-2 that are not captured by clinical sampling and sequencing[19].

Our contribution centers on a novel cryptic lineage detection tool, Crykey, designed for the analysis of wastewater sequencing data. Specifically, Crykey detects the co-occurence of the SNVs that are present on the same reads or read pairs and efficiently cross-references them against databases of SARS-CoV-2 genomes. Our method is fully compatible with standard variant calling pipeline for SARS-CoV-2 genomic analysis, and takes advantage of variant calling results across the entire SARS-CoV-2 genome. We highlight the utility of genome-wide identification of cryptic mutations with Crykey via the identification of multiple novel cryptic lineages in the Houston wastewater.

Our findings show that cryptic lineages emerge during massive transmission events as a circulating lineage of the SARS-CoV-2 in the communities becomes dominant. We found that the waves of newly emerging cryptic lineages from Houston tie well with number of Delta strain genomes submitted from Texas in the GISAID database, Omicron (B.1.1.529 and its descendants, excluding BA.2), BA.2 (B.1.1.529.2 and its descendants, excluding BA.5), and BA.5 in Texas, USA (see Figure 3). As the viral load in wastewater increased, the detection of cryptic lineages also increased The BA.2 and BA.5 waves (May, 2022 to October, 2022) were exceptions to this trend, where there was a marked increase in wastewater viral load, however, new cryptic lineages appeared only at the beginning and towards the end of the viral load spike. 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 is a consequence of the lower quality of the sequencing data.

We also show that most of the cryptic lineages appeared over approximately 2 weeks, however we also observed some persistent cryptic lineages (Supplementary Figure 1). Short-duration cryptic lineages were also generally found in only a few wastewater treatment plants and with low allele frequencies. One possible explanation is that the viral shedding from a community was dominated by the established variant of concern (VOC) lineages that have spread across the city, and thus cryptic lineages make up a small fraction of the total abundance. Due to this, cryptic lineages in wastewater are more likely to remain below the detection limit until they reach some degree of prevalence in the population. We suspect that the short duration of cryptic lineages being observed indicates peaks of local spread of the cryptic lineage within a community. Once their prevalence in the population drops, those cryptic lineages were no longer detectable in wastewater samples.

On the other hand, some cryptic lineages lasted for months and reappeared multiple times. One of the cryptic lineages we identified present in 33 weeks of wastewater data across multiple wastewater treatment plants around Houston (see Figure 5). The cryptic lineage contains two nucleotide mutations, A29039T and G29049A, and cause K256* and R259Q amino acid changes on the N gene. The combination of these SNVs is rare, and there are only 3 sequences in the GISAID database that contain both mutations, and none of those sequences originated from the United States. Previous study shows that 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[27]. A29039T variant causes a stop codon to be generated, and this variant may affect the linker region suppressing the immunogenic domain of the nucleocapsid protein, which leads to possible vaccine escape[27,28]. This may explain the extreme persistence of this cryptic lineage. R259 in the protein N of SARS-CoV-2 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 in MERS and SARS-CoV-1[29]. However, R259Q mutation alone is reported multiple times at low prevalence rates in some of the SARS-CoV-2 lineages, mostly belonging to Delta variant[30]. To the best of our knowledge, the effect of the combination of these two mutations is still unknown and requires further investigation.

Why are cryptic lineages not captured by clinical surveillance?

One possible explanation is that cryptic lineages have low community prevalence rates[19,24,23]. As only a small portion of samples from individuals infected with SARS-CoV-2 are sequenced, cryptic lineages infecting a small proportion of the population 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. Supplementary Figure 1 shows that most of the cryptic lineages detected in Houston wastewater were only found over 1 to 3 weeks, and these short duration lineages may represent those not captured by clinical testing.

We also report the detection of multiple cryptic lineages that persisted for more than 10 weeks.These lineages more likely originate from cryptic lineages that were present in the clinical samples at low allele frequencies, and hence were part of the intra-host viral diversity[19,24]. 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 database such as GISAID. As a result, although the cryptic lineages were sampled, they remained unreported. To test this hypothesis, we evaluated a dataset of 5060 patient derived SARS-CoV-2 sequencing samples collected within Greater Houston (including Houston and its surrounding residential areas) from 12/06/2021 to 01/31/2022 (8 weeks), and performed read alignment for these data. Thus, we showed that cryptic lineages detected in wastewater originated from intra-host low frequency co-occurring variants in the clinical samples (Figures 5 and 6). Further, these cryptic lineages show strong geographic specificity. Even though some of the cryptic lineages found in Houston clinical samples were also found in Maryland and Massachusetts, they were specific to clinical samples from those two states at minor allele frequencies.

Challenges and limitations

One of the limitations of Crykey is that the detection of cryptic lineages is heavily dependent on the quality of the sequenced wastewater samples[31,32]. As shown in Figure 3, the number of newly emerging cryptic lineages follows the same pattern as the viral load until June 2022, where the samples collected afterwards had worse quality in terms of breadth of coverage. The performance of Crykey is limited because the samples do not have enough sequencing depth across most of the regions of SARS-CoV-2 genome during those weeks.

Due to the limitation of the short read platform and protocol used for sequencing the sample, there is a natural limit on the distance between the SNVs we can use for the co-occurrence analysis. Furthermore, regions corresponding to sequencing adapters create gaps along the genome and pose a challenge for identifying sets of cryptic mutations that span longer regions. However, these limitations could be addressed using long read sequencing, assuming intact longer amplicons are recoverable from wastewater samples.

The sources of cryptic lineages in wastewater and their relative contributions requires further research. We found that a subset of the cryptic lineages were supported by reads from clinical sequencing. However, there are some cases in which a cryptic lineage persisted for weeks in wastewater samples, but had little to no trace

in clinical samples. The key reasons why these cryptic lineages were not captured by clinical surveillance remains unknown. One hypothesis put forward by a previous study suggests that such cryptic lineages are carried by non-human hosts[19].

Cryptic lineages were detected in wastewater for time periods ranging from a single week to months of continuous detection. Explanations for these patterns and their variability requires further investigation. We observed that the number of unique emerging cryptic lineages was related to massive transmission events. We suspect the specific cryptic lineages were derivatives of the dominant circulating VOC strain in the population, and as new VOC stains became dominant, the cryptic lineages associated with the previous VOC strain faded out as well.

The concept of searching for combinations of mutations inside of a viral genome that have never, or rarely been reported is universal, and Crykey is not only limited to finding SARS-CoV-2 cryptic lineages in wastewater. The usage of the method could be expanded to multiple pathogens, such as influenza A virus, as long as the pathogens are being monitored in wastewater, and we have established a database of known assemblies for such pathogens[33,34].

In conclusion, Crykey is a software framework designed to identify cryptic lineage of SARS-CoV-2 from wastewater samples, by leveraging the detection co-occurrence of SNVs on the same sequencing reads of the full SARS-CoV-2 genome. We apply this tool to detect numerous novel cryptic lineages present in the population, some persisting for months. We show that for a subset of cryptic lineages, their source is likely intrahost minor variants that are widespread in the population and present a low allele frequencies within an individual infected with SARS-CoV-2. This represents the first study to show that wastewater monitoring can be used to detect population-level intrahost minor variants, which could be used to better understand transmission events at a large scale[35–37].

Methods

The workflow of the Crykey pipeline can be divided into 3 steps, including database construction, sample processing to find cryptic lineage candidates, and rarity calculations for each of the candidates found in the previous step. The following are the details for each of the steps, as well as the filtering and analysis methods used in this manuscript.

SNP Database Construction

The database used in Crykey is built based on the multiple sequence alignment (MSA) generated by the GISAID. We extracted the SNPs for each of the SARS-CoV-2 genome in the MSA using vdb with the command *vdbCreate -N input.msa*[38]. 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 lineages of SARS-CoV-2, as well as build a SNP database containing SNP information for each individual SARS-CoV-2 sequence.

Searching for Possible Cryptic Mutation Combinations

In order to identify the cryptic lineage, Crykey first builds a default mutation lookup table where each mutation in the GISAID is associated with a set of lineages and the 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 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.

Then for a given sample, Crykey takes its associated alignment file (BAM) and variant calling output (VCF) file as input. It first extracts and filters the SNPs from VCF file with a user defined minimum depth of coverage (default: 10) and a user defined minimum allele frequency (default: 0.02). Then, it annotates each SNPs with snpEff and removes mutations not in the coding region. For each sample, Crykey searches through the BAM file and extracts read pairs that contain a combination of multiple SNPs. Using the mutation lookup table of prevalence rate greater than 0.5, Crykey is able to quickly identify whether the SNP combination from the read may belong to a single lineage of a certain week by using set intersection. By the pigeonhole principle, if the intersection is non-empty, it is guaranteed that the SNP combination has been reported to the public database. If the intersection is an empty set, we consider the combination as a cryptic candidate set.

Rarity of Cryptic Mutation Combinations

For each cryptic candidate set of mutations, an exact search is performed by querying the SNPs in the set against the default mutation lookup table and the SNP database. By using the default mutation lookup table, the lineage and specific week of the co-occurrence of all SNPs in the set can be quickly determined, and allows Crykey to minimize the search space while querying the SNP database while searching for exact assemblies that contain such SNPs.

Prevalence of Cryptic Mutations

By evaluating the prevalence of the mutations in each of the cryptic lineages, we categorized the cryptic lineages into 3 categories: (1) cryptic lineages containing no signature variants of any known PANGO lineages; (2) cryptic lineages containing signature variants of a single PANGO lineage; and (3) cryptic lineages containing signature variants of multiple different PANGO lineages, where signature variant is defined as a mutation with prevalence rate over 0.5 in a known PANGO lineage. As expected, most of the cryptic lineages contained signature variants of a single PANGO lineage. This also indicates that most of the cryptic lineages were associated with the circulating strains in the population, as it incorporates the mutations from the known parent strains, with some additional new variants. Additional information on cryptic lineages containing signature variants specific to multiple PANGO lineages can be found in Supplementary section 3.

Filtering Results from Houston Wastewater

Crykey applies both within-sample and cross-sample filtering on the cryptic candidates. We filtered each candidate mutation combination by keeping the ones with both the number of supported reads above a minimum threshold (default: 5), and allele frequency above a minimum threshold (default: 0.01). The cross-sample filtering is based on the minimum number of samples supporting the candidate (default: 2).

In the next step, the remaining candidates are queried against the SNP database, and Crykey outputs a complementary report on whether or not the SNP combinations found in the sample are truly novel, which means no sequences in the database supports the combination, or if the SNP combination is rarely seen in the database. If the SNP combinations can be found in the database, Crykey reports the number of sequences found containing such combinations in each of the lineages, as well as the total number of sequences in those

lineages. We further removed cryptic candidates with prevalence rate above 0.0001 in the GISAID database, and performed our analysis based on the filtered results.

In the final step, we excluded the cryptic candidates that contain variants located between reference genome positions 1 to 55 and 29,804 to 29,903. In addition, we masked cryptic candidates that contain variants located on 25 nucleotide positions between 56 and 29,804 based on suggestions from previous studies, as those locations are highly homoplasic and variants are likely to be recurrent artifacts [39,40].

Searching Cryptic Lineages in Clinical Samples

The samples from Houston were downloaded from the NCBI SRA database under the BioProject PRJNA764181. We collected all 5060 SRA runs with sample collection dates between 12/06/2021 and 01/31/2022 without additional filtering. In addition, 8969 out-of-state SRA runs were sampled from the NCBI database. We first performed a standard read mapping process including quality control. The sequencing reads from the clinical samples were first filtered using fastp v0.23.2 with parameter "--cut_front, --cut_tail, --cut_window_size, 4, --cut_mean_quality 25, --qualified_quality_phred 25 --unqualified_percent_limit 40 --n_base_limit 5 --length_required 15 --low_complexity_filter --complexity_threshold 30" to remove low quality bases and low quality reads[41]. 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[41,42]. The alignment files were sorted and indexed with samtools v1.14. Variant 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"[43]. Consensus genomes were generated and PANGO calling was done using pangolin v4.2 with default parameters[44].

After read mapping, the BAM files are collected for searching cryptic lineages detected in wastewater samples. 20 wastewater cryptic lineages detected during the 8 week sampling period were selected for testing. 10 of the 20 wastewater cryptic lineages are cryptic lineages that occurred exactly 2 out of 8 weeks, representing the cryptic lineages with a short burst detection pattern, and we chose those which had the most occurrence in terms of wastewater treatment plants where the cryptic lineage had been detected. The rest of the 10 wastewater cryptic lineages are cryptic lineages that occurred most of the weeks during the sampling period, ranging from 4 to 8 weeks occurrence, representing the cryptic lineages with a long lasting detection pattern.

We examined the alignments in the clinical samples, and counted the total number of reads spanning the regions that the cryptic lineages located, and counted the number of reads supporting all mutations from the cryptic lineages 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 cryptic lineage was calculated as the number of cryptic lineage supporting reads over the number of total reads in the region.

During the analysis, we further filtered the results, and samples with cryptic lineage supporting read count less than 5 are considered as cryptic lineage 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[45,46]. Samples with cryptic lineage reads containing strain bias scores greater or equal to 1 or p-value of the Fisher's exact test less than 0.05 are considered as cryptic lineage absent.

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 in 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. were supported in part by the National Science Foundation (CBET 2029025), and seed funds from Rice University.

Figures

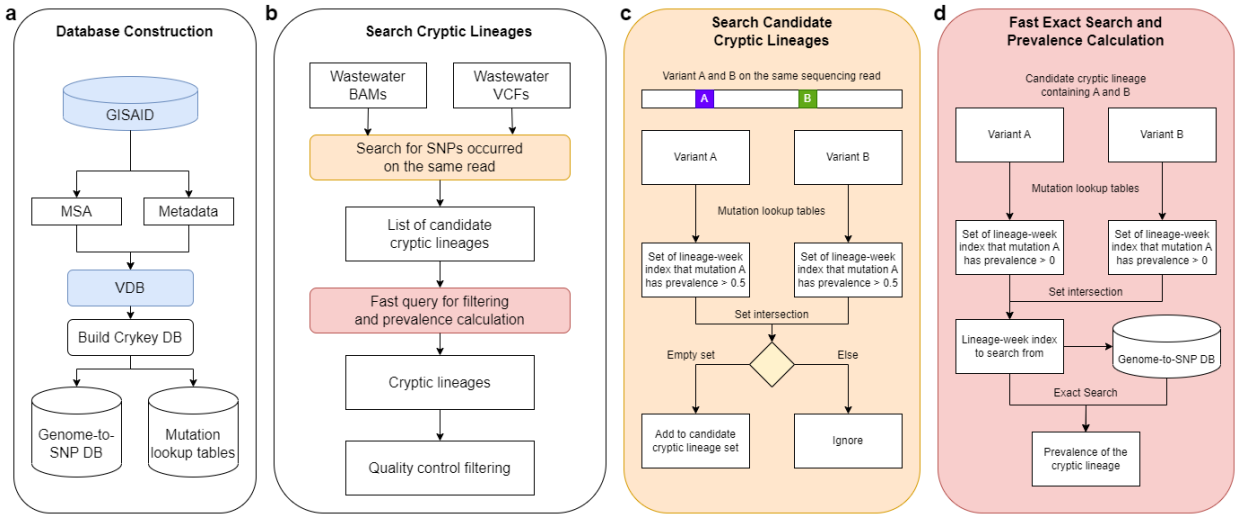

Figure 1. Workflow and algorithms of Crykey. a) Crykey constructs a genome-to-SNP database, as well as a set of mutation lookup tables using GISAID provided multiple sequence alignments and metadata. b) Crykey searches for multiple variants located on the same read, and uses the mutation lookup table to identify whether the combination of the variants is a candidate cryptic lineage. Each candidate cryptic lineage is queried against the genome-to-SNP database to calculate its prevalence rate. c) Algorithm to search candidate cryptic lineages with an example of a read containing two variants A and B at the same time. d) Algorithm for the exact search once the candidate cryptic lineages are generated. The example shows a candidate cryptic lineage containing variant A and B.

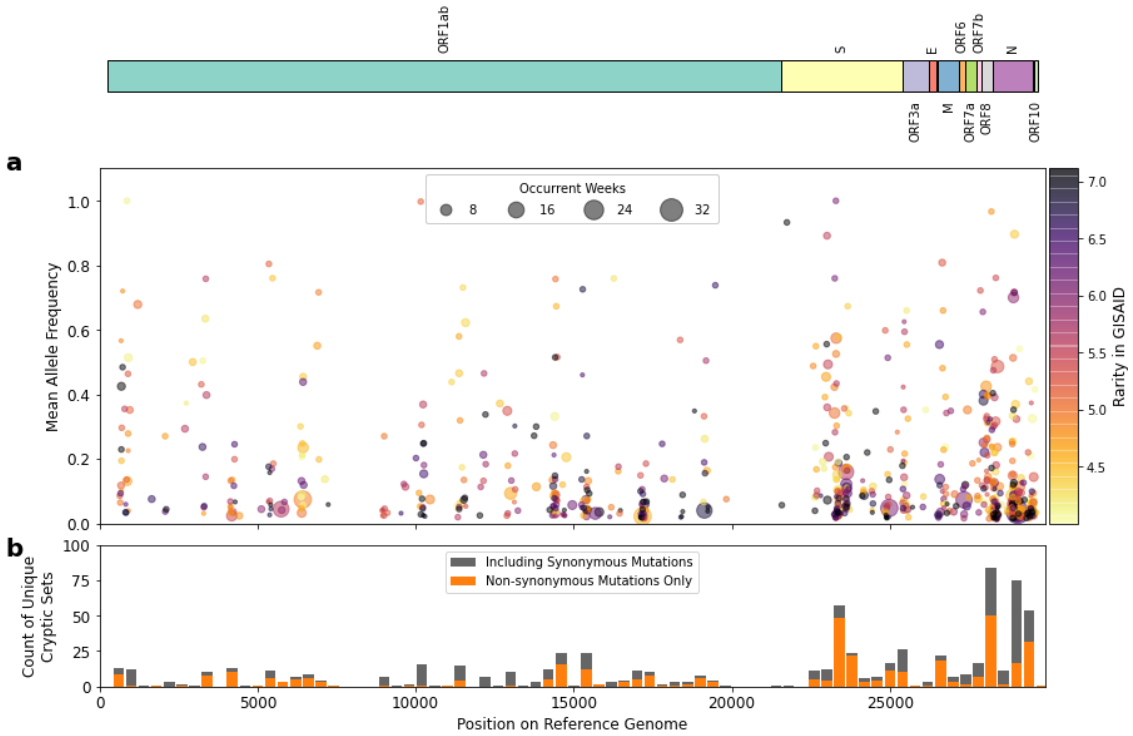

Figure 2. Distribution of cryptic lineages found in Houston wastewater on SARS-CoV-2 genome. In both a) and b), the locations of cryptic lineages (cryptic mutation sets) found in Houston wastewater samples on the SARS-CoV-2 reference genome are shown on the x-axis, with gene annotation on the top of the figures. In panel a), each cryptic lineage is represent by a colored dot, the y-axis indicates its mean allele frequency in the wastewater sample, and the color indicates its rarity , defined as -log10 * (n+1), where n is the number of genomes supporting the cryptic lineage in the GISAID EpiCoV database. Darker color suggests that the cryptic lineage is rarely reported or even never been reported. The size of the dot shows the number of weeks the cryptic lineage has been detected. Larger dots indicate the cryptic lineage persists longer in the community. Panel b) is a histogram showing the count of cryptic lineages found in different regions of the reference genome, by dividing the genome into bins of 400 bp. The cryptic lineages containing exclusively non-synonymous mutations are marked in orange, and the cryptic sets containing at least one synonymous mutation are marked in gray. Higher bars indicating the more cryptic mutations are found in the associated region. The higher ratio between orange and gray bars indicating the associated region tends to have more cryptic lineages only containing non-synonymous mutations.

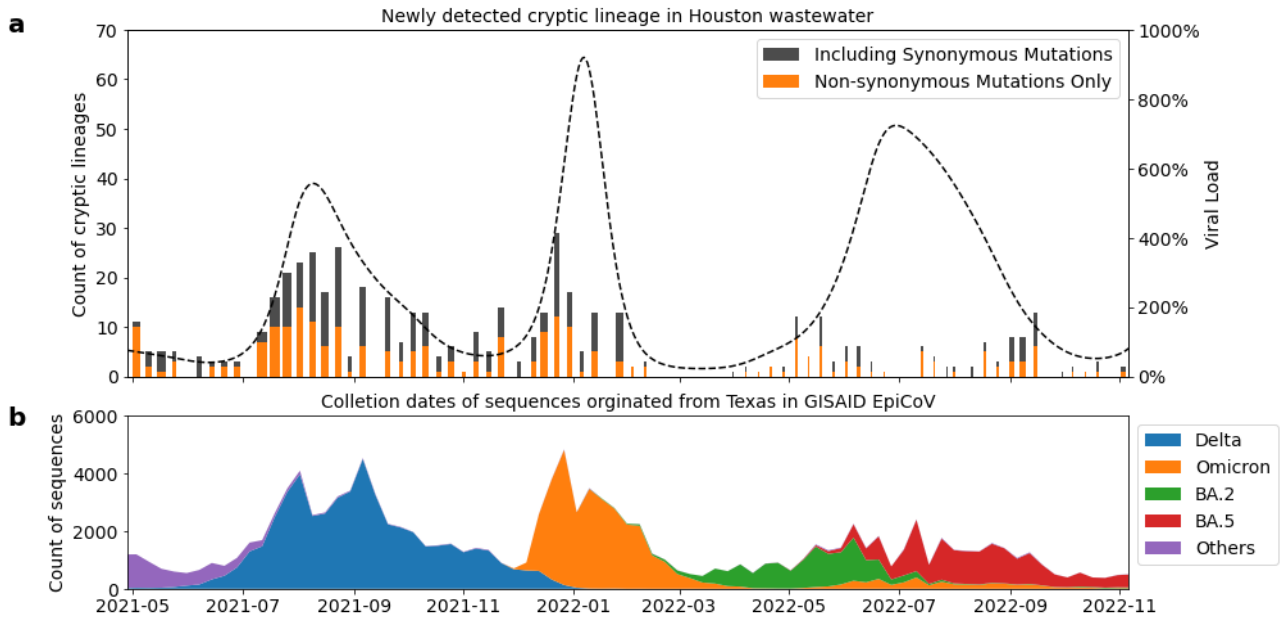

Figure 3. Crykey cryptic mutation sets and associated viral load in Houston wastewater. In both panel a) and b), the x-axis shows dates from May, 2021 to November 2022. Panel a) shows the number of cryptic lineages that are newly detected in Houston wastewater during each sampling cycle (week) on the left y-axis as bars. The Crykey cryptic lineages containing exclusively non-synonymous mutations are marked in orange, and the cryptic lineages containing at least one synonymous mutation are marked in gray. Higher bar indicates that more cryptic lineages are being detected for the first time during a week in Houston wastewater samples. The width of the bin indicates normalized sample qualities derived from breadth of the genome coverage. Wider bin suggests better wastewater sequencing quality, with a higher proportion of the genome being covered after read alignment. The corresponding normalized viral load in the wastewater samples is shown as a dotted line with the values on the right y-axis, and the viral load is normalized based on the viral load from samples collected on July 6, 2020 in Houston. Panel b) shows the weekly number of SARS-CoV-2 sequences in the GISAID EpiCoV database that are originating from Texas, USA. The count of sequences is calculated based on sample collection date. The sequences are binned and colored based on their PANGO lineage assignments.

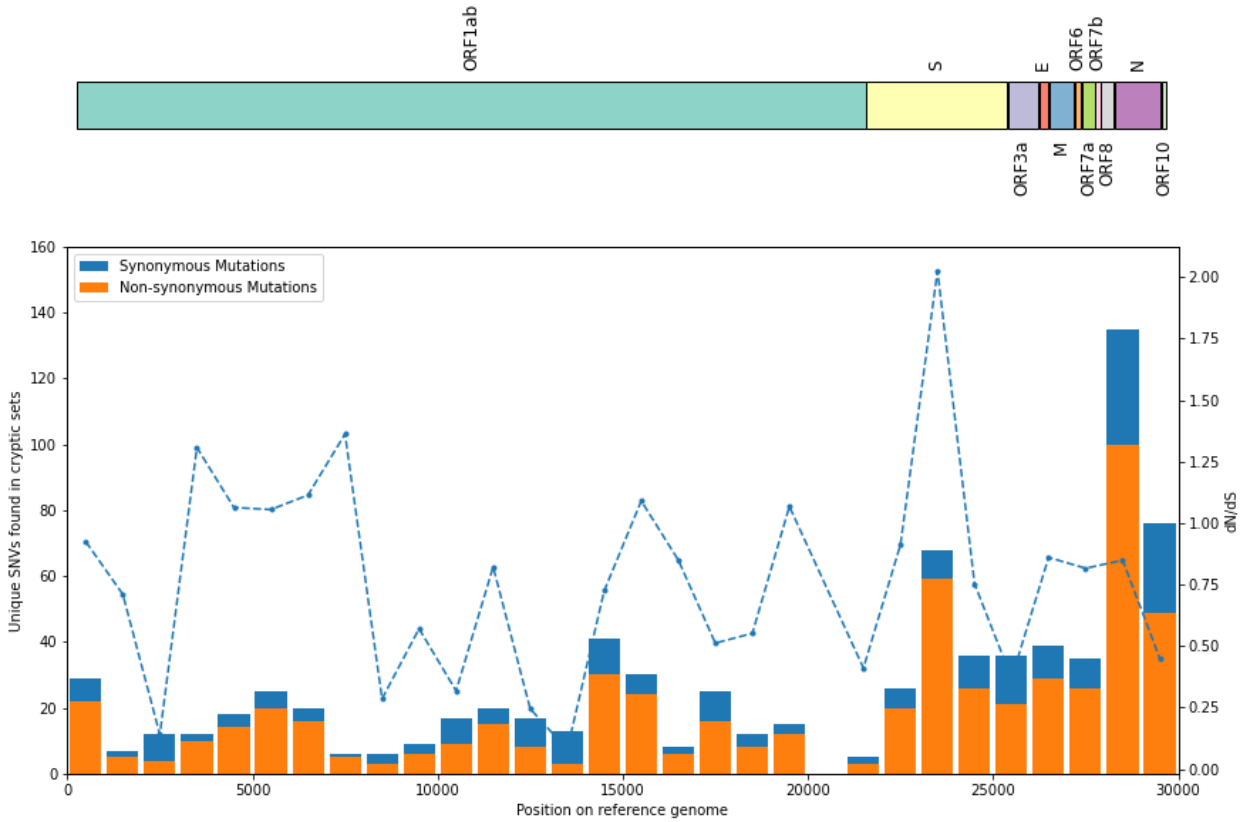

Figure 4. dN/dS ratio of unique SNVs found in cryptic lineage. The SARS-CoV-2 reference genome is partitioned into bins with length of 1000 bp. The figure shows the count of unique SNVs in the cryptic mutation sets detected at each bin, as well as the corresponding dN/dS ratio. The non-synonymous SNVs are marked in orange, and the synonymous SNVs are marked in blue.

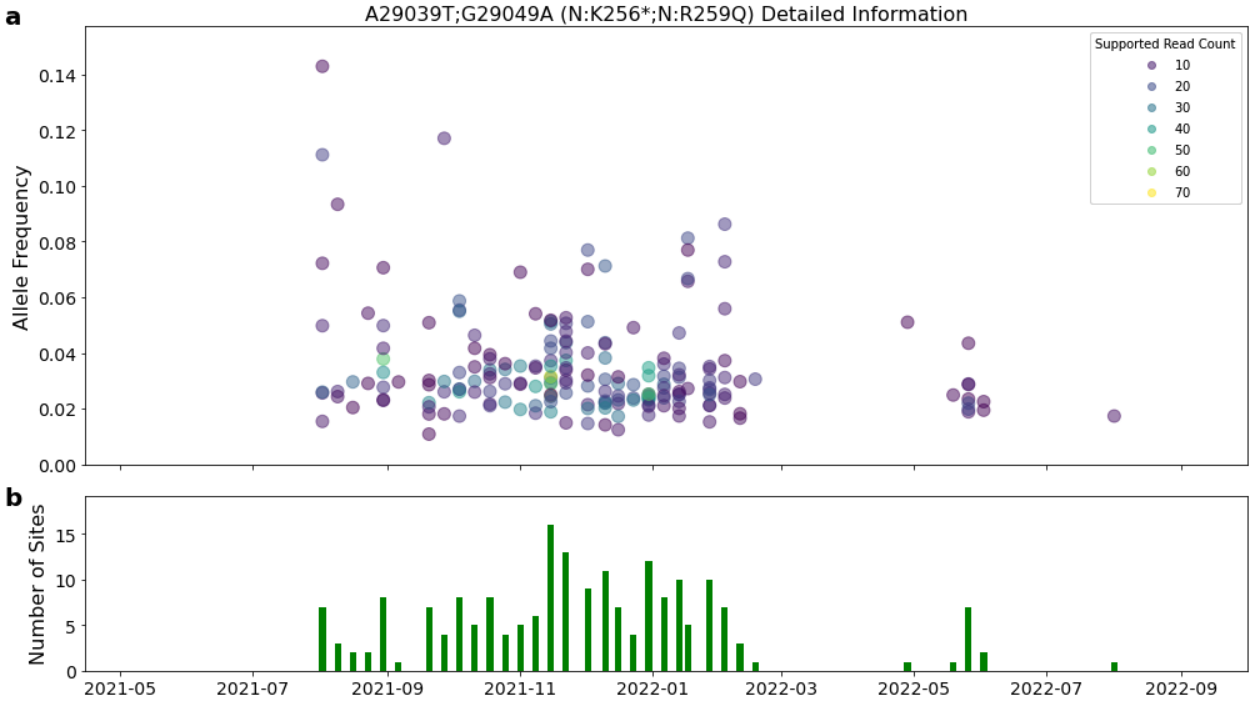

Figure 5. A long lasting cryptic lineage involving A29039T and G29049A mutations. The figures show the most long lasting cryptic lineage with the co-occurrence of A29039T and G29049A mutation. The shared x-axis shows the date of associated samples being collected. a) The y-axis shows the mean allele frequency of the SNVs. The count of supported reads containing both SNVs at the same time are shown in different colors. b) The y-axis shows the number of samples from different wastewater treatment plants containing this cryptic lineage on the specific date.

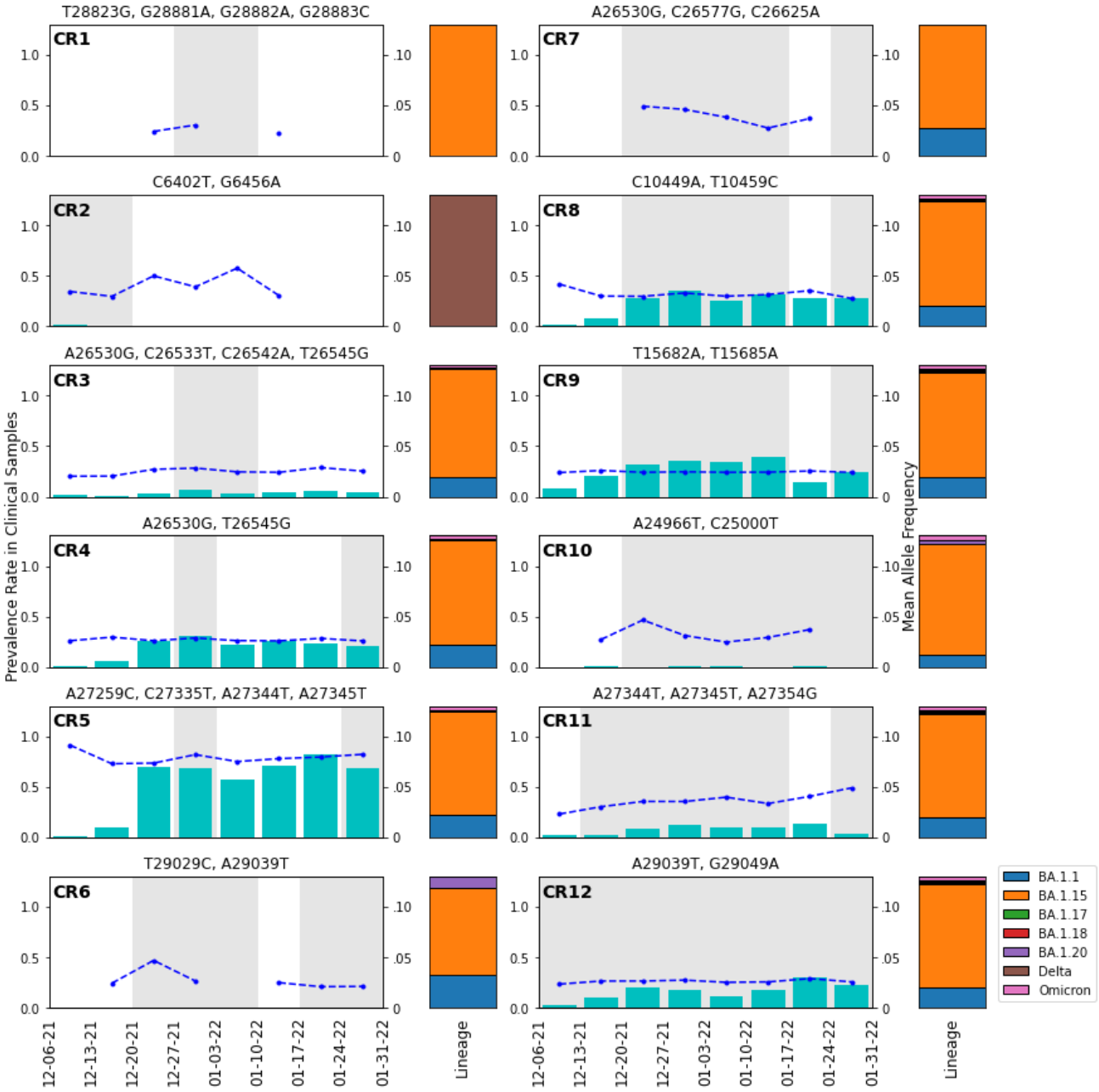

Figure 6. Cryptic lineages detected in Houston clinical samples. 12 sets of cryptic lineages (CR1 - CR12) appeared in Houston wastewater from 12/06/2021 to 01/31/2022 and their detection in clinical samples collected during the same 8 week period from the Greater Houston region. The mutation combinations for each cryptic lineage are labeled on top of each subfigure, and bar height (left y-axis) indicates the prevalence rate of cryptic lineage in the clinical samples each week, and the dotted blue lines (right y-axis) indicate the mean allele frequency of the cryptic lineage in clinical samples. For each cryptic lineage, the shaded areas indicate whether the cryptic lineage is detected in Houston wastewater treatment plants on a specific week, and the stacked bars with different colors show the distribution of the PANGO lineages of the consensus genomes of the cryptic supported clinical samples. All PANGO lineages of Delta are combined. All PANGO lineages of Omicron other than BA.1.1, BA.1.15, BA.1.17, BA.1.18, BA.1.20 are combined and denoted as Omicron.

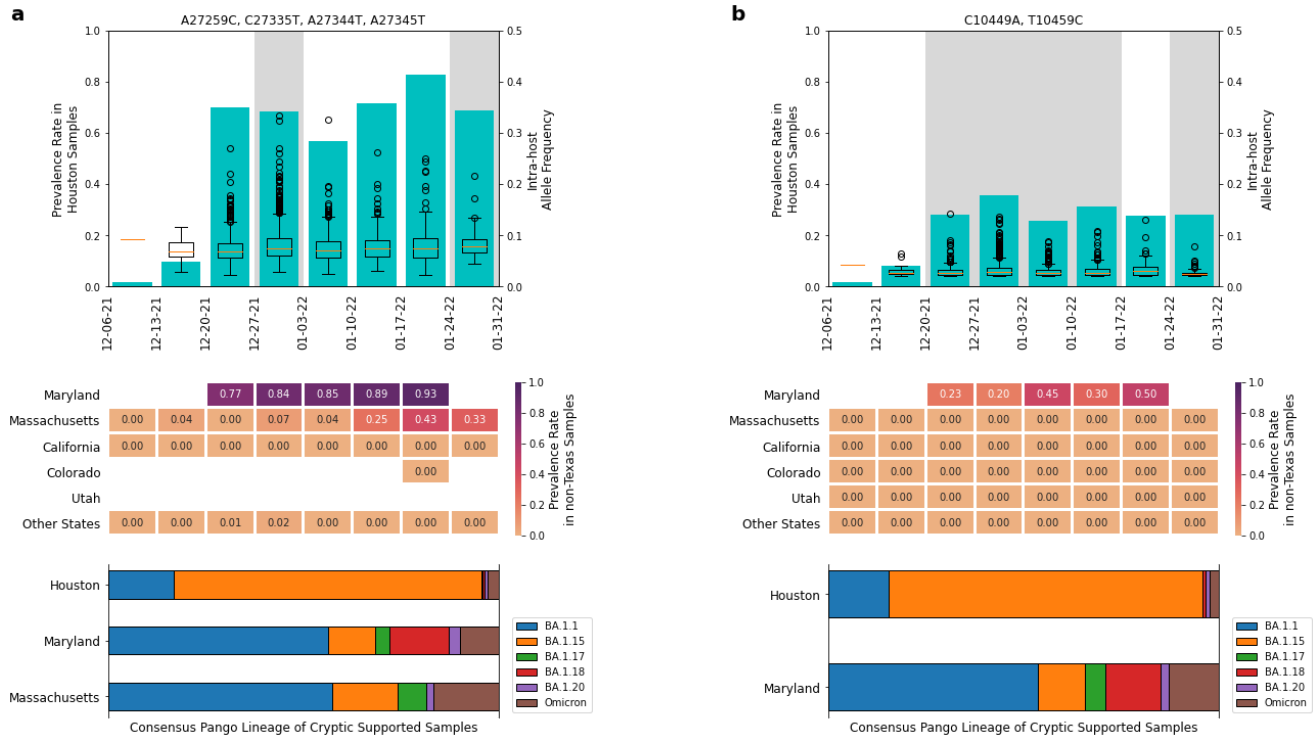

Figure 7. Cryptic lineage found in clinical samples in Houston and other US cities. The figure shows detailed information of a) cryptic lineage A27259C-C27335T-A27344T-A27345T and b) cryptic lineage C10449A-T10459C, which are two of the cryptic lineages detected in Houston wastewater that are supported by clinical samples from Houston as well as from other US cities. For each of the cryptic lineages, the top figure shows the prevalence rate in Houston clinical samples in bars and the intra-host allele frequency of the cryptic lineage in cryptic supported clinical samples in the box plot. The center heatmap shows the prevalence rate in other US regions, where white cells indicate no data. The bottom figure shows the distribution of the PANGO lineages of the consensus genomes of the cryptic supported clinical samples in Houston and other US regions. All PANGO lineages of Omicron other than BA.1.1, BA.1.15, BA.1.17, BA.1.18, BA.1.20 are combined and denoted as Omicron.

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. Smyth, D. S. *et al.* Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. *Nat. Commun.* **13**, 1–9 (2022).
20. 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.
21. 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).
22. Gregory, D. A. *et al.* Genetic Diversity and Evolutionary Convergence of Cryptic SARS-CoV-2 Lineages Detected Via Wastewater Sequencing. *medRxiv* doi:10.1101/2022.06.03.22275961.
23. Gregory, D. A. *et al.* Genetic diversity and evolutionary convergence of cryptic SARS- CoV-2 lineages detected via wastewater sequencing. *PLoS Pathog.* **18**, (2022).
24. Karthikeyan, S. *et al.* Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* **609**, 101–108 (2022).
25. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nature Microbiology* **5**, 1403–1407 (2020).
26. Markov, P. V. *et al.* The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* 1–19 (2023).
27. 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).

28. 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).
29. 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).
30. 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.
31. Factors influencing SARS-CoV-2 RNA concentrations in wastewater up to the sampling stage: A systematic review. *Sci. Total Environ.* **820**, 153290 (2022).
32. 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.
33. 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).
34. Wolfe, M. K. *et al.* Wastewater-Based Detection of Two Influenza Outbreaks. *Environmental Science & Technology Letters* (2022) doi:10.1021/acs.estlett.2c00350.
35. Ortiz, A. T. *et al.* Within-host diversity improves phylogenetic and transmission reconstruction of SARS-CoV-2 outbreaks. *bioRxiv* 2022.06.07.495142 (2022) doi:10.1101/2022.06.07.495142.
36. Wang, Y. *et al.* Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. *Genome Med.* **13**, 1–13 (2021).
37. Gonzalez-Reiche, A. S. *et al.* Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. *Nat. Commun.* **14**, 1–13 (2023).
38. 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).
39. Sapoval, N. *et al.* SARS-CoV-2 genomic diversity and the implications for qRT-PCR diagnostics and transmission. *Genome Res.* **31**, 635–644 (2021).
40. De Maio, N. *et al.* Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* **13**, evab087 (2021).
41. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890

(2018).

42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. 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).
44. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* **7**, veab064 (2021).
45. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**, 1–11 (2012).
46. 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).
47. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science* **372**, (2021).
48. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. (2021) doi:10.7554/eLife.66857.
49. Shen, Z. *et al.* Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clin. Infect. Dis.* **71**, 713–720 (2020).
50. Berrio, A., Gartner, V. & Wray, G. A. Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. *PeerJ* **8**, (2020).
51. Hulswit, R. J. G., de Haan, C. A. M. & Bosch, B.-J. Coronavirus Spike Protein and Tropism Changes. *Adv. Virus Res.* **96**, 29 (2016).
52. 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).
53. 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).