HHS Public Access

Author manuscript

Lancet Microbe. Author manuscript; available in PMC 2024 April 26.

Published in final edited form as:

Lancet Microbe. 2024 April ; 5(4): e335–e344. doi:10.1016/S2666-5247(23)00372-5.

Tracing the origin of SARS-CoV-2 omicron-like spike sequences detected in an urban sewershed: a targeted, longitudinal surveillance study of a cryptic wastewater lineage

Martin M Shafer[*],
Max J Bobholz[*],
William C Vuyk[*],
Devon A Gregory[*],
Adelaide Roguet,
Luis A Haddock Soto,
Clayton Rushford,
Kayley H Janssen,
Isla E Emmen,
Hunter J Ries,
Hannah E Pilch,
Paige A Mullen,
Rebecca B Fahney,
Wanting Wei,
Matthew Lambert,
Jeff Wenzel,
Peter Halfmann,
Yoshihiro Kawaoka,
Nancy A Wilson,
Thomas C Friedrich,

Correspondence to: Prof Marc C Johnson, School of Medicine, University of Missouri, Columbia, MO 65211, USA, marcjohnson@missouri.edu.

Contributors

MMS, MJB, WCV, and DAG all contributed equally to this manuscript, especially in conceptualisation, data curation, formal analysis, investigation, method, visualisation, and writing. LAHS contributed to formal analysis, visualisation, and manuscript review. AR, CR, and KHJ contributed to method, data curation, and formal analysis. AR also contributed to manuscript writing, visualisation, and review. HEP, PAM, and RBF contributed to method and investigation. IEE, HJR, and WW contributed to manuscript review and validation. PH and YK contributed to the investigation and validation of the project. ML, JW, NAW, IWP, and RW contributed to the investigation, supervision, and validation of the study, and review of the manuscript. TCF, DHO, and MCJ contributed to conceptualisation, formal analysis, funding acquisition, investigation, writing (original draft), and supervision. All authors had full access to all the data in the study and accept responsibility for the decision to submit for publication. MMS, MJB, WCV, DAG, AR, CR, KHJ, TCF, DHO, and MCJ have verified the data in the study, and MMS, MJB, WCV, DAG, IWP, RW, TCF, DHO, and MCJ had final responsibility for the decision to submit for publication.

[*]Contributed equally

See **Online** for appendix 1

For the **project data portal** see https://go.wisc.edu/4134pl

See **Online** for appendix 2

Ian W Pray,

Ryan Westergaard,

David H O'Connor*,

Marc C Johnson*

Wisconsin State Laboratory of Hygiene (M M Shafer PhD, A Roguet PhD, K H Janssen PhD, H E Pilch MS, P A Mullen BS, R B Fahney BS), Department of Pathology and Laboratory Medicine (M J Bobholz BS, W C Vuyk BS, I E Emmen BS, N A Wilson PhD, Prof D H O'Connor PhD), Department of Pathobiological Sciences (L A Haddock Soto MS, H J Ries BS, W Wei BS, P Halfmann PhD, Prof Y Kawaoka PhD, Prof T C Friedrich PhD), and Department of Medicine (M Lambert MD, R Westergaard MD PhD), University of Wisconsin-Madison, Madison, WI, USA; Wisconsin Department of Health Services, Madison, WI, USA (M Lambert, R Westergaard); School of Medicine, University of Missouri, Columbia, MO, USA (D A Gregory PhD, C Rushford MS, Prof M C Johnson PhD); Missouri Department of Health and Senior Services, Jefferson City, MO, USA (J Wenzel BS); Wisconsin Department of Health Services, Madison, WI, USA (I W Pray PhD)

Summary

Background—The origin of novel SARS-CoV-2 spike sequences found in wastewater, without corresponding detection in clinical specimens, remains unclear. We sought to determine the origin of one such cryptic wastewater lineage by tracking and characterising its persistence and genomic evolution over time.

Methods—We first detected a cryptic lineage, WI-CL-001, in municipal wastewater in Wisconsin, USA, in January, 2022. To determine the source of WI-CL-001, we systematically sampled wastewater from targeted sub-sewershed lines and maintenance holes using compositing autosamplers. Viral concentrations in wastewater samples over time were measured by RT digital PCR. In addition to using metagenomic 12s rRNA sequencing to determine the virus's host species, we also sequenced SARS-CoV-2 spike receptor binding domains, and, where possible, whole viral genomes to identify and characterise the evolution of this lineage.

Findings—We traced WI-CL-001 to its source at a single commercial building. There we detected the cryptic lineage at concentrations as high as $2{\cdot}7 \times 10^9$ genome copies per L. The majority of 12s rRNA sequences detected in wastewater leaving the identified source building were human. Additionally, we generated over 100 viral receptor binding domain and whole-genome sequences from wastewater samples containing the cryptic lineage collected over the 13 consecutive months this virus was detectable (January, 2022, to January, 2023). These sequences contained a combination of fixed nucleotide substitutions characteristic of Pango lineage B.1.234, which circulated in humans in Wisconsin at low levels from October, 2020, to February, 2021. Despite this, mutations in the spike gene and elsewhere resembled those subsequently found in omicron variants.

Interpretation—Wepropose that prolonged detectionofWI-CL-001 in wastewater indicates persistent shedding of SARS-CoV-2 from a single human initially infected by an ancestral B.1.234 virus. The accumulation of convergent omicron-likemutations in WI-CL-001's ancestral B.1.234 genome probably reflects persistent infection and extensive within-host evolution. People who

shed cryptic lineages could be an important source of highly divergent viruses that sporadically emerge and spread.

Funding—The Rockefeller Foundation, Wisconsin Department of Health Services, Centers for Disease Control and Prevention, National Institute on Drug Abuse, and the Center for Research on Influenza Pathogenesis and Transmission.

Introduction

SARS-CoV-2-infected hosts shed viral RNAin their stool and urine. Furthermore, the virus is known to infect the gastrointestinal tract and kidney tissues.[1,2] Accordingly, wastewater surveillance has become an important complement to clinical nasal swab testing in monitoring SARS-CoV-2 and has enabled the identification of unique genetic lineages not detected via clinical testing.[3–6] SARS-CoV-2 lineages that have been detected in wastewater but not in clinical specimens have been termed cryptic lineages. Although most cryptic lineages are not related to omicron, they frequently have evolved mutational landscapes that are convergent with those of omicron lineages.[6]

Compared with previously circulating SARS-CoV-2 viruses, the first detected omicron lineage (BA.1) had a highly divergent spike protein.[7] The BA.1 spike protein had 12 lineage-defining amino acid substitutions in the receptor binding domain (RBD) between residues 412 and 579 (Lys417Asn, Asn440Lys, Gly446Ser, Ser477Asn, Thr478Lys, Glu484Ala, Gln493Arg, Gly496Ser, Gln498Arg, Asn501Tyr, Tyr505His, and Thr547Lys).[7–9] Although global circulation of BA.1 did not begin until late 2021, ten of these RBD polymorphisms were observed in various combinations in sequence reads fromNew York City (NY, USA) wastewater samples collected in the first half of 2021. One wastewater sample, collected in May, 2021, had sequences encoding 24 amino acid substitutions in the spike RBD. Additional cryptic lineages were detected in wastewater samples from Missouri and California.[6]

There are two leading hypotheses for the source of these cryptic lineage sequences. First, an animal reservoir might be introducing these viruses into wastewater. SARS-CoV-2 exhibits a broad host range, including wild animals, livestock, and household pets.[10–12] Multiple studies have found evidence of SARS-CoV-2 transmission between human and animal populations;[11–14] thus, it is plausible that animal reservoirs of cryptic lineages exist undetected, with ongoing virus transmission and exchange between animal species.[15]

Alternatively, cryptic SARS-CoV-2 lineage sequences in wastewater could be derived from humans with unsampled infections.[16] In a recent cohort study, 49·2% of participants had viral RNA (vRNA) in stool in the first week of infection; vRNA remained detectable in stool 4 months later in 12·7% of individuals although all had cleared vRNA from the nasopharynx.[17] Such individuals could contribute vRNA to wastewater even while testing negative via nasal swabs. People with immuno-compromising conditions are at high risk for prolonged infections, and suboptimal immune responses in such individuals could select for antigenic variation over the course of infection, driving diversification of SARS-CoV-2 within these hosts.[18,19] Such selection could account for the observation that cryptic lineages tend to accumulate high levels of non-synonymous variation in the spike protein

while otherwise maintaining the characteristic mutations from viruses that are no longer common in circulation. In this study we aimed to investigate the origin of a specific cryptic wastewater lineage, named WI-CL-001, after its initial detection in sequences collected from a metropolitan publicly owned treatment works (POTW) in Wisconsin, USA, in January, 2022, as part of routine SARS-CoV-2 wastewater surveillance.

Methods

Study design

Wastewater samples for this study were collected from one metropolitan area in Wisconsin, USA, by wastewater engineers from the city wastewater utility. All wastewater samples were integrated over a 24-h period using compositing 6712 and 6712C autosamplers (ISCO, Lincoln, NE, USA). The Wisconsin State Laboratory of Hygiene (WSLH), in consultation with utility engineers and the Wisconsin Department of Health Services (WDHS), determined specific testing locations in the wastewater collection system, allowing for the gradual narrowing down of the origin of WI-CL-001. Samples used for this investigation included daily wastewater samples from the central POTW and twice per week monitoring of the primary subdistrict lines (January, 2022, through the end of March, 2023), which were collected through the Wisconsin Wastewater Monitoring Program. Additional sampling for this investigation included targeted maintenance hole testing from March, 2022, through mid-May, 2022, directed at isolating the source of the cryptic lineage within the wastewater collection system. Further targeted sampling occurred at a commercial building sewer line access point serving six toilets, henceforth referred to as facility line B, on June 16, Aug 16, Sept 21, and Sept 27, 2022. Using this strategy, the sampled human source population of the WI-CL-001 signal was narrowed from more than 100 000 people to fewer than 30 people (figure 1A). After consulting with local public health officials and facility managers, employees present at the facility were offered RT-PCR testing for SARS-CoV-2 via nasal swabs; 19 of approximately 30 employees provided samples. Further methodological information about wastewater and clinical sample collection can be found in appendix 1 (pp 2, 4).

This activity was reviewed by the US Centers for Disease Control and Prevention (CDC) and the WDHS and was conducted consistent with applicable federal law and CDC policy. Details about the city, sewer plant, and locations of individual sampling sites have been concealed to protect the privacy of participants and residents. The authors representing the state public health authority had numerous conversations with the leadership of the identified source building in person and by telephone, to explain the rationale for the investigation and describe the ways we would protect the identity of the company and employees. Health authorities further consulted with a university-based bioethicist not affiliated with the Department of Health Services, who concurred that the investigation should be considered a public health surveillance activity and did not constitute human participant research. Specifically, we intentionally did not describe the investigation as a response to a known public health hazard, but rather as an activity to learn more about new technologies, and make sense of an unexpected finding that could someday be of public health significance. The wastewater findings were taken as a signal that people who worked in the building

might benefit from testing and linkage to care; once this testing was made available, we communicated to the building's occupants that the results would have the same privacy and confidentiality protections as all other testing.

Procedures

Wastewater samples were shared between the WSLH and the University of Missouri, with the WSLH focusing on virus quantitation and whole-genome sequencing (WGS), and the University of Missouri focusing on RBD-targeted sequencing. These two institutions had different procedures for viral RNA isolation, quantification, and sequencing.

At the WSLH, after the addition of a bovine coronavirus (BCoV) viral recovery control and the concentration of virus using Nanotrap Magnetic Virus Particles (Ceres Nanosciences, Manassas, VA, USA) on a Kingfisher Apex instrument (ThermoFisher Scientific, Waltham, MA, USA), total nucleic acids were extracted using Maxwell HT Environmental TNA kits (Promega, Madison, WI, USA) on a Kingfisher Flex instrument (ThermoFisher Scientific) following manufacturer instructions.

The WSLH quantified the concentration of SARS-CoV-2, BCoV (viral recovery control), and pepper mild mottle virus, a fecal marker, in each sample using RT digital PCR (RT-dPCR). PCR inhibition was probed with bovine respiratory syncytial virus RNA spiked into each PCR reaction. Further details on our viral RNA isolation and quantification procedures are provided in appendix 1 (pp 2–3).

For SARS-CoV-2 WGS at the WSLH, 13 μL of total nucleic acids from the WSLH's wastewater extracts were used as input to Qiagen's Direct SARS-CoV-2 Enhancer kit (Qiagen, Germantown, MD, USA). Amplicon libraries were prepared on a Biomek i5 liquid handler (Beckman Coulter, Brea, CA, USA). Libraries were quantified using a High Sensitivity Qubit 1X dsDNA HS Assay Kit (ThermoFisher Scientific), and fragment size was analysed by a QIAxcel Advanced instrument and QX DNA Screening Kit (Qiagen). Sequencing was performed on an Illumina MiSeq instrument using MiSeq Reagent v2 (300 cycles) kits (Illumina, San Diego, CA, USA).

Fastq files were analysed with the nf-core/viralrecon 2.5 workflow using the SARS-CoV-2 Wuhan-Hu-1 reference genome (Genbank accession number MN908947.3).[20] The workflow was initiated as outlined on the project's data portal. Additional details are provided in appendix 1 (pp 3–4).

The University of Missouri concentrated the virus using a polyethylene glycol 8000 (PEG) protocol on pre-filtered wastewater samples (0·22 μM polyethersulfone membrane [Millipore, Burlington, MA, USA]). Samples were incubated with PEG and 1·2 M NaCl, centrifuged, and the RNA was isolated from the pellet with the QIAamp Viral RNA Mini Kit (Qiagen). Further information on these procedures can be found in appendix 1 (p 2).

A nested RT-PCR approach was used to selectively amplify non-omicron spike protein RBD regions from wastewater samples. Amplified RBD regions were then sequenced using an Illumina MiSeq instrument (Illumina) and analysed using SAMRefiner software version

1.4.[21] WI-CL-001's unique RBD sequences were used to identify and track the lineage across time and space. Additional details are provided in appendix 1 (p 3).

The University of Missouri also conducted 12S rRNA sequencing to assess what species were contributing to facility line B in June, 2022. Additional details are provided in appendix 1 (p 4).

Virus culture

To remove debris, samples were centrifuged twice at 3500 rpm at 4°C for 15 min and then passed through a 0.8 μM syringe filter (Agilent, Santa Clara, CA, USA) or left unfiltered. Samples (1 mL) were incubated on nearly confluent Vero E6-TMPRSS2 (JCRB1819) or Vero E6-TMPRSS2/hACE2 cells seeded the day before in TC25 cm^2 flasks for 1 hour at 37°C. After the incubation, cells were washed twice and media was added back to the cells. The media contained 20 000 μg/mL of penicillin and streptomycin, 50 μg/mL of amphotericin, and 10 μg/mL of chloramphenicol. Cells were monitored daily for potential virus-induced cytopathic effects. After 10 days, a blind passage was performed using the entire volume of media (~4 mL) to fresh, nearly confluent cells seeded the day before in TC175 cm^2 flasks.

Statistical analysis

Variant proportions were assessed from WGS data using Freyja version 1.3.11, a tool previously developed to estimate the proportions of SARS-CoV-2 variants in deep sequence data containing mixed populations.[22] Additional details are provided in the appendix 1 (p 5).

We generated root-to-tip regressions using iqtree (version 2.2.0.3) to infer a maximum likelihood phylogenetic tree of all full SARS-CoV-2 consensus genomes belonging to Pango lineage B.1.234 (the inferred parent of WI-CL-001) from specimens collected in a 12-state Midwest region available in Genbank. Molecular clock rates and genetic distances were obtained through TreeTime (version 0.9.3). Additional details and citations are provided in appendix 1 (p 5).

To assess the selective environment in which WI-CL-001 evolved, variant calls against reference genome MN908947.3 obtained through the nf-core/viralrecon workflows were processed using custom Python (version 3.8) scripts. Variants were classified using SnpEff (version 5.0). The proportion of non-synonymous variants per site was calculated using SNPGenie (version 2019.10.31), and a binomial probability distribution was implemented using SciPy's binomtest function (version 1.9.3). To characterise genetic diversity within each sample, we used the summary statistic π, which quantifies the number of pairwise differences per non-synonymous (πN) and synonymous (πS) site within a set of sequences. A Mann-Whitney two-sided test was applied to test the difference between πN and πS in each gene, while a one-sided test was used to test for an enrichment of the πN value of spike against the πN value on the other genes. The average Hamming distance between B.1.234 isolates and the ancestral sequence Wuhan-Hu-1 (MN908947.3 reference sequence) was calculated to obtain synonymous and non-synonymous divergence values. Additional details and citations are provided in appendix 1 (p 5).

Role of the funding source

This Article underwent CDC's clearance review process due to the involvement of CDC coauthors; however, the funders did not play a direct role in the study design, data collection, data analysis, or Article preparation.

Results

On Jan 11, 2022, a cryptic lineage containing at least six unusual spike RBD substitutions (Val445Ala, Tyr449His, Asn460Lys, Glu484Gln, Phe490Tyr, Gln493Lys) was first detected in a composite wastewater sample from a metropolitan POTW in Wisconsin (figure 1C). Over the following 6 months, the source of the RBD sequences was narrowed to a single commercial building, and subsequently a single sewer line serving six toilets (facility line B;figure 1A). As the sampling effort progressively narrowed down the source, an increasing proportion of total SARS-CoV-2 sequences at each sample site was classified as WI-CL-001 (labelled B.1.234 in figure 1B). Accordingly, higher B.1.234 proportions were seen in subdistrict 5 than in the POTW over many months, corresponding to the closer proximity of subdistrict 5 to the source of WI-CL-001 (figure 1C).

Consistently high wastewater SARS-CoV-2 RNA concentrations were observed in all samples collected from facility line B. Concentrations of approximately $5{\cdot}2 \times 10^8$, $1{\cdot}6 \times 10^9$, $2{\cdot}7 \times 10^9$, and $5{\cdot}5 \times 10^8$ genome copies (gc) per L undiluted wastewater were quantified using RT-dPCR on samples collected on June 16, Aug 16, Sept 23, and Sept 27, respectively (figure 1B). Despite these high vRNA concentrations, viable virus could not be cultured from wastewater after multiple attempts. 12s rRNA sequencing detected predominantly human rRNA sequences from this source. Chicken 12s rRNA sequences, the next largest taxon identified, were less than 0·05% of the sample (appendix 1 p 9). Despite these findings, none of the building occupants who volunteered for nasal swab testing organised by the local public health department were positive for SARS-CoV-2 in June, 2022. The wastewater signal from the subdistrict 5 line became undetectable in January, 2023, after a decline in signal concentration beginning the 46th week of 2022 (figure 1C).

The high levels of WI-CL-001 vRNA in facility line B facilitated the amplification and sequencing of the lineage's entire genome from each facility line B sample. At the consensus level, all sequences were classified as lineage B.1.234 by pangolin. In SARS-CoV-2 genomic surveillance using clinical specimens, B.1.234 viruses were detected in Wisconsin between Sept 2, 2020, and March 30, 2021.[23]

The B.1.234 lineage does not have any characteristic spike RBD amino acid changes relative to the reference Wuhan-Hu-1 (figure 2A). Sequencing single amplicons spanning the RBD allowed us to define haplotypes (ie, specific combinations of mutations found together in a single RNA molecule). We repeatedly sequenced spike RBD in wastewater samples from the subdistrict 5 line and haplotypes of WI-CL-001 were detected every month from January, 2022, to January, 2023 (figure 2B). We detected 54 RBD haplotypes during this time, with a mean of 1·35 haplotypes (range 1–3) being detected at any one timepoint (appendix 1 p 7). Some of the amino acid substitutions on these haplotypes predated the emergence of the

same amino acid changes, or different changes at the same amino acid residues, in globally circulating omicron lineages (figure 3).[24]

The cryptic lineage is also divergent outside of the spike RBD. When plotted on a radial phylogenetic tree using Nextclade, Illumina whole-genome consensus sequences from facility line B show divergence from the Wuhan-Hu-1 reference similar to clade 22B and XBB* omicron lineages (appendix 1 p 8).[25] To investigate this further, we used iVar output from the nf-core/viralrecon workflow to identify variants at a frequency of 25% or more (appendix 1 p 10; appendix 2).[20] One interesting mutation is in the N-terminal ectodomain of the membrane protein, where a 15 nucleotide in-frame insertion (Ile8delinsSerAsnAsnSerGluPhe) is present at an average frequency of 92·4% in all facility line B whole-genome sequences (appendix 1 pp 11–13).

We next asked whether the unusual combinations of mutations present in WI-CL-001 could be the result of natural selection favouring non-synonymous (ie, amino acid-changing) mutations. First, we found that mutations accumulated in WI-CL-001 faster than expected on the basis of the nucleotide substitution rate that prevailed when B.1.234 viruses were circulating in the US midwest ($4{\cdot}24 \times 10^{-4}$ substitutions per site per year [SD 7×10^{-5}]; figure 4A). Across the four timepoints with available genome sequences, there was a notable excess of non-synonymous nucleotide substitutions (mean 121·8 [SD 16·3]) relative to synonymous ones (mean 22·5 [SD 4·7]; figure 4B). When estimating π, we found that within the spike gene, πN was significantly greater than πS at each timepoint, which could indicate ongoing diversifying selection on spike (figure 4C). Spike also had significantly higher non-synonymous diversity compared with ORF1AB, ORF3A, M, ORF6, ORF7A, and N at each timepoint (figure 4C). Because π counts pairwise differences per site within a sample, mutations that have become fixed or nearly fixed within the virus population do not contribute to π values. We next calculated divergence between each sequenced virus (either B.1.234 variants or WI-CL-001) and the ancestral sequence Wuhan-Hu-1 (figure 4D). Both synonymous and non-synonymous divergence values were substantially higher for WI-CL-001 than for B.1.234 viruses.

Discussion

We traced the source of a cryptic SARS-CoV-2 lineage, first detected in wastewater from a metropolitan POTW, to a sewer line within a commercial building (facility line B). Non-human animal sequences made up a negligible proportion of 12s rRNA sequences detected within facility line B, making an animal source highly improbable. The lineage was not detected by voluntary nasal swab testing at the source building, suggesting that a multiperson upper-respiratory outbreak was not a probable source of the cryptic signal. Combining our observations, we posit that the simplest explanation for the appearance and persistence of WI-CL-001 is that a single person, originally infected when B.1.234 was in circulation, developed a persistent infection and continued to excrete viruses into wastewater throughout 2022. The WI-CL-001 signal became undetectable after 53 weeks, at the end of a multiweek decline in signal strength. This is one of the longest periods of continuous detection that we are aware of for a SARS-CoV-2 cryptic lineage.[6]

The average SARS-CoV-2 concentration we detected in samples from facility line B ($1{\cdot}3 \times 10^9$ gc/L) was 7 $\log_2$ fold change higher by RT-dPCR than the highest signal previously measured ($8{\cdot}8 \times 10^6$ gc/L) out of over 12 000 wastewater samples collected throughout Wisconsin between 2020 and 2023. Upon searching PubMed for comparable results in other studies using the terms "(((SARS-CoV-2) AND (wastewater)) AND (concentration)) AND (building)" on Sept 15, 2023, none of the 40 articles we found reported a single-building wastewater SARS-CoV-2 concentration above 1×10^8 gc/L. The high source concentration of WI-CL-001 might help to resolve a paradox from earlier cryptic lineage studies: if cryptic lineages come from only a single source, how could they be detected in a dilute municipal wastewater sample? On the basis of wastewater flow data from the subdistrict 5 line and estimations of typical toilet use, we would expect the WI-CL-001 vRNA to be diluted from a wastewater volume of approximately 200 gallons at facility line B into a volume of 8 million gallons at the subdistrict 5 line. Thus, if there were 2 billion gc/L at facility line B, we would expect to detect approximately 50 000 gc/L at the subdistrict 5 line. This value is similar to what we actually observed over 13 months. Hence, our observations are consistent with a persistently infected individual shedding high amounts of SARS-CoV-2 RNA into wastewater at the source building throughout 2022.

The large preponderance of non-synonymous substitutions in the facility line B viral genomes suggests that this virus has undergone diversifying selection on spike, and perhaps other genes. This possibility is consistent with reports of individuals with prolonged SARS-CoV-2 infections, in whom weak immunity and persistent virus replication result in the selection of immune escape variants.[18,19] Many RBD amino acid changes present in WI-CL-001 have eventually appeared in omicron variants circulating in human populations. In the RBD region of the spike gene, Arg346Thr, Val445Pro, Leu452Gln, Leu452Arg, Asn460-Lys, Phe486Val, and Phe486Pro emerged in circulating omicron variants globally between January, 2022, and January, 2023. Some of these spike mutations—specifically Arg346Thr, Val445Pro, and Asn460lys—emerged in WI-CL-001 5–6 months before becoming highly prevalent globally, largely associated with the spread of BQ.1.1* and XBB.1.5. In WI-CL-001, a Phe486Ala substitution in the spike protein appeared approximately 4 months before the rise of the Phe486Val substitution (found in BA.5*/BQ.1* variants) and 10 months before the rise of the Phe486Pro substitution (found in XBB.1.5*) at the same spike residue. The RBD mutations Tyr453Phe and Val483Ala were detected in WI-CL-001, respectively, from January and February, 2022, until January, 2023, but were found in less than 1% of global sequences during that same time.[23] We therefore speculate that those two substitutions, or other mutations at these sites, might become more prevalent in circulating viruses in the future. Since we made this prediction in May, 2023, the mutation Val483Δ has arisen at the same residue in BA.2.86.[26]

In addition to the divergent spike, there was a cluster of nearly fixed variants in the region that encodes the ectodomain of the viral membrane protein. The mutation cluster includes a 15-nucleotide insertion (5′-GCAACAACTCAGAGT-3′) that encodes the amino acids SerAsnAsnSerGluPhe by splitting the A and TT of an existing Ile codon. Interestingly, the insertion is identical to the sequence found between positions 11 893 and 11 907 in ORF1AB, which suggests an intramolecular recombination event. Additionally, WI-CL-001 has Ala2Glu, Gly6Cys, and Leu17Val amino acid substitutions in the membrane protein.

One possible explanation for these substitutions is that they could confer escape from membrane protein-directed antibodies. This region of the membrane protein is exposed outside of the SARS-CoV-2 virion and is a known target for binding antibodies.[27,28] A previous study using linear peptide binding arrays found that antibody binding on the membrane protein ectodomain was the most intense, on average, of all epitopes in the SARS-CoV-2 proteome.[28] Together these observations suggest that the accumulation of mutations in WI-CL-001, particularly in the spike protein, is the result of adaptive evolution.

WI-CL-001, BA.1, and BA.2.86 are all spike saltations (ie, evolutionary jumps involving the concurrent appearance of a large number of substitutions). These viruses share two common features; namely, in-frame insertions of multiple amino acids in regions frequently recognised by antibodies, and an extraordinary excess of non-synonymous substitutions in the spike protein.[8,26] BA.1 and BA.2.86 have insertions in the spike N-terminal domain that are reminiscent of the membrane protein ectodomain insertion observed in WI-CL-001. The characteristic BA.2.86 insertion of TCATGCCGCTGT is, like WI-CL-001's membrane protein insertion, an apparent intramolecular recombination that is identical to a nucleotide sequence from ORF1AB (17 166–77). The excess of non-synonymous variants and near absence of synonymous variants in the spike protein for these three lineages suggests that they have each evolved under diversifying selection. Although the original sources of BA.1 and BA.2.86 will never be known definitively, a leading hypothesis for the origin of some divergent SARS-CoV-2 variants is that they arise in immunocompromised individuals with prolonged infections.[18,19]

Accordingly, more frequent global wastewater viral surveillance and sequencing of catchment areas would probably detect more examples of cryptic SARS-CoV-2 lineages. Traceback studies such as these, if conducted within appropriate ethical and privacy constraints, might be useful for determining the origins, as well as epidemiological and clinical significance, of these lineages. We speculate that omicron-derived cryptic lineages will be detectable in wastewater in the future. Given the extensive spread of omicron, we expect the number of prolonged infections that give rise to these cryptic lineages to increase, making the emergence of cryptic lineages more common. Although RBD sequencing covers only a small segment of the SARS-CoV-2 genome, we believe this method will continue to be valuable in wastewater surveillance due to its high sensitivity. Cryptic wastewater lineages like WI-CL-001 might have the potential for wider community transmission. And regardless of this possibility, the fact that these lineages frequently exhibit specific mutations, or changes at specific sites, that are later found in circulating variants could be used to aid in forecasting the future evolutionary trajectory of SARS-CoV-2. Such a forecast would be useful in evaluating the cross-protection of existing and future vaccines and treatments. In the present, wastewater sequencing surveillance has become a valuable approach for tracking the emergence of novel SARS-CoV-2 variants in the context of waning clinical sequencing and otherwise-unsampled prolonged infections.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was made possible by the generous support of the Rockefeller Foundation's Regional Accelerators for Genomics Surveillance (DHO and TCF), WDHS Epidemiology and Laboratory Capacity funds (144 AAJ8216) to DHO, CDC contract 75D30121C11060 (DHO and TCF), WDHS ELC Wastewater Surveillance funds (130:AAI8627) to the University of Wisconsin-Madison Wisconsin State Laboratory of Hygiene, and National Institute on Drug Abuse contract 1U01DA053893-01 (MCJ), and the Center for Research on Influenza Pathogenesis and Transmission (75N93021C00014) from the National Institutes of Allergy and Infectious Diseases to YK. The authors thank Roger Wiseman, Nick Minor, David Baker, and CDC SPHERES for helpful discussions. The authors also thank Sarah Abu Kamal, Maansi Bhasin, Sydney Wolf, and Aanya Virdi for help with sequence generation and data organisation. They also acknowledge and thank the wastewater engineers from the city wastewater utility for their sewershed sampling prowess. Additional thanks to Katia Koelle and Michael Martin of Emory University for helpful discussions on the quantitative analysis of viral evolution. Adrian Creanga (National Institutes of Health) provided Vero E6-TMPRSS2/hACE2 cells. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC.

Declaration of interests

YK has received unrelated funding support from Daiichi Sankyo Pharmaceutical, Toyama Chemical, Tauns Laboratories, Shionogi, Otsuka Pharmaceutical, KM Biologics, Kyoritsu Seiyaku, Shinya Corporation, and Fuji Rebio. DHO, MCJ, and TCF have a provisional patent (US patent application number 63/394,159, Method for selecting antigenic viral vaccine and therapeutic sequences) that describes the use of cryptic lineage sequences in variant forecasting. All other authors declare no competing interests.

Data sharing

Sequencing data are available in National Center for Biotechnology Information SRA and GenBank. Additional data are available from https://go.wisc.edu/4134pl. All sequences used for the phylogenetic inferences shown in figure 4 were obtained from GenBank and can be accessed using the accession numbers available on the GitHub repository accompanying this Article, which also contains the scripts used to analyse them (https://github.com/tcflab/wisconsin_cryptic_lineages).

References

1. Xiao F, Tang M, Zheng X, Liu Y, Li X, Shan H. Evidence for gastrointestinal infection of SARS-CoV-2. Gastroenterology 2020; 158: 1831–33. [PubMed: 32142773]
2. Anjos D, Fiaccadori FS, Servian C do P, et al. SARS-CoV-2 loads in urine, sera and stool specimens in association with clinical features of COVID-19 patients. J Clin Virol 2022; 2: 100059.
3. Ahmed W, Tscharke B, Bertsch PM, 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; 761: 144216. [PubMed: 33360129]
4. Vo V, Harrington A, Afzal S, et al. Identification of a rare SARS-CoV-2 XL hybrid variant in wastewater and the subsequent discovery of two infected individuals in Nevada. Sci Total Environ 2023; 858: 160024. [PubMed: 36356728]
5. Smyth DS, Trujillo M, Gregory DA, et al. Tracking cryptic SARS-CoV-2 lineages detected in NYC wastewater. Nat Commun 2022; 13: 635. [PubMed: 35115523]
6. Gregory DA, Trujillo M, Rushford C, et al. Genetic diversity and evolutionary convergence of cryptic SARS-CoV-2 lineages detected via wastewater sequencing. PLoS Pathog 2022; 18: e1010636. [PubMed: 36240259]
7. Viana R, Moyo S, Amoako DG, et al. Rapid epidemic expansion of the SARS-CoV-2 omicron variant in southern Africa. Nature 2022; 603: 679–86. [PubMed: 35042229]
8. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 2020; 581: 215–20. [PubMed: 32225176]
9. covlineages.org. BA.1 2021-12-09. https://cov-lineages.org/global_report_BA.1.html (accessed July 27, 2022).

10. Fritz M, Rosolen B, Krafft E, et al. High prevalence of SARS-CoV-2 antibodies in pets from COVID-19+ households. One Health 2021; 11: 100192. [PubMed: 33169106]
11. Hale VL, Dennis PM, McBride DS, et al. SARS-CoV-2 infection in free-ranging white-tailed deer. Nature 2022; 602: 481–86. [PubMed: 34942632]
12. Lu L, Sikkema RS, Velkers FC, et al. Adaptation, spread and transmission of SARS-CoV-2 in farmed minks and associated humans in the Netherlands. Nat Commun 2021; 12: 6802. [PubMed: 34815406]
13. Pickering B, Lung O, Maguire F, et al. Divergent SARS-CoV-2 variant emerges in white-tailed deer with deer-to-human transmission. Nat Microbiol 2022; 7: 2011–24. [PubMed: 36357713]
14. Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. Science 2021; 371: 172–77. [PubMed: 33172935]
15. Domanska-Blicharz K, Oude Munnink BB, Orłowska A, et al. Cryptic SARS-CoV-2 lineage identified on two mink farms as a possible result of long-term undetected circulation in an unknown animal reservoir, Poland, November 2022 to January 2023. Euro Surveill 2023; 28: 2300188.
16. Pérez-Cataluña A, Chiner-Oms Á, Cuevas-Ferrando E, et al. Spatial and temporal distribution of SARS-CoV-2 diversity circulating in wastewater. Water Res 2022; 211: 118007. [PubMed: 35033744]
17. Natarajan A, Zlitni S, Brooks EF, et al. Gastrointestinal symptoms and fecal shedding of SARS-CoV-2 RNA suggest prolonged gastrointestinal infection. Med 2022; 3: 371–87. [PubMed: 35434682]
18. Corey L, Beyrer C, Cohen MS, Michael NL, Bedford T, Rolland M. SARS-CoV-2 variants in patients with immunosuppression. N Engl J Med 2021; 385: 562–66. [PubMed: 34347959]
19. Wilkinson SAJ, Richter A, Casey A, et al. Recurrent SARS-CoV-2 mutations in immunodeficient patients. Virus Evol 2022; 8: veac050. [PubMed: 35996593]
20. Patel H, Varona S, Monzón S, et al. nf-core/viralrecon: nf-core/viralrecon v2·5 - Manganese Monkey. July 13, 2022. https://zenodo.org/records/6827984 (accessed Feb 14, 2023).
21. 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; 13: 19.
22. Karthikeyan S, Levy JI, De Hoff P, et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature 2022; 609: 101–08. [PubMed: 35798029]
23. outbreak.info. SARS-CoV-2 data explorer. https://outbreak.info/ (accessed Sept 8, 2022).
24. Chen C, Nadeau S, Yared M, et al. CoV-spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics 2022; 38: 1735–37. [PubMed: 34954792]
25. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw 2021; 6: 3773.
26. Rasmussen M, Møller FT, Gunalan V, et al. First cases of SARS-CoV-2 BA.2.86 in Denmark, 2023. Euro Surveill 2023; 28: 2300460.
27. Jörrißen P, Schütz P, Weiand M, et al. Antibody response to SARS-CoV-2 membrane protein in patients of the acute and convalescent phase of COVID-19. Front Immunol 2021; 12: 679841. [PubMed: 34421894]
28. Heffron AS, McIlwain SJ, Amjadi MF, et al. The landscape of antibody binding in SARS-CoV-2 infection. PLoS Biol 2021; 19: e3001265. [PubMed: 34143766]

Research in context

Evidence before this study

To identify other studies that characterised unusual wastewater-specific SARS-CoV-2 lineages, we conducted a PubMed search of the English-language literature using the keywords ((cryptic SARS-CoV-2 lineages) OR (novel SARS-CoV-2 lineages)) AND (wastewater)) on May 9, 2023. From the 18 full-text articles retrieved, only two reported wastewater-specific cryptic lineages. These lineages were identified by members of our author team in wastewater from California, Missouri, and New York City (USA). None of these could be definitively traced to a specific source. A third study in Nevada identified a unique recombinant variant (designated Pango lineage XL) in wastewater, which was also discovered in two clinical specimens from the same community. However, it was unclear whether the clinical specimens collected were from the same individual or individuals responsible for the virus detected in the wastewater. To our knowledge, no previous study has successfully traced novel SARS-CoV-2 lineages detected in wastewater back to a specific location. How and where cryptic lineages are introduced into wastewater is not known.

Added value of this study

This study documents the presence and probable source of a novel and highly divergent cryptic SARS-CoV-2 lineage detected in Wisconsin (USA) wastewater for 13 months. In contrast to previously reported cryptic lineages, we successfully traced the lineage (WI-CL-001) to a single commercial building with approximately 30 employees. The exceptionally high viral RNA concentrations at the source building facilitated the tracing effort and allowed for the sequencing of WI-CL-001's whole genome, expanding our view of the lineage's mutational landscape beyond the spike gene.

Implications of all the available evidence

WI-CL-001's persistence in wastewater, its heavily mutated omicron-like genotype, and its identified point source at a human-occupied commercial building all support the hypothesis that cryptic wastewater lineages can arise from persistently infected humans. Because cryptic wastewater lineages have some amino acid changes that subsequently emerge in circulating viruses, increased global monitoring of such lineages could help forecast variants that might arise in the future.

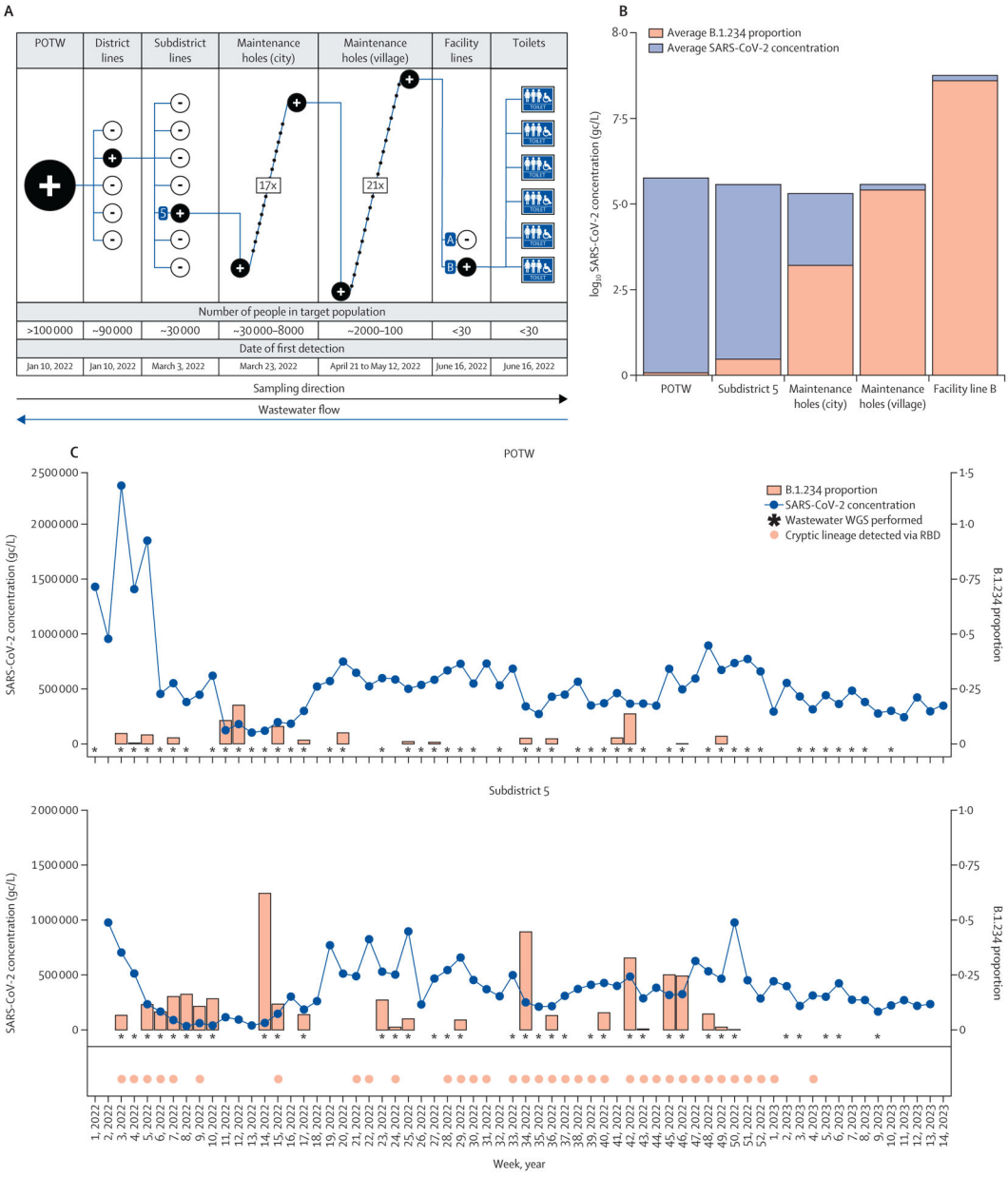

Figure 1: Tracing the source of the cryptic SARS-CoV-2 lineage

(A) After initial detection at a POTW the source of the WI-CL-001 signal was narrowed to positive district and subdistrict lines. Within the positive subdistrict (subdistrict 5) the signal was further traced to a single commercial building via city and village maintenance hole sampling. Of two collecting lines at the identified source facility (line A and line B) only line B, which serves six toilets, was positive for WI-CL-001. All location names are pseudonyms and all population values are approximate to conceal the identity of the sampled sewershed. (B) Mean SARS-CoV-2 RNA concentrations (on a $\log_{10}$ scale) obtained by RT-dPCR detected in five sampling areas. WI-CL-001's contribution (B.1.234 proportion as estimated by Freyja within WGS data) to the total mean SARS-CoV-2 concentrations at each sampling level is shown in tan. (C) SARS-CoV-2 RNA concentrations measured

by RT-dPCR throughout 2022 for the POTW and subdistrict 5 are shown as a blue line. The percent contribution of WI-CL-001 (B.1.234 proportion in WGS data) is shown as tan bars. Asterisks mark the weeks in which samples underwent WGS, regardless of whether or not WI-CL-001was detected in those sequences. Dates WI-CL-001 was detected in subdistrict 5 wastewater samples via targeted RBD amplicon sequencing are marked by tan circles. Weeks without tan circles were either negative or not tested by RBD sequencing. gc=genome copies. POTW=publicly owned treatment works facility. RBD=receptor binding domain. WGS=whole-genome sequencing.

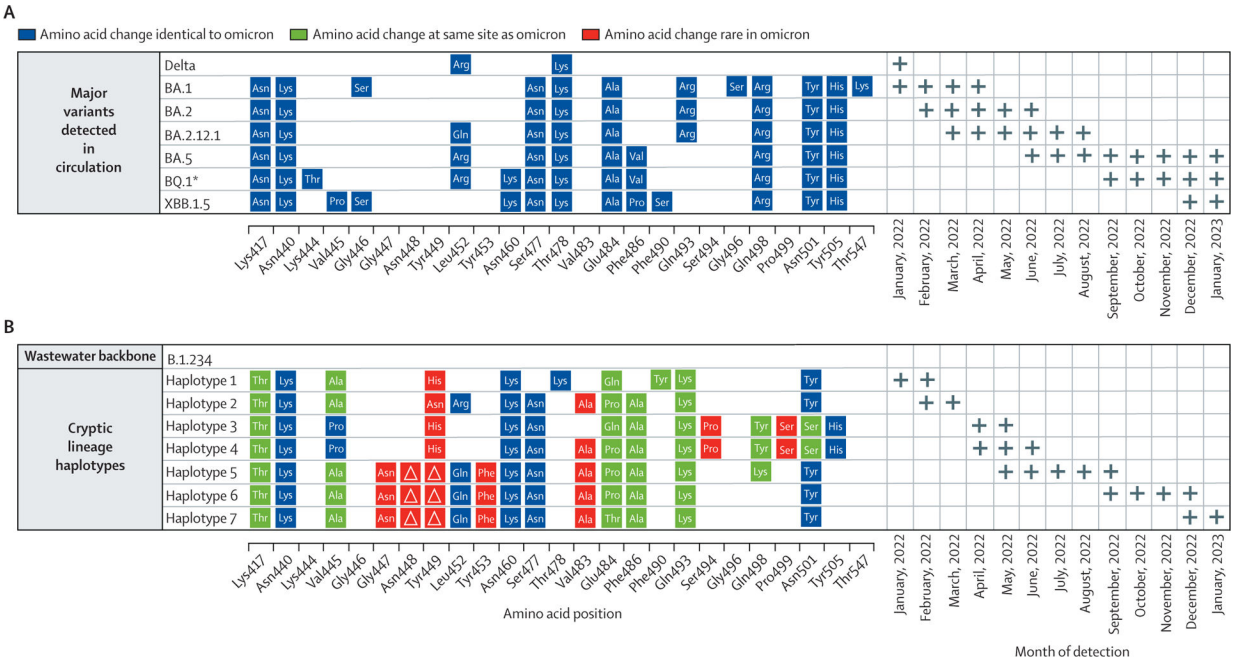

Figure 2: Representative haplotypes of WI-CL-001 sequences

(A) Delta and omicron amino acid changes (blue) represent characteristic changes relative to the Wuhan-Hu-1 (MN908947.3) reference seen in at least 90% of all US sequences of the specified sublineage.[23] Plus signs on the right-hand side of the panel denote months in which these lineages were identified in at least 1% of all US clinical sequences.

(B) Representative haplotypes detected from subdistrict 5 using non-omicron spike RBD sequencing, each of which were at least 25% of the total reads in at least one sample. Amino acid changes are relative to the Wuhan-Hu-1 reference. Green boxes indicate amino acid sites that are also altered in major omicron lineages, blue boxes indicate amino acid sites that have mutations identical to major omicron lineages, and red boxes indicate amino acid sites that are altered in WI-CL-001 but altered in less than 0·1% of omicron sequences.[24] Plus signs on the right-hand side of the panel indicate months in which these haplotypes were detected.

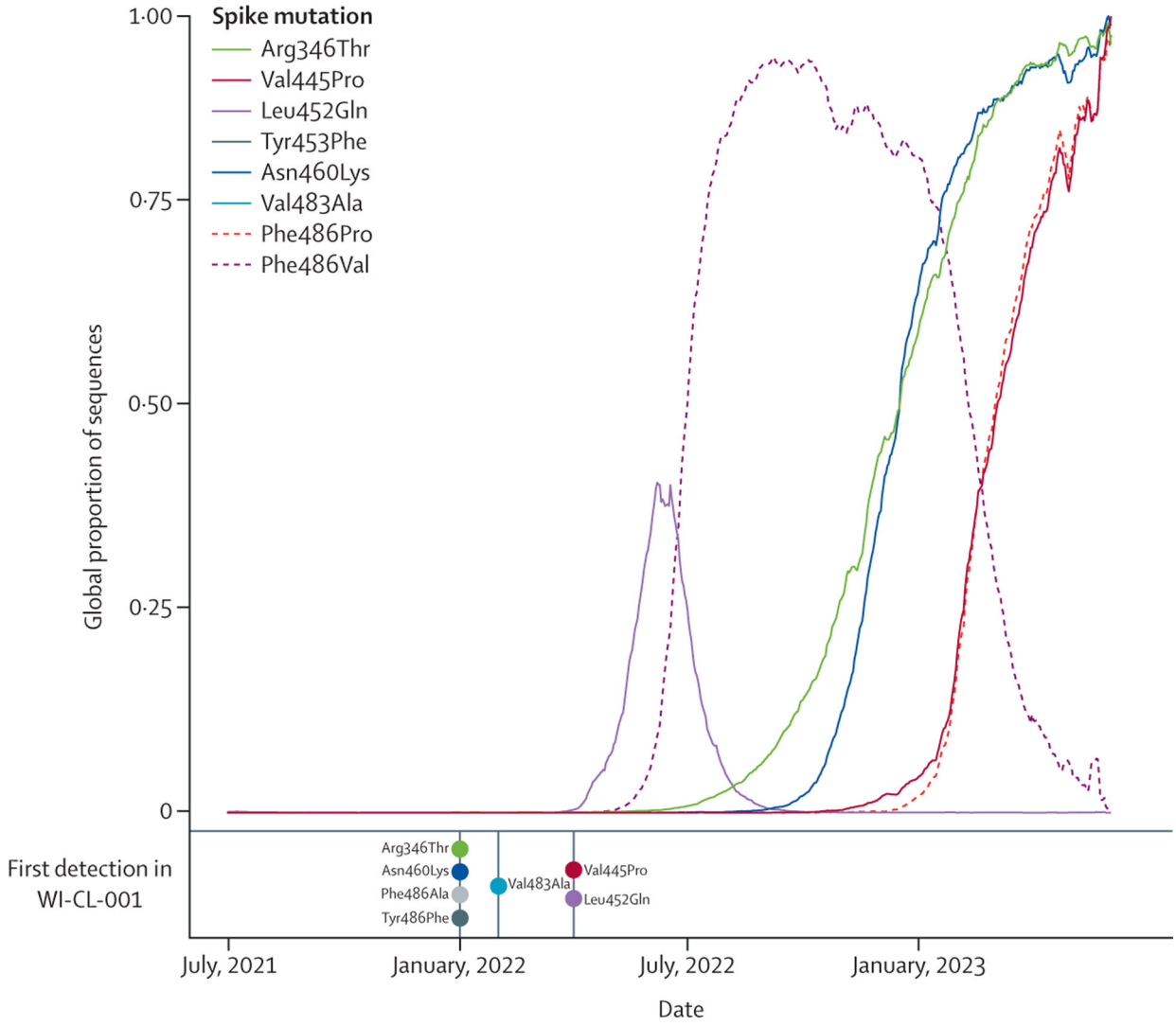

Figure 3: Prevalence of key cryptic lineage mutations in global sequences

The global proportions of sequences uploaded to National Center for Biotechnology Information GenBank between June 31, 2021, and May 1, 2023, for key mutations in the spike gene of the Wisconsin wastewater lineage over time.[23] The spike mutations Arg346Thr, Val445Pro, and Asn460Lys were all detected in the Wisconsin cryptic lineage months before becoming predominant in global sequences. WI-CL-001 also harboured Phe486Ala from the time of initial detection in January, 2022. Two other substitutions at spike amino acid residue 486 have since become dominant in global sequences (dotted lines).

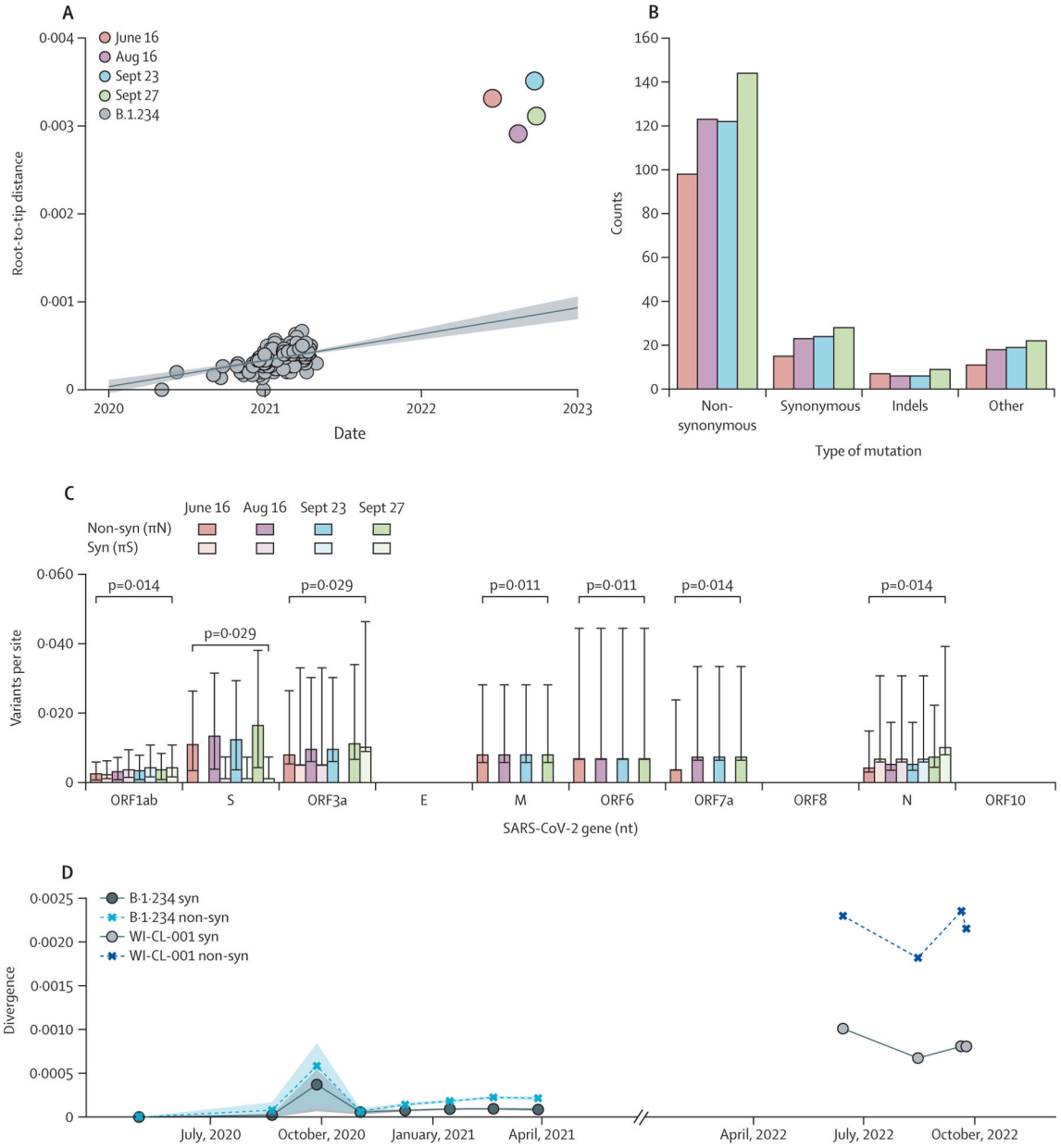

Figure 4: Analysis of wastewater genomic sequences from all facility line B timepoints

(A) Root-to-tip regression analysis (distance) of B.1.234 sequences via Tree Time based on a maximum likelihood phylogenetic tree inferred with iqtree (not shown) and alignedto the Wuhan-Hu-1 (MN908947.3) reference sequence. The mean nucleotide substitution rate for B.1.234 sequences was estimated to be $4{\cdot}24 \times 10^{-4}$ substitutions per site per year (SD 7×10^{-5}). (B) Number of intra-sample single nucleotide polymorphisms (y-axis) for the wastewater timepoints for each mutation type relative to the reference MN908947.3. Mutations were classified as non-synonymous, synonymous, indels, or others (including nonsense and frameshift mutations outside of coding regions). (C) Within-sample pairwise sequence diversity assessed for non-synonymous sites (πN) and synonymous sites (πS) in each SARS-CoV-2 gene at each timepoint. p values for differences between πN and πS on

each gene are marked closer to the bars; p values for enrichment of the πN value of spike against the πN value of the other gene are marked further from the bars. (D) The divergence (Hamming distance; y axis) between B.1.234 isolates from panel A and the MN908947.3 reference sequence over a sliding window of 36 days (x-axis) compared with WI-CL-001 isolates. Except for WI-CL-001, data are only plotted when windows contain at least two B.1.234 sequences. Indels=insertions-deletions. WW=wastewater.