



Viruses and Viral Diseases

Revealing patterns of SARS-CoV-2 variant emergence and evolution using RBD amplicon sequencing of wastewater



Xingwen Chen^{a,b}, John Balliew^c, Cici X. Bauer^{a,b}, Jennifer Deegan^{a,b}, Anna Gitter^{a,b}, Blake M. Hanson^{a,b}, Anthony W. Maresso^d, Michael J. Tisza^e, Catherine L. Troisi^{a,b}, Janelle Rios^{a,b}, Kristina D. Mena^{a,b}, Eric Boerwinkle^{a,b,*}, Fuqing Wu^{a,b,*}

^a School of Public Health, University of Texas Health Science Center at Houston, TX, USA

^b Texas Epidemic Public Health Institute (TEPHI), UTHealth Houston, Houston, TX, USA

^c El Paso Water Utility, El Paso, TX, USA

^d TAILOR Labs, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA

^e The Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA

ARTICLE INFO

Article history:

Accepted 20 September 2024

Available online 26 September 2024

Keywords:

Receptor-binding domain
Amplicon sequencing
Wastewater surveillance
SARS-CoV-2
Variant emergence
Viral evolution

SUMMARY

Objectives: Rapid evolution of SARS-CoV-2 has resulted in the emergence of numerous variants, posing significant challenges to public health surveillance. Clinical genome sequencing, while valuable, has limitations in capturing the full epidemiological dynamics of circulating variants in the general population. This study aimed to monitor the SARS-CoV-2 variant community dynamics and evolution using receptor-binding domain (RBD) amplicon sequencing of wastewater samples.

Methods: We sequenced wastewater from El Paso, Texas, over 17 months, compared the sequencing data with clinical genome data, and performed biodiversity analysis to reveal SARS-CoV-2 variant dynamics and evolution.

Results: We identified 91 variants and observed waves of dominant variants transitioning from BA.2 to BA.2.12.1, BA.4&5, BQ.1, and XBB.1.5. Comparison with clinical genome sequencing data revealed earlier detection of variants and identification of unreported outbreaks. Our results also showed strong consistency with clinical data for dominant variants at the local, state, and national levels. Alpha diversity analyses revealed significant seasonal variations, with the highest diversity observed in winter. By segmenting the outbreak into lag, growth, stationary, and decline phases, we found higher variant diversity during the lag phase, likely due to lower inter-variant competition preceding outbreak growth.

Conclusions: Our findings underscore the importance of low transmission periods in facilitating rapid mutation and variant evolution. Our approach, integrating RBD amplicon sequencing with wastewater surveillance, demonstrates effectiveness in tracking viral evolution and understanding variant emergence, thus enhancing public health preparedness.

© 2024 The Author(s). Published by Elsevier Ltd on behalf of The British Infection Association. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Introduction

SARS-CoV-2, the virus responsible for COVID-19, has undergone significant evolution since its emergence, leading to waves of outbreaks by multiple variants such as Alpha, Beta, Gamma, Delta, and Omicron. Each of these variants exhibits distinct characteristics in terms of transmissibility, pathogenicity, and ability to evade immune responses. A key driver of this evolution is the mutation of the spike

protein-encoding gene, particularly in the receptor-binding domain (RBD), which is crucial for the virus's ability to bind to host cell receptors and initiate infection.^{1–3} Mutations in the RBD can enhance the virus's ability to spread and escape immune detection,⁴ making it a critical focus for variant tracking and public health monitoring. For example, the Delta variant harbors the L452R and T478K mutations which increases binding affinity to the ACE2 receptor on host cells and contributes to reduced vaccine efficacy.^{5–9} Understanding and tracking these variants are crucial for managing ongoing and future outbreaks.

Traditional surveillance methods, such as genome sequencing of clinically ascertained samples, have been instrumental in identifying

* Corresponding author at: School of Public Health, University of Texas Health Science Center at Houston, TX, USA.

E-mail address: fuqing.wu@uth.tmc.edu (F. Wu).

and tracking SARS-CoV-2 variants. However, these methods are often time-consuming, expensive, and resource-intensive.^{10–12} Moreover, clinical testing primarily captures data from symptomatic individuals, potentially missing large portions of the infected population who are asymptomatic or pre-symptomatic.^{13–16} In contrast, wastewater surveillance has emerged as a promising approach to monitor community-level prevalence and diversity of SARS-CoV-2. By analyzing wastewater samples, one can detect viral RNA shed by infected individuals, regardless of their symptom status.^{17–19} Wastewater surveillance offers several advantages, including non-invasiveness, cost-effectiveness, and the ability to capture data from a broad segment of the population. This method has now been widely implemented globally, providing early warnings of outbreaks and offering a comprehensive view of community-level viral transmission.

Wastewater sequencing is a crucial tool for early detection of viral variants. By collecting and analyzing the nucleic acids in wastewater using next-generation sequencing, researchers can identify mutations and new variants weeks before they appear in clinical samples.^{20–22} This method allows continuous tracking of variant dynamics and provides a more complete picture of the viral landscape.^{23–26} However, the complex nature of wastewater, containing a mixture of genetic material from various sources, complicates data analysis and interpretation, posing challenges in genome assembly, variant assignment/classification, and sequencing accuracy.^{27–32} Typically, wastewater sequencing relies on tiled amplicon amplification to recover the whole genome due to low concentrations. A simpler approach targets the spike genes, particularly the RBD, with specific target areas varying among studies.^{20,21,33–39} While existing research has primarily focused on the detection and surveillance of variants, there has been limited exploration into the epidemiological dynamics and patterns of variant emergence and evolution within community wastewater.

In this study, we targeted the receptor-binding domain region and developed a streamlined bioinformatic pipeline to track the variant dynamics of SARS-CoV-2 over 17 months using wastewater samples from El Paso, a border city in west Texas, US, and Ciudad Juárez, Mexico. We compared our wastewater data with clinical genome sequencing results at local, state, and national levels to assess consistency and reliability. Additionally, we conducted epidemiological dynamic analyses to uncover patterns of variant evolution and emergence. By integrating RBD amplicon sequencing with wastewater surveillance, this research offers an innovative approach to understanding variant emergence and evolution, enriching the surveillance toolkit to anticipate and mitigate emerging threats in the post-pandemic era.

Materials and methods

Wastewater sampling

Composite wastewater samples (24-hr) were collected weekly from March 14th, 2022, to August 8th, 2023, from three wastewater treatment plants (WWTPs) in the City of El Paso, Texas, USA. El Paso is a major municipal city along the U.S.-Mexico border in western Texas, neighboring Ciudad Juárez, Mexico, and forming a large binational metropolitan area. The Franklin Mountains run through the city, creating distinct urban regions. The three WWTPs, referred to as A, B, and C for anonymity, serve these regions: A serves the northern region (~98,754 residents), B covers the central area (~125,462 residents), and C manages wastewater from the northwest sector (~128,003 residents). The daily total wastewater flow volume was measured and provided by El Paso Water. Samples were shipped overnight in ice boxes to the laboratory and processed on the day of receipt. The rest of the raw samples were stored at -80 °C.

Wastewater sample processing, RNA extraction, and RT-qPCR

Samples were processed based on previous methods.^{40–42} Briefly, 50 mL wastewater samples were first centrifuged at 3000 g for 10 min at 4 °C. The supernatant was collected and further filtered with MilliporeSigma™ Steriflip™ Sterile Disposable Vacuum Filter Units. Then, the 15 mL filtrates were concentrated to ~200 µL with MilliporeSigma™ Amicon™ Ultra-15 Centrifugal Filter Units. The enriched samples were subject to RNA extraction with QIAamp Viral RNA Mini Kit according to the manufacturer's protocol. RNA was eluted with 100 µL nuclease-free water and stored at -20 °C.

SARS-CoV-2 was tested and quantified using the US CDC N1 real-time PCR assay (Biorad, CFX96 C1000 Touch) using the following program: 50 °C 10 min for reverse transcription, 95 °C 20 s for RT inactivation and initial denaturation, and 48 cycles of denature (95 °C, 1 s) and anneal/extend (60 °C, 30 s). At least two negative controls were included in every PCR run. Pepper mild mottle virus (PMMoV) was also measured as an internal reference for sample processing and variations in wastewater flow and/or fecal materials. Two technical replicates were performed for all RT-qPCR reactions. SARS-CoV-2 RNA concentrations were adjusted by the corresponding PMMoV concentrations.^{41,43} Briefly, the median of PMMoV Ct values was used as the baseline for normalization. Assuming that environmental changes affect the Ct values of both PMMoV and SARS-CoV-2 equally in each sample, we normalize the SARS-CoV-2 Ct values by adjusting for the deviation of PMMoV Ct from its median. This normalized SARS-CoV-2 Ct is then converted to viral load using a standard curve function (slope: -3.1629, intercept: 40.678).⁴³ To enable data comparison across sewersheds, we used the total viral load per capita per day using the following formula:

Total viral load per day

$$= \text{Measured viral concentration} \times \text{Dilution factor} \times \text{Influent flow}$$

$$\text{Viral load per capita per day} = \frac{\text{Total viral load per day}}{\text{Population covered by the WWTP}} \times 100,000$$

The formula incorporates the measured SARS-CoV-2 concentration (genome copies per µL), dilution factor, wastewater influent flow volume (mL), and the population size served by each WWTP. The “dilution factor” accounts for the sample concentrating process where 15 mL of raw wastewater samples was processed to yield 100 µL of RNA, which is 6.7 (100/15).

Preparation of synthetic SARS-CoV-2 variant communities

The SARS-CoV-2 Wuhan-Hu strain and Omicron BA.4&5 genomes were purchased from TWIST Bioscience. Original stocks were diluted one thousand-fold before mixing with four different ratios: 10%:90%, 25%:75%, 50%:50%, and 90%:10%. 10 µL of each mixture sample were then subjected to RNA extraction with QIAamp Viral RNA Mini Kit.

RBD amplicon amplification of wastewater samples

Primers for amplifying the receptor-binding domain (RBD) of the SARS-CoV-2 spike glycoprotein were provided in Table S1 and adapted from a previous study.²⁰ This amplicon amplifies 168 amino acids in spike gene from position 337 Pro to 504 Gly, covering almost the entire receptor-binding motif of SARS-CoV-2 (spike gene from position 437 Asn to 508 Tyr) and most receptor-binding domain (spike gene from position 319 Arg to 541 Phe). Briefly, wastewater samples and positive control RNA were transcribed into cDNA using SuperScript™ IV Reverse Transcriptase (ThermoFisher Scientific, USA). The RBD amplicon was then amplified from the cDNA using primers S008 and S012 in a first PCR with Q5® High-Fidelity 2X

Master Mix (New England Biolabs, USA), under the following conditions: polymerase activation at 95 °C for 2 min, followed by 30 cycles of denaturation at 95 °C for 1 s, annealing at 72 °C for 30 s, and extension at 60 °C for 30 s. A second PCR was performed to add Illumina overhangs to the RBD amplicon using primers S009 and S013, with the same conditions but for 40 cycles. The PCR product was purified using AMPure XP beads and eluted with 30 µL of nuclease-free water. For testing the protocol, we used synthetic SARS-CoV-2 Wuhan-Hu strain and Omicron BA.4&5 genomes from TWIST Bioscience. They were diluted 1,000-fold and mixed in water in ratios of 10%:90%, 25%:75%, 50%:50%, and 90%:10% to create positive control mock communities. We used 10 µL of each mixture for RNA extraction, amplicon amplification, library preparation, and sequencing. Note that the mock communities were not spiked into raw wastewater to avoid interference from existing SARS-CoV-2, and thus potential bias from wastewater sample processing was not estimated in this study.

Library preparation and sequencing data analysis

The purified PCR amplicons were dual-indexed using NEBNext Multiplex Oligos (New England Biolabs, USA) for Illumina and further purified using AMPure XP beads. Libraries were pooled based on their concentrations and sequenced on a 300 paired-end Miseq run at the ATGC Facility at MD Anderson Cancer Center. The raw sequencing data in fastq format were de-multiplexed. DADA2 package (version 4.3.0)⁴⁴ was used to filter and trim sequences, merge paired ends, and remove chimeras in Rstudio version 4.3.2. The unique amplicon sequence variants (ASVs) were aligned and identified using the BLAST online tool selecting the Betacoronavirus database provided by the BLAST tool for dataset option based on GenBank nucleotide sequence database (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>).⁴⁵ The database was updated on 1/31/2024 for the analysis in this study. For each ASV, the top 10 BLAST results containing the Pango Lineage name with the highest alignment scores (ranging from 99% to 100%) were downloaded, and the most frequent variant was identified as the variant for that sequence. The relative abundance of each variant was calculated by the ratio of the count of each variant to the total sequence count. Variants with relative abundance higher than 10% on at least one sample were designated as major variants. The remaining variants were grouped as 'others' for plotting and visualization purposes (Table S1). The diversity analysis was carried out with the 'phyloseq' (version 3.19) package in R. The alpha diversity was measured with four indexes: the Chao1, Shannon, Observed, and Inverse Simpson indexes. Beta diversity analysis was performed using the non-metric multidimensional scaling (NMDS) method with Bray-Curtis distances to assess the dissimilarity in variant communities among the three sewersheds.

COVID-19 cases and genome data

The reported COVID-19 case data by lab-testing diagnostics was downloaded from healthdata.gov, and the 7-day rolling average was used for further analysis and data visualization. The Texas-specific data was extracted according to the state name label. SARS-CoV-2 genome data for the study period was obtained from the GISAID EpiCoV™ database using the GISARD (version 0.9.9) package in R.⁴⁶ The variants were assigned by Pango v.4.3 by GISAID. Only records with complete genomes were downloaded. The number of each variant per week was counted, and the relative abundance was calculated accordingly. Note that the wastewater data in this study is confined to El Paso city and thus does not fully reflect the entire population of Texas. Due to insufficient clinical data in El Paso, we used the Texas-wide clinical genome data to compare the temporal trends of the variants. Associations between variant abundance in wastewater and variant abundance in clinical genome sequencing

data were analyzed, with the method described in the [Supplemental Materials](#).

Results

RBD amplicon sequencing of wastewater reveals temporal dynamics of SARS-CoV-2 variants circulating in the city

To validate the RBD amplicon sequencing approach for variant identification, we utilized synthetic mock variant communities. Four mock communities were prepared by mixing the original SARS-CoV-2 Wuhan-Hu strain (Wild Type) and the BA.4&5 strain at ratios of 10%:90%, 25%:75%, 50%:50%, and 90%:10%. Each mock community underwent RNA extraction, reverse transcription, amplicon amplification, library preparation, and sequencing, followed by bioinformatic analysis. As a negative control, nuclease-free water was processed identically, except for RNA extraction. The sequencing results showed that nearly all reads (98.5%~100%) were accurately identified as either Wild Type or Omicron BA.4&5 strains across all four mock communities. The estimated relative abundances for each variant in the mock communities were 6.7%:92.3%, 21.9%:76.6%, 51.3%:47.5%, and 93.1%:6.9% (Fig. 1B), closely matching the original mixing ratios with a deviation range of -3.1% to 3.1%. These results demonstrate that the RBD amplicon sequencing method is effective for identifying variants and estimating the composition and structure of SARS-CoV-2 variants in a sample.

Next, we applied the RBD amplicon sequencing approach to weekly wastewater samples collected from the three wastewater treatment plants. All samples tested positive for SARS-CoV-2. Consistent with clinically reported case data, viral load results revealed three major waves of infection in the city: from April 2022 to September 2022, from November 2022 to December 2022, and from February 2023 to the end of the experiment period (August 2023) (Fig. 1C and Fig. S1).

Across the 216 samples analyzed, 91 SARS-CoV-2 variants were identified, each with over 99% identity to the reported variant sequences. For visualization, we plotted the relative abundance of the variants of interest (VOIs) listed by WHO and those with an abundance of 10% or higher in at least one wastewater sample. All other variants were grouped under 'others'. Fig. 1D-F show the relative abundance of SARS-CoV-2 variants over the 17 months across the three sewersheds, highlighting the evolving composition of variant communities. Overall, each wastewater sample contained multiple variants, and we observed waves of dominant variants transitioning from BA.2 to BA.2.12.1, BA.4&5, BQ.1, and to XBB.1.5. While there were some differences, the overall variant profiles for each WWTP were similar, indicating that the spread and evolution of the virus followed a comparable trajectory across the three sewersheds in the city. These results demonstrate that RBD amplicon sequencing of wastewater samples can be used to monitor the temporal dynamics and community structure of SARS-CoV-2 variants circulating in the population.

Consistency and novel insights from RBD amplicon sequencing data compared to clinical genome sequencing data

We further validated the wastewater findings using reported clinical genome sequencing data from the USA, Texas, and El Paso. As shown in Fig. 1G, the overall trends between the wastewater RBD amplicon sequencing data and clinical genome sequence data in Texas are consistent, with multiple waves of major variants emerging sequentially. Moreover, the composition of variants in wastewater samples is more diverse, suggesting that some variants circulating in the city may go undetected in clinical testing. We further compared the temporal dynamics of major variants during the study period: BA.2, BA.2.12.1, BA.4&5, BQ.1, and XBB.1.5 (Fig. 2) in

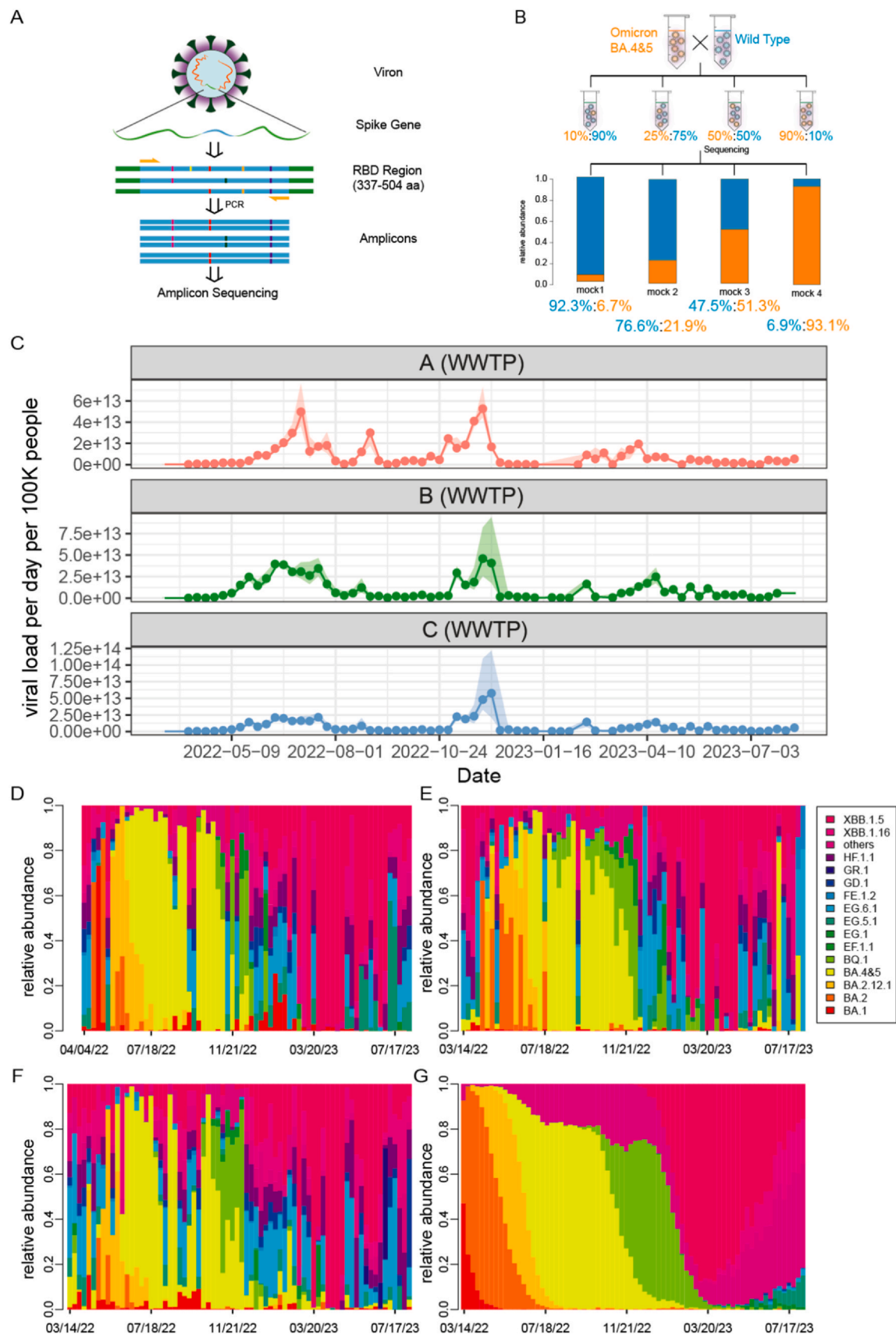


Fig. 1. RBD amplicon sequencing of wastewater reveals temporal dynamics of SARS-CoV-2 variants. (A) Diagram of the RBD amplicon sequencing procedure. We specifically amplify the spike gene 337–504 aa sequence in the RBD region of SARS-CoV-2 spike gene. (B) Validation of the RBD amplicon sequencing protocol using the wild-type SARS-CoV-2 strain and BA.4&5. The bar plot shows the sequencing results of these samples, demonstrating that the resultant ratios closely match the original mixing ratios. (C) Viral load of SARS-CoV-2 per day per 100k people in the three sewersheds in the City of El Paso, TX, from March 2022 to August 2023. Shaded area represents the range between technical replicate values for viral load. (D–F) Relative abundance of SARS-CoV-2 variants revealed by RBD amplicon sequencing of wastewater samples from WWTPs A (D), B (E), and C (F) in the City of El Paso. (G) Relative abundance of SARS-CoV-2 variants from genomic sequencing of clinical samples in Texas.

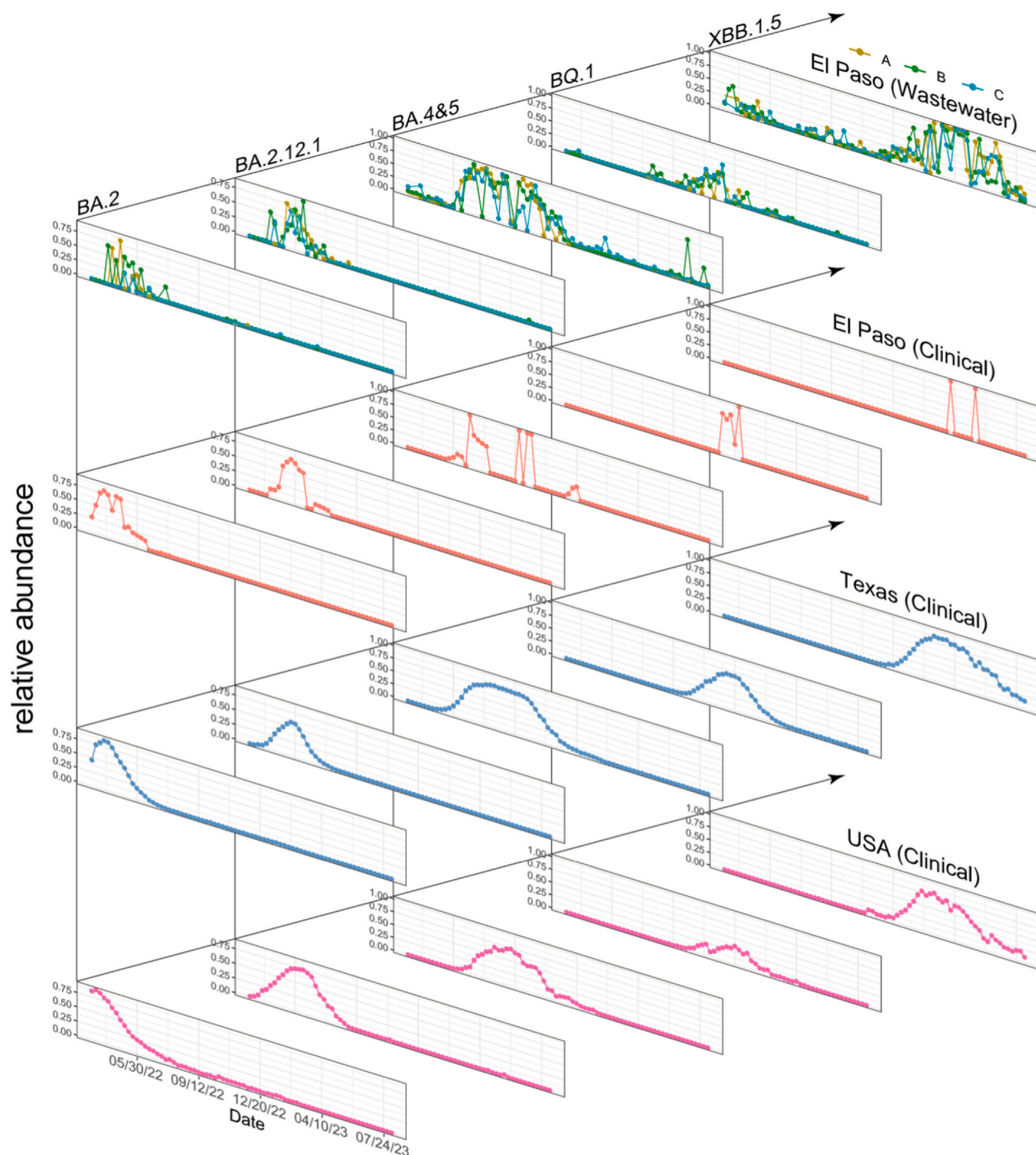


Fig. 2. Time-course of major SARS-CoV-2 variants identified from RBD amplicon sequencing of wastewater. This figure presents the temporal dynamics of five major variants identified from three WWTPs alongside clinical data from El Paso, Texas, and the USA. Rows depict the progression of variant waves over time, while columns allow for a comparative analysis between the wastewater treatment plant data and clinical data, highlighting their similarities and trends.

both wastewater and clinical sequencing data at the El Paso, Texas, and national levels. All five variant waves were consistently found in both datasets, showing similar transmission waves and relative abundance at the peaks. The correlations between wastewater and reported data are strongly positive for four of the five variants (Spearman's r ranging from 0.61 to 0.88, [Table S3](#)), except for BA.2, as

our experiment period only covered part of the wave. We also noticed some differences between Texas and national trends. For example, the BA.2 peak occurred later, and the BA.4&5 peak arrived earlier and lasted longer in Texas compared to the national data. In general, the correlations between wastewater data and Texas clinical data are higher than those between the wastewater and the USA

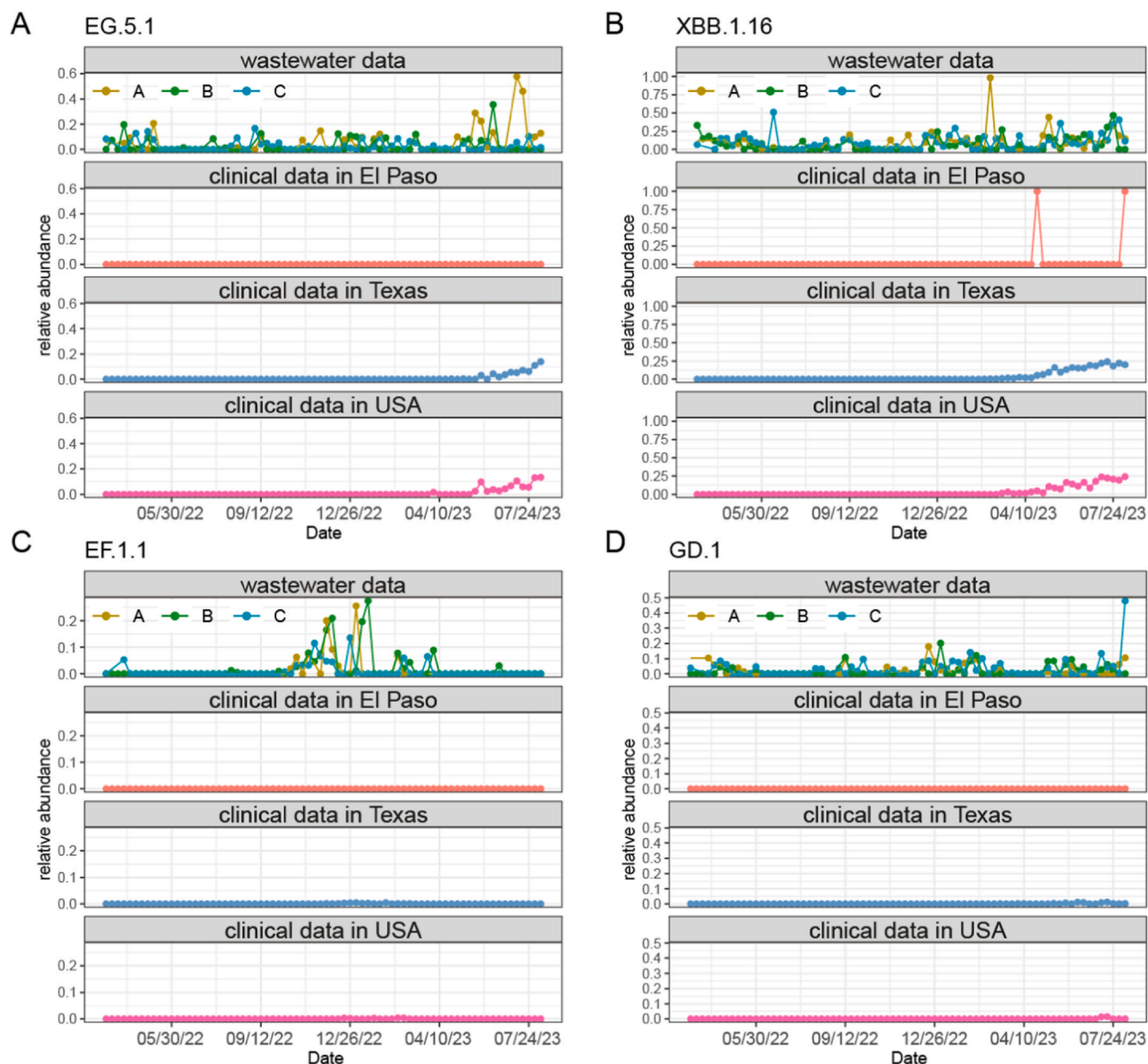


Fig. 3. Early detection and identification of unreported SARS-CoV-2 variants from RBD amplicon sequencing data. (A–B) The EG.5.1 and XBB.1.16 variants were detected circulating in the city earlier than their corresponding waves in clinical genome sequencing data from El Paso, Texas, and the USA. (C–D) The EF.1.1 and GD.1 variants were found in the city but were rarely reported in clinical genome sequencing data.

clinical data (Table S3). Therefore, wastewater data in El Paso closely matched the clinical sequencing data in Texas, more broadly.

In addition, we found that some variants circulated in the population earlier than their prevalence observed in clinical tests. For example, the XBB.1.5 wave began in December 2022, as shown in both clinical and wastewater data. However, this variant was detectable in wastewater samples before the major wave, with small peaks observed in March and September 2022 (Figs. 2 and S4). Similarly, variants XBB.1.16, EG.5.1, and EG.6.1 were detected in wastewater over 30 weeks before clinical identification, as they were present at the start of our study (Figs. 3A and B, and S5–S7, and Table S4). For BQ.1 and BA.4&5, wastewater detection preceded clinical records by seven weeks (Figs. 2, S2, and S3, and Table S4). These findings strengthen the existing idea that wastewater surveillance could provide early warnings to the community.^{28,47,48}

We also identified some variants that were not detected in the clinical data. Those variants, including EF.1.1, GD.1, EG.6.1, and FE.1.2, were barely noticeable in clinical records due to low abundance but were significant enough to exceed 10% in at least one wastewater sample (Fig. 3C and D, and Fig. S8). The EF.1.1 variant even had a small outbreak in all of the three sewersheds, with a relative abundance of up to 25% from October 2022 to February 2023. The EG.6.1, FE.1.2, and GD.1 variants were detected throughout the experiment period, but very few records were found in clinical data. This discrepancy likely occurs because these variants have a moderate transmission rate but do not cause severe symptoms, leading to fewer clinical tests.

In summary, the results from RBD amplicon sequencing of wastewater samples are consistent with clinical genome sequencing data in capturing the variants' dynamics. This consistency highlights

the effectiveness and accuracy of this approach in monitoring SARS-CoV-2 variants circulating in the sewershed. Additionally, the method allows for the early detection of emerging variants and provides a comprehensive view of variant diversity, emphasizing its importance for community-level surveillance efforts.

Winter and lag-phase are crucial periods for variant emergency and evolution

To better understand the epidemiological dynamics of SARS-CoV-2, we performed alpha diversity analyses of the variants identified from the RBD amplicon sequencing of wastewater samples.⁴⁹ Specifically, we examined how variant diversity changes with the seasons, during high transmission periods (in outbreak waves) versus low transmission periods (out of wave), and across different outbreak phases including lag, growth, stationary, and decline. Metrics including Observed, Chao1, Shannon, and Inverse Simpson indices were used to measure the richness and evenness of SARS-CoV-2 variants in the wastewater samples. The Observed index counts the actual number of unique variants, Chao1 estimates total species richness, Shannon quantifies the diversity considering both abundance and evenness, and Inverse Simpson emphasizes the dominance of the most abundant variants.^{49–54}

Results revealed significant seasonal variations in alpha diversity in all four metrics (Fig. 4A and Fig. S9). An increasing trend of alpha diversity was observed from spring to winter, with the highest alpha diversity observed in winter, suggesting that many variants emerged during the winter months. And summer months exhibited the lowest diversity. Moreover, higher diversity was also observed during the low transmission (out of wave) period across the four metrics (Fig. 4B and C and Fig. S10). This suggests that during these low transmission periods, a more diverse pool of variants circulates at low levels in the population, potentially due to reduced selective pressures compared to the outbreak period. When grouping data into different outbreak phases, we found that the highest diversity emerged in the lag phase, followed by the decline and growth phases, with the lowest diversity observed in the stationary phase (Fig. 4D and E). Beta diversity analysis showed no significant difference across the three sewersheds (Fig. 4F). Together, these results showed that RBD amplicon sequencing of wastewater is a powerful tool for uncovering the patterns of variant emergence and evolution in the sewershed.

Discussion

One significant insight from this study is the ability of RBD amplicon sequencing data to reveal patterns of viral evolution and variant emergence. Previous studies using whole genome sequencing of SARS-CoV-2 in wastewater have highlighted the early-warning potential of this approach and its ability to uncover variant dynamics.^{55–58} However, our study emphasizes that wastewater data can also be viewed from an ecological and evolutionary perspective, reflecting the community's viral variant landscape.

Our seasonal analysis revealed significant variations in alpha diversity, with higher diversity observed in winter and lower diversity in summer. This suggests that colder months may facilitate the emergence and spread of more variants, likely due to increased indoor activities, enhanced viral stability, and seasonal variations in immune responses.^{59–61} It is noteworthy that higher variant diversity was also observed during the lag phase of outbreaks, a period of low transmission. This seeming contradictory finding may suggest different underlying mechanisms at play. In winter, increased transmission events likely facilitate more frequent mutations and recombination, leading to higher variant diversity. But during lag phases, reduced viral competition may allow a wider range of

variants to coexist. From an ecological perspective, this phase may represent as “hot periods” of biodiversity and allow the accumulation and persistence of a heterogeneous variant population (viral quasi-species⁶²). While these variants may not immediately lead to outbreaks, they serve as a genetic pool that could contribute to future waves if conditions become favorable. Further research is also needed to fully understand variant dynamics across different epidemiological contexts. These insights underscore the importance of continuous monitoring of viral diversity to predict future outbreaks and implement timely public health interventions.

Our findings also demonstrate that RBD amplicon sequencing of wastewater samples effectively captures the dynamics of SARS-CoV-2 variants, showing strong correlation with clinical genome sequencing data for dominant variants at local, state, and national levels. Notably, our study identified variants circulating in the population earlier than observed in clinical sequencing data and detected variants and outbreaks that were not reported or noticed locally. This consistency between wastewater and clinical data, along with the earlier detection of variants, underscores the accuracy and reliability of RBD amplicon sequencing of wastewater in variant surveillance. In the post-pandemic period, as reporting of COVID-19 cases is not required or ceased, integrating wastewater data with available genome sequencing data offers a novel approach to validate and enhance wastewater surveillance findings. It also strengthens the role of wastewater surveillance as a complementary tool to clinical testing,^{24–26,63–66} especially valuable in areas with limited access to clinical sequencing resources.

Despite its advantages, RBD amplicon sequencing has limitations. While the RBD region is rich in information and useful for targeted analysis, it represents only a small portion of the viral genome. This limitation means that mutations outside the RBD region will be missed, and closely related variants with mutations in other regions may not be differentiated. Additionally, evolutionary changes occurring outside the targeted region, including recombination events, are not captured. The wastewater concentrating and PCR amplification steps in the protocol have the potential to introduce biases and errors,^{67–69} which could affect variant detection and quantification. Our analyses of synthetic variant communities showed the deviation ranges of –3.1% to 3.1% between measured and designed variant ratios (Fig. 1B). The high read mapping rates observed in these synthetic communities, ranging from 98.5% to 100%, suggest that PCR and sequencing error rates were low (0–1.5%). While these findings indicate overall reliability of this methodology, further testing in different wastewater samples will help confirm its robustness. On the other hand, RBD amplicon sequencing is faster, easier, and cheaper than whole genome sequencing,^{20,70,71} making it accessible for widespread use. Its high throughput capability with relatively lower costs allows for the rapid analysis of numerous samples, which is critical during pandemic conditions. By balancing these trade-offs, RBD amplicon sequencing offers a practical and efficient tool for ongoing variant surveillance, particularly when rapid data collection is necessary and resources are constrained.

Viruses have various mechanisms to enhance transmission between hosts. One significant strategy is mutating the receptor-binding domain in the spike gene, as observed with SARS-CoV-2, which has produced multiple waves of variants from Alpha to Omicron.^{5–9,72,73} These mutations in RBD enable the virus to evade adaptive immune responses from prior infections and/or vaccinations.^{74–77} Consequently, monitoring the RBD region is crucial for tracking viral evolution, predicting the emergence of new variants, and informing updated vaccine composition. This study highlights the utility of RBD amplicon sequencing and the integration of wastewater surveillance to track SARS-CoV-2 evolution and variant emergence. Our findings demonstrate that this tool captures the

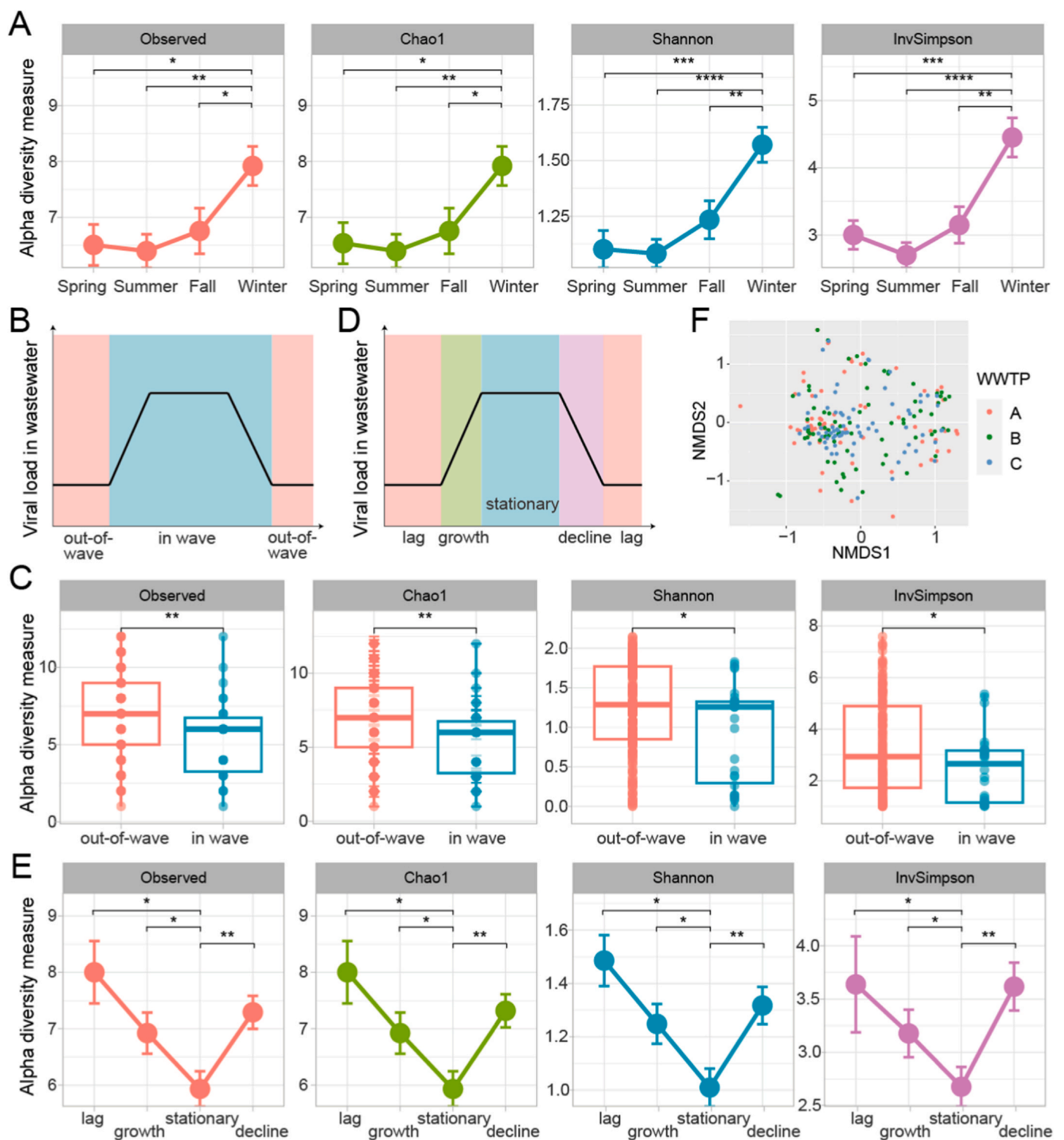


Fig. 4. Epidemiological patterns of variant emergence and evolution. (A) Alpha diversity comparison across seasons in the three sewersheds. (B) Schematic diagram of the high transmission (in-wave) and low transmission periods (out-of-wave) in an epidemic cycle. (C) Comparison of alpha diversity between in-wave and out-of-wave samples. (D) Schematic diagram of an epidemic cycle: lag, growth, stationary, and decline. (E) Alpha diversity comparison across the four phases: lag, growth, stationary, and decline. (F) Beta diversity analysis across all three WWTPs. No significant differences in viral diversity among the three sewersheds. Error bar: standard error; Wilcoxon rank-sum test was used for the group comparisons with significance: *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$.

dynamics of variant community, provides early detection of variants, aligns well with clinical genome sequencing data, and reveals epidemiological patterns in viral diversity. The concept of monitoring the receptor-binding domain can also be generalized to other viruses for viral evolution. This tool can improve our ability to predict future outbreaks and enhance public health preparedness.

Data availability

Sequencing data in this study has been uploaded to NCBI Sequence Read Archive, with the Bioproject accession ID: **PRJNA1156230**. Scripts will be shared on GitHub with the paper publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to Teresa T. Alcala at El Paso Water for sample collection and shipping. This work is supported by the PRIME award from the UTHealth Houston School of Public Health, National Science Foundation (DMS-2421257), National Institutes of Health RADxUP (1U01TR004355-01), and the Texas Epidemic Public Health Institute (TEPHI).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jinf.2024.106284.

References

- Shin HJ, Ku KB, Kim HS, Moon HW, Jeong GU, Hwang I, et al. Receptor-binding domain of SARS-CoV-2 spike protein efficiently inhibits SARS-CoV-2 infection and attachment to mouse lung. *Int J Biol Sci* 2021 Aug 28;17(14):3786–94.
- Tai W, He L, Zhang X, Pu J, Voronin D, Jiang S, et al. Characterization of the receptor-binding domain (RBD) of 2019 novel coronavirus: implication for development of RBD protein as a viral attachment inhibitor and vaccine. *Cell Mol Immunol* 2020 Jun;17(6):613–20.
- Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, et al. Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci* 2020 May 26;117(21):11727–34.
- Chakraborty C, Bhattacharya M, Sharma AR. Present variants of concern and variants of interest of severe acute respiratory syndrome coronavirus 2: their significant mutations in S-glycoprotein, infectivity, re-infectivity, immune escape and vaccines activity. *Rev Med Virol* 2022;32(2):e2270.
- Dhawan M, Sharma A, Priyanka, Thakur N, Rajkhowa TK, Choudhary OP. Delta variant (B.1.617.2) of SARS-CoV-2: mutations, impact, challenges and possible solutions. *Hum Vaccin Immunother* 2022;18(5):2068883.
- Adhikari P, Jawad B, Rao P, Podgornik R, Ching WY. Delta variant with P681R critical mutation revealed by ultra-large atomic-scale Ab Initio simulation: implications for the fundamentals of biomolecular interactions. *Viruses* 2022 Feb 24;14(3):465.
- Bhattacharya M, Chatterjee S, Sharma AR, Lee SS, Chakraborty C. Delta variant (B.1.617.2) of SARS-CoV-2: current understanding of infection, transmission, immune escape, and mutational landscape. *Folia Microbiol* 2023 Feb 1;68(1):17–28.
- Gohar SS, Ali F, Amin M. The Delta variant mutations in the receptor binding domain of SARS-CoV-2 show enhanced electrostatic interactions with the ACE2. *Med Drug Discov* 2022 Mar 1;13:100114.
- Pondé RAA. Physicochemical effect of the N501Y, E484K/Q, K417N/T, L452R and T478K mutations on the SARS-CoV-2 spike protein RBD and its influence on agent fitness and on attributes developed by emerging variants of concern. *Virology* 2022 Jul 1;572:44–54.
- Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med* 2018 Oct 1;20(10):1122–30.
- Christensen KD, Dukhovny D, Siebert U, Green RC. Assessing the costs and cost-effectiveness of genomic sequencing. *J Pers Med* 2015 Dec 10;5(4):470–86.
- Nurchis MC, Riccardi MT, Damiani G. Health technology assessment of whole genome sequencing in the diagnosis of genetic disorders: a scoping review of the literature. *Int J Technol Assess Health Care* 2022 Jan;38(1):e71.
- Long QX, Tang XJ, Shi QL, Li Q, Deng HJ, Yuan J, et al. Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat Med* 2020 Aug;26(8):1200–4.
- Sah P, Fitzpatrick MC, Zimmer CF, Abdollahi E, Juden-Kelly L, Moghadas SM, et al. Asymptomatic SARS-CoV-2 infection: a systematic review and meta-analysis. *Proc Natl Acad Sci USA* 2021 Aug 24;118(34):e2109229118.
- You Y, Yang X, Hung D, Yang Q, Wu T, Deng M. Asymptomatic COVID-19 infection: diagnosis, transmission, population characteristics. *BMJ Support Palliat Care* 2024 May 1;14(e1):e220–7.
- Kimball A. Asymptomatic and presymptomatic SARS-CoV-2 infections in residents of a long-term care skilled nursing facility—King County, Washington, March 2020. *MMWR Morb Mortal Wkly Rep* 2020;69:377–81. [cited 2024 Jun 26]. (<https://www.cdc.gov/mmwr/volumes/69/wr/mm6913e1.htm>).
- Thompson JR, Nanchaiah YV, Gu X, Lee WL, Rajal VB, Haines MB, et al. Making waves: wastewater surveillance of SARS-CoV-2 for population-based health management. *Water Res* 2020 Oct 1;184:116181.
- Xagorarakis I. Can we predict viral outbreaks using wastewater surveillance? *J Environ Eng* 2020 Nov 1;146(11):01820003.
- Kilaru P, Hill D, Anderson K, Collins MB, Green H, Kmush BL, et al. Wastewater surveillance for infectious disease: a systematic review. *Am J Epidemiol* 2023 Feb 1;192(2):305–22.
- Kuroiwa M, Gahara Y, Kato H, Morikawa Y, Matsui Y, Adachi T, et al. Targeted amplicon sequencing of wastewater samples for detecting SARS-CoV-2 variants with high sensitivity and resolution. *Sci Total Environ* 2023 Oct 1;893:164766.
- Iwamoto R, Yamaguchi K, Katayama K, Ando H, Setsukinai K, Kobayashi H, et al. Identification of SARS-CoV-2 variants in wastewater using targeted amplicon sequencing during a low COVID-19 prevalence period in Japan. *Sci Total Environ* 2023 Aug 20;887:163706.
- Lou EG, Sapoval N, McCall C, Bauhs L, Carlson-Stadler R, Kalvapalle P, et al. Direct comparison of RT-ddPCR and targeted amplicon sequencing for SARS-CoV-2 mutation monitoring in wastewater. *Sci Total Environ* 2022 Aug 10;833:155059.
- Environmental surveillance for SARS-COV-2 to complement public health surveillance – Interim Guidance [Internet]. [cited 2024 Jun 26]. Available from: (<https://www.who.int/publications/i/item/WHO-HEP-ECH-WSH-2022.1>).
- Bisseux M, Debroas D, Mirand A, Archimbaud C, Peigue-Lafeuille H, Bailly JL, et al. Monitoring of enterovirus diversity in wastewater by ultra-deep sequencing: an effective complementary tool for clinical enterovirus surveillance. *Water Res* 2020 Feb 1;169:115246.
- Fantilli A, Cola GD, Castro G, Sicilia P, Cachi AM, de los Angeles Marinza M, et al. Hepatitis A virus monitoring in wastewater: a complementary tool to clinical surveillance. *Water Res* 2023 Aug 1;241:120102.
- Smith T, Holm RH, Yeager R, Moore JB, Rouchka EC, Sokoloski KJ, et al. Combining community wastewater genomic surveillance with state clinical surveillance: a framework for SARS-CoV-2 Public Health Practice. *Food Environ Virol* 2022 Dec 1;14(4):410–6.
- Li J, Hosegood I, Powell D, Tschärke B, Lawler J, Thomas KV, et al. A global aircraft-based wastewater genomic surveillance network for early warning of future pandemics. *Lancet Glob Health* 2023 May 1;11(5):e791–5.
- Karthikeyan S, Levy JI, De Hoff P, Humphrey G, Birmingham A, Jepsen K, et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* 2022 Sep;609(7925):101–8.
- Jiang M, Wang ALW, Be NA, Mulakken N, Nelson KL, Kantor RS. Evaluation of the impact of concentration and extraction methods on the targeted sequencing of human viruses from wastewater. *Environ Sci Technol* 2024 May 14;58(19):8239–50.
- Levy JI, Andersen KG, Knight R, Karthikeyan S. Wastewater surveillance for public health. *Science* 2023 Jan 6;379(6627):26–7.
- Parkins MD, Lee BE, Acosta N, Bautista M, Hubert CRJ, Hudey SE, et al. Wastewater-based surveillance as a tool for public health action: SARS-CoV-2 and beyond. *Clin Microbiol Rev* 2023 Dec 14;37(1):e00103–22.
- Li Y, Miyani B, Childs KL, Shiu SH, Xagorarakis I. Effect of wastewater collection and concentration methods on assessment of viral diversity. *Sci Total Environ* 2024 Jan 15;908:168128.
- Gregory DA, Trujillo M, Rushford C, Flury A, Kannoly S, San KM, et al. Genetic diversity and evolutionary convergence of cryptic SARS-CoV-2 lineages detected via wastewater sequencing. *PLoS Pathog* 2022 Oct 14;18(10):e1010636.
- Baaijens JA, Zulli A, Ott IM, Nika I, van der Lugt MJ, Petrone ME, et al. Lineage abundance estimation for SARS-CoV-2 in wastewater using transcriptome quantification techniques. *Genome Biol* 2022 Nov 8;23(1):236.
- Cancela F, Ramos N, Smyth DS, Etchebehere C, Berois M, Rodríguez J, et al. Wastewater surveillance of SARS-CoV-2 genomic populations on a country-wide scale through targeted sequencing. *PLoS One* 2023 Apr 21;18(4):e0284483.
- Gregory DA, Wieberg CG, Wenzel J, Lin CH, Johnson MC. Monitoring SARS-CoV-2 populations in wastewater by amplicon sequencing and using the Novel Program SAM Refiner. *Viruses* 2021 Aug;13(8):1647.
- Adamopoulos PG, Diamantopoulos MA, Boti MA, Zafeiriadou A, Galani A, Kostakis M, et al. Spike-Seq: an amplicon-based high-throughput sequencing approach for the sensitive detection and characterization of SARS-CoV-2 genetic variations in environmental samples. *Sci Total Environ* 2024 Mar 1;914:169747.
- Rodríguez GM, Volpato FCZ, Wink PL, Paiva RM, Barth AL, de-París F. SARS-CoV-2 variants of concern: presumptive identification via Sanger sequencing analysis of the receptor binding domain (RBD) region of the S Gene. *Diagnostics* 2023 Jan;13(7):1256.
- Aguayo-Acosta A, Oyervides-Muñoz MA, Rodríguez-Aguilón KO, Ovalle-Carcano A, Romero-Castillo KD, Robles-Zamora A, et al. Omicron and Delta variant prevalence detection and identification during the fourth COVID-19 wave in Mexico using wastewater-based epidemiology. *IJID Reg* 2023 Nov 10;10:44–51.
- Oghuan J, Chavarria C, Vanderwal SR, Gitter A, Ojarega AA, Monserrat C, et al. Wastewater analysis of Mpox virus in a city with low prevalence of Mpox disease: an environmental surveillance study. *Lancet Reg Health Am* 2023 Dec 1;28:100639 [cited 2023 Nov 26]. Available from: ([https://www.thelancet.com/journals/lanam/article/PIIS2667-193X\(23\)00213-2/fulltext](https://www.thelancet.com/journals/lanam/article/PIIS2667-193X(23)00213-2/fulltext)).
- Wu F, Xiao A, Zhang J, Moniz K, Endo N, Armas F, et al. SARS-CoV-2 RNA concentrations in wastewater foreshadow dynamics and clinical presentation of new COVID-19 cases. *Sci Total Environ* 2022 Jan 20;805:150121.
- Wu F, Xiao A, Zhang J, Moniz K, Endo N, Armas F, et al. Wastewater surveillance of SARS-CoV-2 across 40 U.S. states from February to June 2020. *Water Res* 2021 Jul 2;202:117400.
- Wu F, Xiao A, Zhang J, Gu X, Lee WL, Kauffman K, et al. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. *mSystems* 2020 Aug 25;5(4):10–128.
- Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016 Jul;13(7):581–3.

45. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction*. BMC Bioinforma 2012 Jun 18;13(1):134.
46. Wirth W, Duchene S. GISAIDR [Internet]. Zenodo; 2022 [cited 2024 Jul 9]. Available from: (<https://zenodo.org/records/6474693>).
47. Amman F, Markt R, Endler L, Hupfauf S, Agerer B, Schedl A, et al. *Viral variant-resolved wastewater surveillance of SARS-CoV-2 at national scale*. Nat Biotechnol 2022 Dec;40(12):1814–22.
48. Gupta P, Liao S, Ezekiel M, Novak N, Rossi A, LaCross N, et al. *Wastewater genomic surveillance captures early detection of Omicron in Utah*. Microbiol Spectr 2023 May 8;11(3):e00391–23.
49. McMurdie PJ, Holmes S. *phyloseq: an R Package for reproducible interactive analysis and graphics of microbiome census data*. PLoS One 2013 Apr 22;8(4):e61217.
50. Fisher RA, Corbet AS, Williams CB. *The relation between the number of species and the number of individuals in a random sample of an animal population*. J Anim Ecol 1943 May;12(1):42.
51. Chao A, Bunge J. *Estimating the number of species in a stochastic abundance model*. Biometrics 2002 Sep 1;58(3):531–9.
52. Shannon CE. *A mathematical theory of communication*. Bell Syst Tech J 1948;27:379–423.
53. Simpson EH. *Measurement of diversity*. Nature 1949 Apr;163(4148):688.
54. Willis AD. *Rarefaction, alpha diversity, and statistics*. Front Microbiol 2019 Oct 23;10:2407.
55. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. *Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands*. Nat Med 2020 Sep;26(9):1405–10.
56. Mao K, Zhang K, Du W, Ali W, Feng X, Zhang H. *The potential of wastewater-based epidemiology as surveillance and early warning of infectious disease outbreaks*. Curr Opin Environ Sci Health 2020 Oct 1;17:1–7.
57. Ahmed W, Tschärke B, Bertsch PM, Bibby K, Bivins A, Choi P, et al. *SARS-CoV-2 RNA monitoring in wastewater as a potential early warning system for COVID-19 transmission in the community: a temporal case study*. Sci Total Environ 2021 Mar 20;761:144216.
58. Mackuľak T, Gál M, Špalková V, Fehér M, Briestenská K, Mikušová M, et al. *Wastewater-based epidemiology as an early warning system for the spreading of SARS-CoV-2 and its mutations in the population*. Int J Environ Res Public Health 2021 Jan;18(11):5629.
59. Landier J, Paireau J, Rebaudet S, Legendre E, Lehot L, Fontanet A, et al. *Cold and dry winter conditions are associated with greater SARS-CoV-2 transmission at regional level in western countries during the first epidemic wave*. Sci Rep 2021 Jun 17;11(1):12756.
60. Neher RA, Dyrda R, Druelle V, Hodcroft EB, Albert J. *Potential impact of seasonal forcing on a SARS-CoV-2 pandemic*. Swiss Med Wkly 2020 Mar 16;150:w20224 [cited 2024 Jul 12]; Available from: (<https://smw.ch/index.php/smw/article/view/2745>).
61. Kissler SM, Tedijanto C, Goldstein E, Grad YH, Lipsitch M. *Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period*. Science 2020 May 22;368(6493):860–8.
62. Jary A, Leducq V, Malet I, Marot S, Klement-Frutos E, Teyssou E, et al. *Evolution of viral quasispecies during SARS-CoV-2 infection*. Clin Microbiol Infect Dis 2020 Nov;26(11):1560.e1–4.
63. Gitter A, Oghuan J, Godbole AR, Chavarria CA, Monserrat C, Hu T, et al. *Not a waste: wastewater surveillance to enhance public health*. Front Chem Eng [Internet] 2023 Jan 9;4: 1112876.
64. CDC. *About CDC's National Wastewater Surveillance System (NWSS)* [Internet]. Centers for Disease Control and Prevention; 2024 [cited 2024 Jun 26]. Available from: (<https://www.cdc.gov/nwss/about.html>).
65. DeJonge PM, Adams C, Pray I, Schussman MK, Fahney RB, Shafer M, et al. *Wastewater surveillance data as a complement to emergency department visit data for tracking incidence of influenza A and respiratory syncytial virus – Wisconsin, August 2022–March 2023*. MMWR Morb Mortal Wkly Rep 2023 Sep 15;72(37):1005–9.
66. Bonanno Ferraro G, Veneri C, Mancini P, Iaconelli M, Suffredini E, Bonadonna L, et al. *A state-of-the-art scoping review on SARS-CoV-2 in sewage focusing on the potential of wastewater surveillance for the monitoring of the COVID-19 pandemic*. Food Environ Virol 2022 Dec 1;14(4):315–54.
67. Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. *PCR-induced sequence artifacts and bias: insights from comparison of Two 16S rRNA clone libraries constructed from the same sample*. Appl Environ Microbiol 2005 Dec;71(12):8966–9.
68. Kanagawa T. *Bias and artifacts in multitemplate polymerase chain reactions (PCR)*. J Biosci Bioeng 2003 Jan 1;96(4):317–23.
69. Krehenwinkel H, Wolf M, Lim JY, Rominger AJ, Simison WB, Gillespie RG. *Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding*. Sci Rep 2017 Dec 15;7(1):17668.
70. Techera C, Tomás G, Grecco S, Williman J, Hernández M, Olivera V, et al. *A rapid and affordable amplicon-based method for next-generation genome sequencing of the infectious bursal disease virus*. J Virol Methods 2023 Dec 1;322:114807.
71. Kong X, Gao P, Jiang Y, Lu L, Zhao M, Liu Y, et al. *Discrimination of SARS-CoV-2 omicron variant and its lineages by rapid detection of immune-escape mutations in spike protein RBD using asymmetric PCR-based melting curve analysis*. Virol J 2023 Aug 25;20(1):192.
72. Zhao Z, Zhou J, Tian M, Huang M, Liu S, Xie Y, et al. *Omicron SARS-CoV-2 mutations stabilize spike up-RBD conformation and lead to a non-RBM-binding monoclonal antibody escape*. Nat Commun 2022 Aug 24;13(1):4958.
73. Farheen S, Araf Y, Tang YD, Zheng C. *The Deltacron conundrum: its origin and potential health risks*. J Med Virol 2022 Nov;94(11):5096–102.
74. Chaudhari AM, Joshi M, Kumar D, Patel A, Lokhande KB, Krishnan A, et al. *Evaluation of immune evasion in SARS-CoV-2 Delta and Omicron variants*. Comput Struct Biotechnol J 2022 Aug 8;20:4501–16.
75. Willett BJ, Grove J, MacLean OA, Wilkie C, De Lorenzo G, Furnon W, et al. *SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway*. Nat Microbiol 2022 Aug;7(8):1161–79.
76. Kudriavtsev AV, Vakhrusheva AV, Novoseletsky VN, Bozdaganyan ME, Shaitan KV, Kirpichnikov MP, et al. *Immune escape associated with RBD Omicron mutations and SARS-CoV-2 evolution dynamics*. Viruses 2022 Jul 22;14(8):1603.
77. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, de Silva TI, et al. *SARS-CoV-2 variant biology: immune escape, transmission and fitness*. Nat Rev Microbiol 2023 Mar;21(3):162–77.