medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Tracing the origin of SARS-CoV-2 Omicron-like Spike sequences detected in wastewater

Martin M. Shafer*, Max J. Bobholz*, William C. Vuyk*, Devon Gregory*, Adelaide Roguet, Luis A. Haddock Soto, Clayton Rushford, Kayley H. Janssen, 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, Ian W. Pray, Ryan Westergaard, David H. O'Connor*, Marc C. Johnson*
* Contributed equally

Corresponding Author

Marc C. Johnson
Phone: 573-882-1519
marcjohnson@missouri.edu
471c Bond Life Sciences Center
1201 Rollins St
Columbia, MO USA 65211

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Abstract

Background: The origin of divergent SARS-CoV-2 spike sequences found in wastewater, but not in clinical surveillance, remains unclear. These "cryptic" wastewater sequences have harbored many of the same mutations that later emerged in Omicron lineages. We first detected a cryptic lineage in municipal wastewater in Wisconsin in January 2022. Named the "Wisconsin Lineage", we sought to determine this virus's geographic origin and characterize its persistence and evolution over time.

Methods: We systematically sampled maintenance holes to trace the Wisconsin Lineage's origin. We sequenced spike RBD domains, and where possible, whole viral genomes, to characterize the evolution of this lineage over the 13 consecutive months that it was detectable.

Findings: The persistence of the Wisconsin Lineage signal allowed us to trace it from a central wastewater plant to a single facility, with a high concentration of viral RNA. The viral sequences contained a combination of fixed nucleotide substitutions characteristic of Pango lineage B.1.234, which circulated in Wisconsin at low levels from October 2020 to February 2021, while mutations in the spike gene resembled those subsequently found in Omicron variants.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Interpretation: We propose that prolonged detection of the Wisconsin Lineage in wastewater represents persistent shedding of SARS-CoV-2 from an infected individual, with ongoing within-host viral evolution leading to an ancestral B.1.234 virus accumulating “Omicron-like” mutations.

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

Research in context

Evidence before this study: To identify other studies characterizing unusual wastewater-specific SARS-CoV-2 lineages, we conducted a PubMed search using the keywords "cryptic SARS-CoV-2 lineages," "novel SARS-CoV-2 lineages," and "wastewater" on May 9, 2023. From the 18 papers retrieved, only two reported detecting wastewater-specific cryptic lineages. These lineages were identified by investigators from this research team in wastewater from California, Missouri, and New York City, but none could be definitively traced to a specific origin.

A third study in Nevada identified a unique recombinant variant (designated Pango lineage XL) in wastewater, which was also discovered in two clinical samples from the same community. However, it remained unclear whether the clinical samples were collected from the same individual(s) responsible for the virus detected in the wastewater. To our knowledge, no prior study has successfully traced cryptic SARS-CoV-2 wastewater lineages back to a specific location. How and where evolutionarily advanced cryptic lineages are introduced into wastewater has remained uncertain.

Added value of this study: This study documents the presence and likely source of a novel and highly divergent cryptic SARS-CoV-2 lineage detected in Wisconsin wastewater for 13 months. In contrast to previously reported cryptic lineages, we successfully traced the Wisconsin Lineage to a single facility serving about 30 people. The exceptionally high viral RNA loads at the source account for the detection of these sequences in wastewater collected at the downstream municipal wastewater treatment facility.

Implications of all the available evidence: Many SARS-CoV-2 lineages have been found exclusively in wastewater. The high number of unusual mutations found in these wastewater-specific cryptic sequences raises the possibility that they originate from individual prolonged shedders or even non-human sources. The Wisconsin Lineage's persistence in wastewater, single-facility origin, and heavily

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

mutated Omicron-like genotype support the hypothesis that cryptic wastewater lineages arise from persistently infected humans. As these divergent sequences contain amino acid changes that eventually emerge in circulating viruses, increased global monitoring of such lineages in wastewater could help anticipate future circulating mutations and/or variants of concern.

Introduction

SARS-CoV-2-infected hosts shed viral RNA in their stool and urine. Furthermore, the virus is known to infect the gastrointestinal (GI) tract and kidney tissues.[1-2] Owing to these features, and the relative stability/persistence of SARS-CoV-2 in wastewater, its RNA is readily detected in wastewater samples.[3] Wastewater surveillance has accordingly become an important complement to clinical testing in monitoring SARS-CoV-2. In some cases, however, wastewater surveillance has also identified unique lineages that clinical nasal swab testing does not. These wastewater-specific "cryptic" lineages are often highly divergent from their pre-Omicron ancestors, showing a level of mutation more similar to Omicron lineages.[4]

Compared to SARS-CoV-2 viruses that circulated in 2020, the Omicron lineage that first emerged in South Africa in November 2021 had a highly divergent spike receptor binding domain (RBD), with 10 lineage-defining amino acid substitutions between residues 412 and 579 (K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, and N501Y).[5–7] Though global circulation of Omicron did not begin until late 2021, 8 of these polymorphisms were observed in RNA from New York City wastewater samples collected in the first half of 2021. One wastewater sample, collected in May 2021, had more extensive mutation than all Omicron lineages known as of March 2022, with one haplotype containing 24 amino acid substitutions between spike residues 412 and 579. Subsequent testing of wastewater samples from Missouri and California resulted in sporadic detection of additional cryptic lineages.[4] It is not clear how or where these divergent SARS-CoV-2 sequences are being introduced into wastewater, or why they share many similarities with distantly related Omicron viruses.

There are two leading hypotheses for the source of these cryptic sequences. First, an un- or under-sampled animal reservoir may be introducing these viruses into wastewater, e.g., through defecation in combined sewer systems, via inflow/infiltration, or from livestock processing waste. SARS-CoV-2 exhibits a broad host range, including infection of household pets.[8] Multiple studies have detected SARS-CoV-2 virus and antibodies in wild white-tailed deer populations around North America.[9,10] Circulation in deer may exert different selection pressures on SARS-CoV-2 than circulation

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

in human populations, resulting in evolution of divergent sequences.[11] SARS-CoV-2 has also caused large-scale outbreaks in farmed mink, some of which have included transmission of novel viral variants back to humans.[12–14] A divergent cryptic SARS-CoV-2 variant recently found in farmed mink may have resulted from epizootic spillover from an unknown animal reservoir into mink.[15] Thus, it is plausible that wild animal reservoirs of cryptic lineages exist undetected, with ongoing virus transmission and exchange between animal species. [9]

The second hypothesis is that cryptic SARS-CoV-2 lineages in wastewater are derived from people with unsampled infections, as has been suggested for less divergent wastewater sequences.[16] Individuals without a detectable SARS-CoV-2 infection in the upper respiratory tract could still have prolonged infections in the GI tract or kidneys that will shed virus into wastewater, thus such individuals could therefore be a significant subgroup of unsampled infections. Persons with immunocompromising 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.[17,18] Such selection could account for the observation that cryptic lineages tend to accumulate high levels of nonsynonymous variation in spike, upon the genetic background of viruses that are no longer common in circulation at the time the cryptic variants are sampled. Notably, human waste (stool and urine) is the predominant source of genetic material in wastewater, making it the most parsimonious explanation for the origin of SARS-CoV-2 sequences detected in wastewater sampling.

Methods

Isolation of viral RNA from wastewater

Wastewater samples were shared between the Wisconsin State Laboratory of Hygiene (WSLH) and the University of Missouri, with the WSLH focusing on virus quantitation and whole genome sequencing, and the University of Missouri focusing on RBD-targeted sequencing. At the WSLH, after the addition of a bovine coronavirus (BCoV) viral recovery control and concentration of virus using Nanotrap Magnetic Virus Particles (Ceres Nanosciences, VA, USA) on a Kingfisher Apex instrument (ThermoFisher Scientific, MA, USA), total nucleic acids were extracted using Maxwell(R) HT Environmental TNA kits (Promega, Madison, WI, USA) on a Kingfisher Flex instrument (ThermoFisher Scientific, Waltham, MA, USA). The University of Missouri concentrated the virus using a PEG protocol on pre-filtered samples (0.22 μM polyethersulfone membrane (Millipore, Burlington, MA, USA)).

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Samples were incubated with PEG (polyethylene glycol 8000) and 1.2 M NaCl, centrifuged, and the RNA was isolated from the pellet with the QIAamp Viral RNA Mini Kit (Qiagen, Germantown, MD, USA). Further information on these procedures can be found in the Supplemental Methods (pages 1-2).

Quantification of viral RNA by RT-dPCR

The WSLH quantified the concentration of SARS-CoV-2, BCoV (viral recovery control), and PMMoV (fecal marker) in each sample using reverse transcriptase digital PCR (RT-dPCR). PCR inhibition was probed with a bovine respiratory syncytial virus (BRSV) spiked into each PCR reaction. A more detailed protocol is provided in the Supplemental Methods (pages 2-3).

Identification of cryptic lineages in wastewater with non-Omicron PCR amplification and amplicon sequencing

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 and analyzed using the SAMRefiner software.[19] The Wisconsin Lineage's unique RBD sequences were used to identify and track the lineage across time and space. Additional details are provided in the Supplemental Methods (pages 3-4).

SARS-CoV-2 whole genome sequencing of wastewater

For SARS-CoV-2 whole genome sequencing (WGS), 13 µL of total nucleic acids from the 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 analyzed by a QIAxcel Advanced and the QX DNA Screening Kit (QIAGEN, Germantown, MD, USA). Sequencing was performed on an Illumina MiSeq instrument using MiSeq Reagent v2 (300 cycles) kits.

Fastq files were analyzed with the nf-core/viralrecon 2.5 workflow[20] (10.5281/zenodo.3901628) using the SARS-CoV-2 Wuhan-Hu-1 reference genome (Genbank accession MN908947.3) The workflow was initiated as outlined on the project's data portal (https://go.wisc.edu/4134pl). Additional details are provided in the Supplemental Methods (pages 4-5).

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Additional Methods

Additional information regarding our protocols for wastewater collection, virus culture, variant proportion assessments, root-to-tip regressions, and natural selection analyses can be found in the Supplemental Methods.

Role of the funding source

This manuscript underwent CDC's clearance review process due to the involvement of CDC co-authors. However, the funders did not play a direct role in the study design, data collection, data analysis, or manuscript preparation.

Results

On January 11, 2022, a cryptic lineage containing at least six unusual spike RBD substitutions was first detected in a composite wastewater sample from a metropolitan publicly owned treatment works (POTW) in Wisconsin (Figure 1c). This initial sample was composited from raw influent from five interceptor districts in the metropolitan area sewershed, effectively sampling a population of more than 100,000 people. The source of the enigmatic RBD sequences was narrowed by testing each of the five interceptor district lines. The Wisconsin Lineage was only detected in one district line, which served seven sub-districts. Of the seven sub-district lines, only the line from Sub-District 5 contained the target (Figure 1c). Additional testing in March 2022 of manholes upstream of the interceptor line within Sub-District 5 confirmed the persistence of the Wisconsin Lineage and further refined the lineage's source. As the sampling effort progressed further upstream in the sewershed, the proportion of the Wisconsin Lineage (labeled B.1.234 in Figure 1), determined using Freyja[21] v.1.3.11 (see page 13 of Supplemental Methods), increased relative to the total SARS-CoV-2 sequences detected at each sampling site (Figure 1b). By May 2022, this investigation had traced the source of the lineage to a single manhole accessing a lateral that collected wastewater from a single building. Subsequent testing of wastewater from sewer service lines within this building in June 2022 further narrowed the source to one sewer line serving six toilets (called "Facility Line B") on one side of the building (Figure 1a). 12s rRNA sequencing detected predominantly human rRNA from this source, supporting the hypothesis that this cryptic lineage was being shed by a human. Chicken rRNA, the next largest taxon identified, was less than 0.05% of the sample (Supplemental Table 1). Facility Line B was retested for the cryptic lineage in August and again in September of 2022, remaining positive at each time point.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

As quantified by digital PCR, unprecedentedly high wastewater SARS-CoV-2 RNA viral loads were observed in samples collected from Facility Line B on June 16th (~520,000,000 genome copies per liter undiluted wastewater), August 16th (~1,600,000,000 copies per liter), September 23rd (~2,700,000,000 copies per liter) and September 27th (~550,000,000 copies per liter) (Figure 1b). Drops of this raw wastewater tested positive for SARS-CoV-2 in a lateral flow antigen test. Despite these high viral loads, viable virus could not be cultured from wastewater after multiple attempts.

The extremely high levels of viral RNA in Facility Line B allowed us to amplify and sequence entire viral genomes to shed further light on the origins and evolution of this unusual lineage. We generated whole genome sequences from the Facility Line B samples taken on June 16th, August 16th, September 23rd and September 27th 2022. All of these sequences were classified as lineage B.1.234 by Pangolin. In SARS-CoV-2 genomic surveillance using clinical specimens, B.1.234 viruses were first detected in Wisconsin on 2 September 2020, and were last detected on 30 March 2021.[22] Combining our observations, we posit that the simplest explanation for the appearance and persistence of the Wisconsin Lineage is that a single individual, originally infected when B.1.234 was in circulation, developed a persistent infection and continued to excrete viruses into wastewater throughout 2022.

While the original B.1.234 lineage does not have any characteristic spike RBD amino acid changes relative to the reference Wuhan-Hu-1, Omicron lineages detected in wastewater concurrently with the Wisconsin Lineage had many (Figure 2a). RBD amplicon sequencing of the Wisconsin Lineage using non-Omicron PCR amplification detected 29 RBD changes at a frequency of at least 25%, and 43 more at a frequency of at least 10%. Sequencing single amplicons that span the RBD allowed us to define haplotypes, i.e., specific combinations of mutations found together in a single RNA molecule. We repeatedly sequenced spike RBD in wastewater samples from the Sub-District 5 interceptor line, and haplotypes of the Wisconsin Lineage were detected every month from January 2022 to January 2023 (Figure 2b, Supplemental Figure 1). In all, we detected 87 RBD haplotypes between Jan 2022 and Jan 2023, but the mean number of haplotypes detected at any one time point was 2 (range, 1-6) (Supplemental Figure 1). The signal became undetectable in January 2023. Of the RDB amino acid changes with a frequency of at least 25%, 11 of these were at the same site as Omicron RBD changes, 9 were identical to Omicron, and 9 are absent from known Omicron lineages to date (Figure 2a,b). Some of these exact amino acid changes, or different changes at the same positions, were initially

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

detected in the Wisconsin Lineage months before becoming predominant in globally circulating dominant Omicron lineages (Figure 3).

The cryptic lineage is also highly divergent outside of the spike RBD. When plotted on a radial phylogenetic tree using Nextclade, our Illumina whole genome consensus sequences from Facility Line B show a similar degree of divergence from Wuhan-Hu-1 to 22B clade and XBB* Omicron lineages (Supplemental Figure 2). To investigate this further, we used iVar within the nf-core/viralrecon workflow to call variants at ≥25% frequency from the Wuhan-Hu-1 reference and turned that output into a pivot table (Supplemental Table 2). From this pivot table, we identified which codons had variants detected at every time point in at least one whole genome sequence replicate. One of the Wisconsin Lineage's most characteristic (and peculiar) mutations is in the N-terminal domain (NTD) of the membrane protein, where there is a 15 nucleotide insertion (I8delinsSNNSEF) found at an average frequency of 92.4% in all sequences (Supplemental Table 3).

We next asked whether the unusual combinations of mutations present in the Wisconsin Lineage could be the result of natural selection favoring nonsynonymous (amino-acid-changing) mutations. First, we calculated the nucleotide substitution rate that prevailed when B.1.234 viruses were circulating in the US Midwest (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin) and found that mutations accumulated in the Wisconsin Lineage faster than expected based on this rate (Figure 4a). Across the four timepoints with available genome sequences, there is a notable excess of nonsynonymous nucleotide substitutions (mean 121.8 ± 16.3) relative to synonymous ones (mean 22.5 ± 4.7) (Figure 4b). Some previous reports have suggested that mutations mediated by APOBEC cytidine deaminases can lead to a dramatic excess of C-to-T substitutions within SARS-CoV-2-infected hosts.[23,24] In our sequences, C to T transitions were the most common mutation type, but did not dramatically outnumber other types (Figure 4c). To further characterize genetic diversity within each sample, we used the summary statistic π, which quantifies the number of pairwise differences per nonsynonymous (πN) and synonymous (πS) site within a set of sequences. Within the spike gene, πN was significantly greater than πS at each timepoint, which could indicate ongoing diversifying selection on spike (Figure 4d). Spike also had significantly higher nonsynonymous diversity compared to ORF1ab, ORF3a, M, ORF6, ORF7a, and N at each timepoint (Figure 4d). 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 therefore

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

next calculated divergence, i.e., the average Hamming distance between each sequenced virus (either B.1.234 variants or the Wisconsin Lineages) and the ancestral sequence Wuhan-Hu-1 (MN908947.3; Figure 4e). Both synonymous and nonsynonymous divergence values were substantially higher for the Wisconsin Lineage than for B.1.234 viruses. Notably, nonsynonymous divergence was also dramatically elevated relative to synonymous divergence in the wastewater lineage. Together these observations suggest that the accumulation of mutations in the Wisconsin Lineage, particularly in spike, is the result of adaptive evolution.

Discussion

Here we traced the source of a cryptic SARS-CoV-2 lineage, first detected in wastewater from a central municipal wastewater treatment facility, to a single point source. At the identified point source of this cryptic lineage (Facility Line B), non-human animal sequences made up a minimal proportion of rRNA detected. Thus, the likelihood that this virus was being shed by an otherwise scarce animal reservoir is low, especially given the high viral RNA concentrations of samples from this site. We conclude that the Wisconsin Lineage, the cryptic SARS-CoV-2 lineage we detected in wastewater throughout 2022, was most likely derived from a single human individual with an unusually prolonged infection.

We detected remarkably high levels of SARS-CoV-2 RNA in Facility Line B, the source of the Wisconsin Lineage signal. In one Canadian study, the highest total viral concentrations in municipal wastewater of the three cities tested during the Omicron surge was 3.4 million gene copies/L.[25] Another study reported peak SARS-CoV-2 concentrations of 1.1 million copies/L coming from a single university residence hall with 328 residents.[26] Average N1/N2 SARS-CoV-2 concentrations detected from Facility Line B in this study peaked at 2.7 billion copies/L. This finding may 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? Based on wastewater flow data from the Sub-District 5 interceptor line and estimations of typical toilet use, we would expect the Wisconsin Lineage viral RNA to be diluted from a wastewater volume of approximately 200 gallons at Facility Line B into a volume of 8 million gallons at the Sub-District 5 interceptor. Thus, if there were 2 billion copies/L at Facility Line B, we would expect to detect ~50,000 copies/L at the Sub-District 5 interceptor. This is a comparable concentration, if a little lower, than what we actually observed over 13 months (Figure 1c). We hypothesize that these high levels of viral RNA result from a prolonged infection involving virus replication in the GI and/or urinary tracts, though the extent to which such infections shed

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

more virus into the wastewater than infections where replication primarily occurs in the upper and/or lower respiratory tracts needs to be investigated further.

The large preponderance of nonsynonymous substitutions in the wastewater viral genomes suggests that this virus has undergone diversifying selection on spike, and perhaps on other genes. This 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.[17,18] Many RBD amino acid changes present in the Wisconsin Lineage have eventually appeared in Omicron variants circulating in human populations. In the RBD region of the spike gene, R346T, V445P, L452Q, Y453F, N460K, V483A, and F486A emerged in circulating Omicron variants globally at various times between January and June 2022 (Supplemental Figure 1). Some of these spike mutations, specifically R346T, V445P, and N460K, emerged in the Wisconsin Lineage five to six months before becoming highly prevalent globally (largely associated with the spread of BQ.1.1* and XBB.1.5). In the Wisconsin Lineage, a phenylalanine-to-alanine substitution at spike residue 486 (F486A) appeared approximately four months before the rise of S:F486V (found in BA.5*/BQ.1* variants) and ten months before the rise of S:F486P (found in XBB.1.5*). The RBD mutations V483A and Y453F have been detected since January and February 2022, respectively, in the Wisconsin Lineage but have yet to be significantly detected in any major Omicron lineages (Figure 4). We could therefore speculate that those two substitutions, or other mutations at these sites, may arise in circulating viruses in the future.

In addition to the highly divergent spike, there was a cluster of 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 SNNSEF by splitting the A and TT of an existing isoleucine codon. Interestingly, the insertion is identical to the sequence found between positions 11,893 and 11,907 in ORF1ab, suggesting intramolecular recombination. Additionally, the Wisconsin Lineage has M:A2E, M:G6C, and M:L17V amino acid substitutions. The phenotypic impact of these substitutions, if any, is unclear. One possible explanation is diversifying selection spurred by host immune pressure. 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] In Heffron et al., 2021, membrane-binding antibodies were present at a higher level than anti-spike antibodies.

Our data strongly suggest that SARS-CoV-2 cryptic lineages in wastewater originate from human sources. While animal sources may contribute in other settings, that is not the case here. This has sev-

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

eral important implications. Such lineages likely exist wherever people are infected with SARS-CoV-2, i.e., worldwide. That is, many, perhaps most, divergent SARS-CoV-2 lineages detected in wastewater likely reflect ongoing human infections, and may therefore pose a transmission risk to others. The elevated number of nonsynonymous substitutions in the wastewater variant, and its accumulation of mutations at a rate faster than expected based on its ancestral lineage resemble attributes of the original Omicron lineage when it emerged.[29] Indeed, a leading hypothesis for the origin of many previous SARS-CoV-2 variants of concern is that they arose in immunocompromised individuals with prolonged infections.[17,18] The fact that the Wisconsin Lineage appears to be derived from a prolonged infection with an ancestral B.1.234 virus further highlights the importance of prolonged infections in the emergence of highly divergent viruses and emphasizes the importance of identifying, tracing, and treating such infections.

To this end, more frequent global wastewater viral surveillance/sequencing of catchment areas would likely detect many more examples of cryptic SARS-CoV-2 lineages. Although RBD sequencing covers only a small segment of the SARS-CoV-2 genome, we believe this method to be valuable in wastewater surveillance due to its high sensitivity. Due to the extremely large number of Omicron infections, we speculate that Omicron-derived cryptic lineages will be detectable in wastewater in the future. Given the extensive spread of Omicron, the number of prolonged infections that give rise to these cryptic lineages is also expected to increase, making the emergence and detection of cryptic lineages more common. We note that individuals with immunocompromising conditions are at increased risk for prolonged infections but may not be the only population in which such infections occur; the facility in which the cryptic lineage was detected was an otherwise unremarkable business, not a healthcare facility or other location with medically fragile occupants. SARS-CoV-2 cryptic lineage sequences could aid in forecasting the future evolutionary trajectory of SARS-CoV-2 to evaluate the cross-protection of existing and future vaccines and monoclonal antibodies. In the present, wastewater surveillance has become an irreplaceable window into the progression of the pandemic as clinical sampling wanes and more human-derived cryptic wastewater lineages await detection.

Data availability

Sequencing data are available in NCBI SRA and Genbank. Additional data is available from https://go.wisc.edu/4134pl. All sequences used for the phylogenetic inferences shown in Figure 4 were

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

obtained from GenBank and can be accessed using the accession numbers available on the GitHub repository accompanying this manuscript (https://github.com/tcflab/wisconsin_cryptic_lineages).

Acknowledgments

This study was made possible by the generous support of the Rockefeller Foundation's Regional Accelerators for Genomics Surveillance (DHO/TCF), Wisconsin Department of Health Services Epidemiology and Laboratory Capacity funds (www.dhs.wisconsin.gov, 144 AAJ8216) to DHO, CDC contract 75D30121C11060 (DHO/TCF), Wisconsin Department of Health Services ELC Wastewater Surveillance funds (www.dhs.wisconsin.gov, 130:AAI8627) to the UW-Madison Wisconsin State Laboratory of Hygiene (WSLH), and NIDA contract 1U01DA053893-01 (MJ), 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 Kamal, Maansi Bhasin, Sydney Wolf, and Aanya Virdi for help with sequence generation and data organization. They would also like to 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.

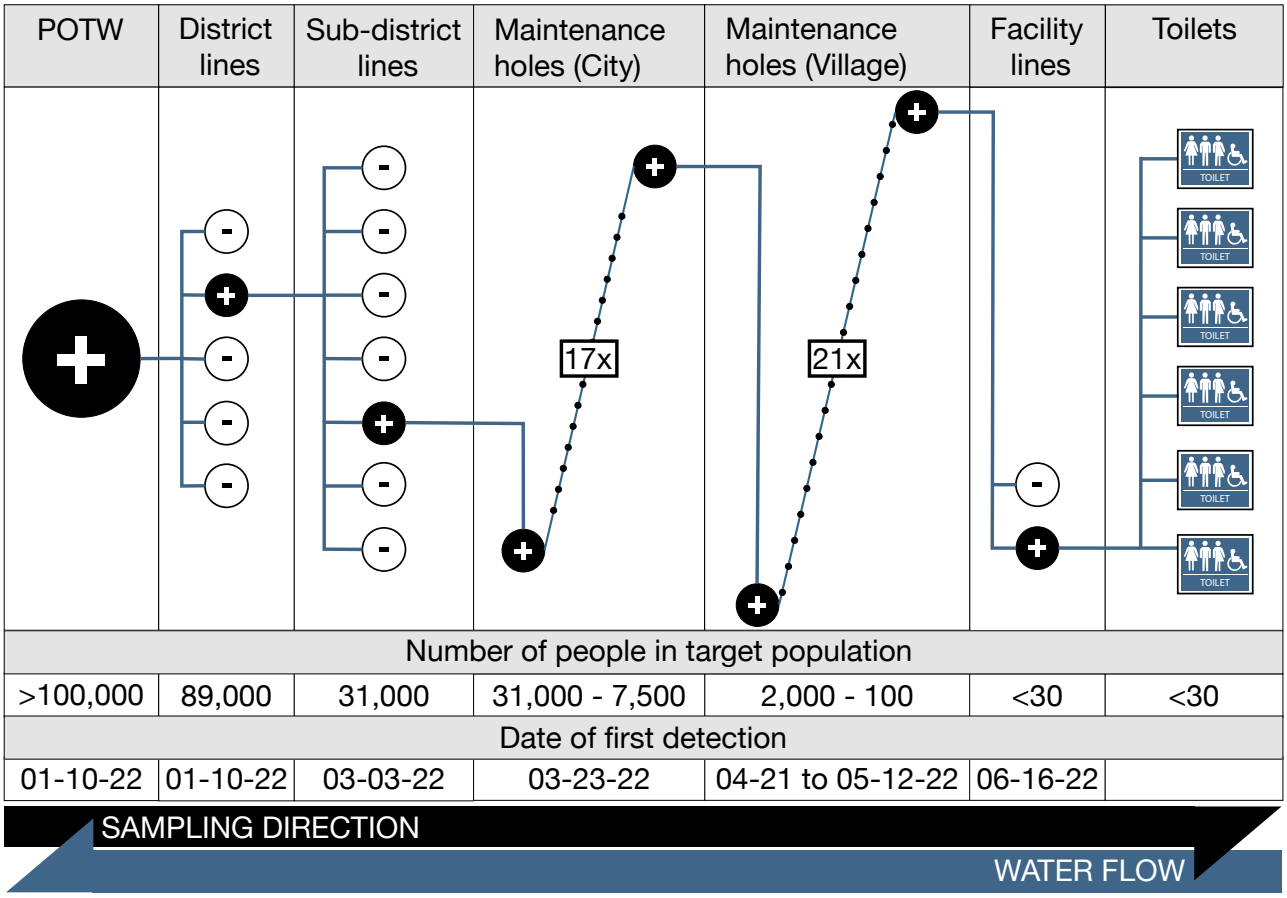
Competing interests

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.

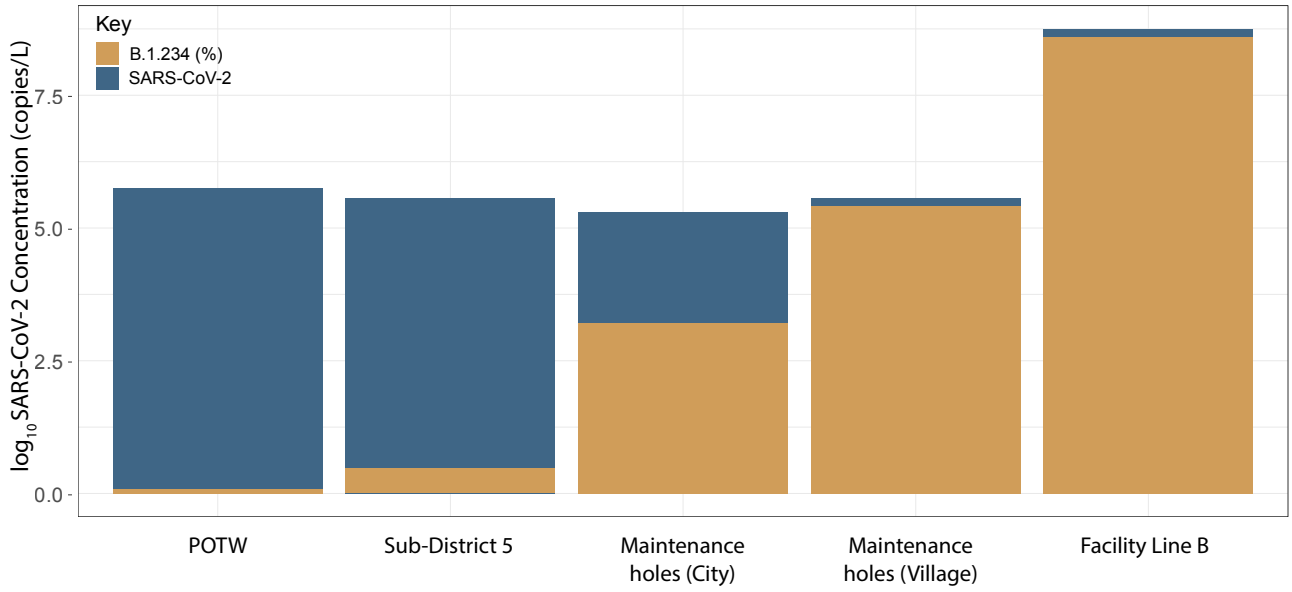
medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Figures

a



b



medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

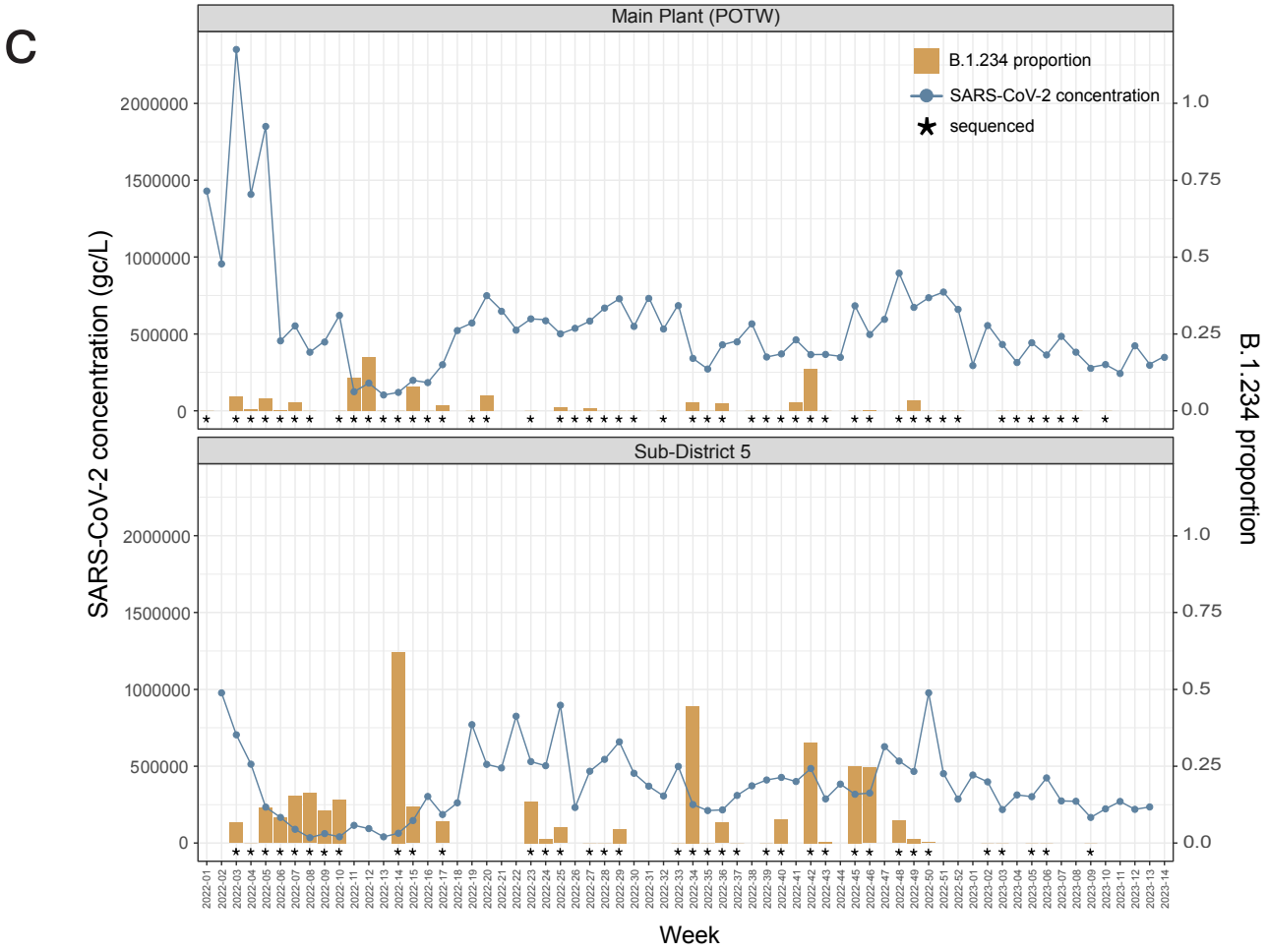


Figure 1. Tracking the source of the cryptic SARS-CoV-2 lineage. (a) The Wisconsin Lineage was first detected at the publicly owned treatment works (POTW) facility from one of the five district lines that serve the POTW sewershed. Continued wastewater sampling at interceptor lines that serve the positive district line isolated the lineage's source to a single sub-district (Sub-District 5). Further detailed source-tracking of maintenance holes pointed to a single place of business as the cryptic lineage's source, with an internal sampling of the facility pinpointing a collecting line servicing 6 toilets used by facility employees. (b) SARS-CoV-2 concentrations over five sampling levels show extremely high levels of SARS-CoV-2 in wastewater from Facility Line B. The Wisconsin Lineage's percent (B.1.234 %) contribution to the SARS-CoV-2 levels (estimated by Freyja) at each sampling level is shown in tan. (c) SARS-CoV-2 concentrations throughout 2022 for the Main Plant (POTW) and Sub-District 5 are shown in blue. The percent contribution of the Wisconsin Lineage (B.1.234 proportion) is shown in tan and showcases the continued detection at both sampling levels for most of 2022. Higher B.1.234 proportions were seen in Sub-District 5 than in the Main Plant (POTW), corresponding to getting nearer to the cryptic lineage's source.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

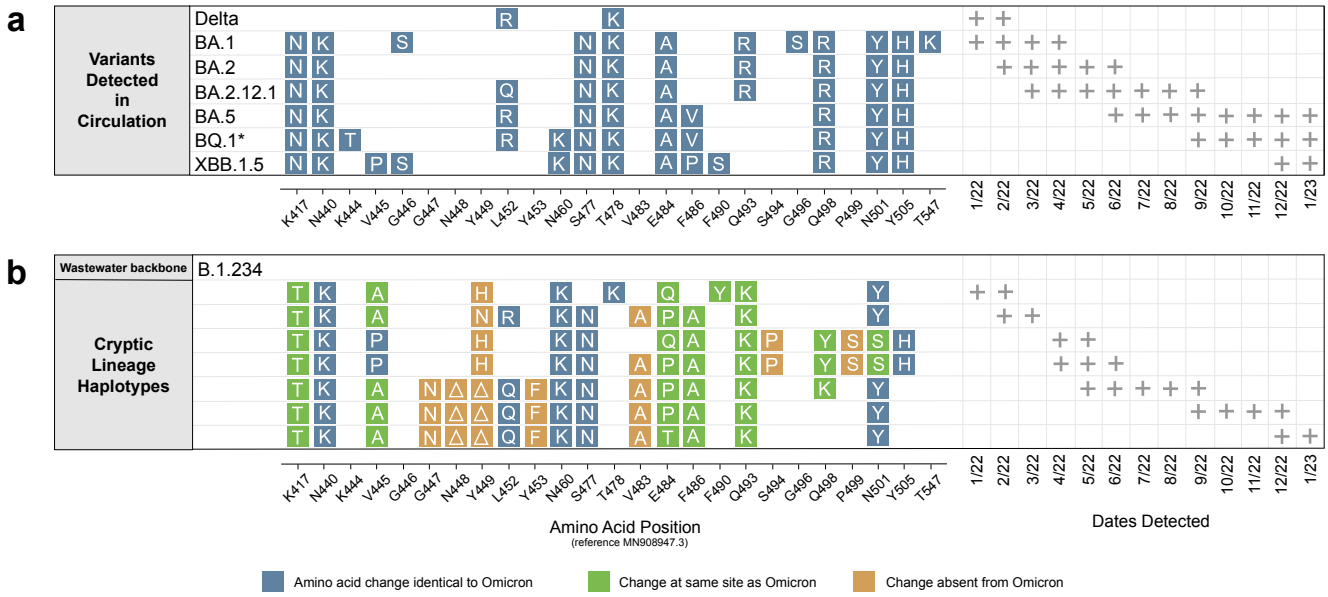


Figure 2. Representative haplotypes of Wisconsin Lineage sequences. (a) The SARS-CoV-2 spike RBD was amplified from Sub-District 5 wastewater samples using primers designed to exclude Omicron lineages. A single amplicon, spanning spike amino acid residues 377-606, was sequenced, and the amino acid changes (relative to Wuhan-1) found in Omicron lineages are shown in blue. (b) Representative haplotypes are displayed, each of which represented at least 25% of the total sequences in at least one sample. Green boxes indicate residues that are also altered in Omicron (BA.1 or BA.2), blue boxes indicate residues that have identical mutations to Omicron lineages, and tan boxes indicate residues that are altered in the Wisconsin Lineage and not in major Omicron lineages. Δ indicates an in-frame amino acid deletion.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

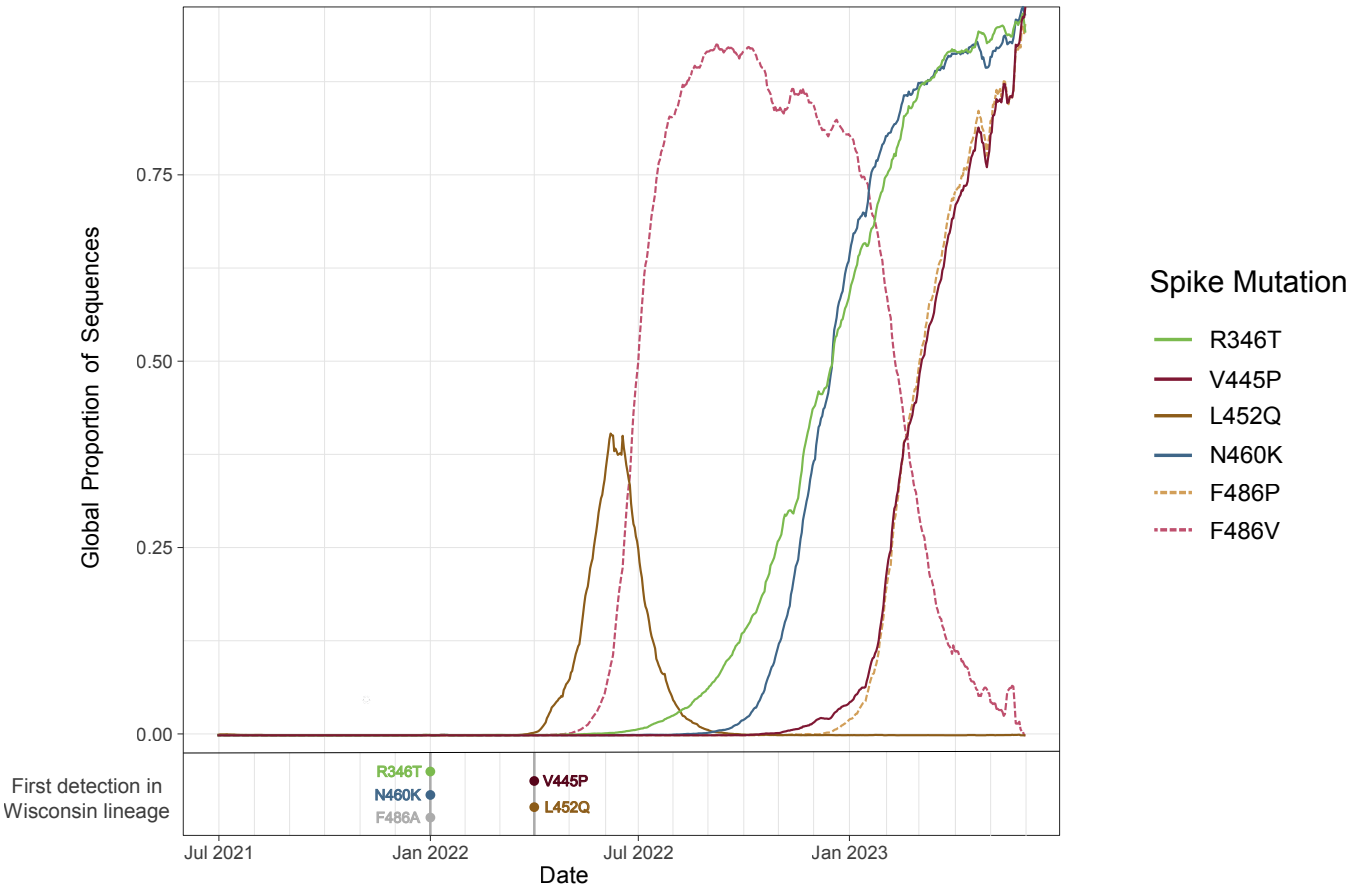
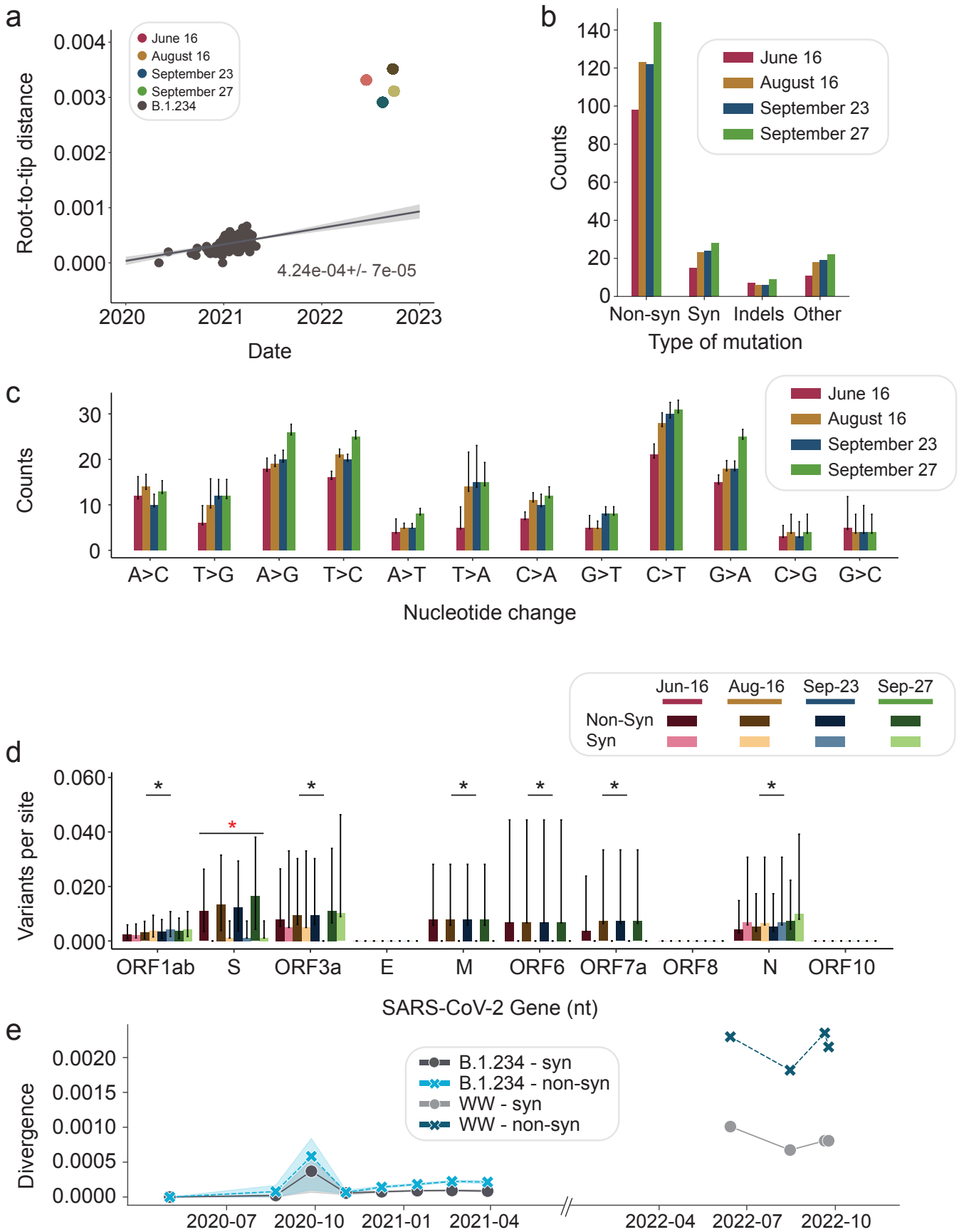


Figure 3. Prevalence of key cryptic lineage mutations in global sequences. The global proportions of sequences uploaded to NCBI GenBank for key mutations in the spike gene of the Wisconsin wastewater lineage are plotted over time. The spike mutations R346T, V445P, and N460K were all detected in the Wisconsin cryptic lineage months before becoming predominant in global sequences. The Wisconsin Lineage also harbored F486A 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). Searching for the global proportion of sequences was done in cov-SPECTRUM (https://cov-spectrum.org/explore/World/AllSamples/AllTimes/variants?&).[30]

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .



medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Figure 4. Diversity analysis of wastewater genomic sequences from all Facility Line B time points. (a) Root-to-tip regression analysis (distance) of B.1.234 sequences via TreeTime based on a maximum likelihood phylogenetic tree inferred with iqtree (not shown) and aligned to the MN908947.3 reference sequence. All sequences were obtained from GenBank and can be accessed using the accession numbers available on the GitHub repository accompanying this manuscript. (b) The enumeration of intra-host single nucleotide polymorphisms (iSNVs; y-axis) for the wastewater timepoints for each mutation type following alignment to the reference genome MN908947.3 (colored as in panel a). Mutations were classified as non-synonymous (Non-syn), synonymous (Syn), insertions-deletions (indels), or others (including nonsense and frameshift mutations outside of coding regions). (c) The number of nucleotide transitions and transversions from all timepoints. The 95% confidence intervals were obtained from the relative risk (RR) of every nucleotide substitution (i.e. RR= `A>C` / `C>A`). (d) We estimated genetic diversity within each sample using the summary statistics πN, which quantifies pairwise nonsynonymous differences per nonsynonymous site (darker bars) and πS, which quantifies pairwise synonymous differences per synonymous site (lighter bars), for each SARS-CoV-2 gene. The 95% confidence intervals were obtained using a binomial probability distribution. A Mann-Whitney two-sided test was applied to test the difference between πN and πS on each gene (red asterisk). 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 (black asterisks). (e) 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 to the WIsconsin Lineage isolates. With the exception of the Wisconsin Lineage, data are only plotted when windows contain at least 2 B.1.234 sequences.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

References

1 Xiao F, Tang M, Zheng X, Liu Y, Li X, Shan H. Evidence for Gastrointestinal Infection of SARS-CoV-2. *Gastroenterology* 05/2020; **158**: 1831–3.e3.

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. *Journal of Clinical Virology Plus* 02/2022; **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* 03/2021; **761**: 144216.

4 Gregory DA, Trujillo M, Rushford C, *et al.* Genetic Diversity and Evolutionary Convergence of Cryptic SARS-CoV-2 Lineages Detected Via Wastewater Sequencing. *medRxiv : the preprint server for health sciences* 2022; published online June 3. DOI:10.1101/2022.06.03.22275961.

5 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.

6 Abdool Karim SS, Karim QA. Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. *Lancet* 2021; **398**: 2126.

7 Cov-Lineages. https://cov-lineages.org/global_report_BA.1.html (accessed July 27, 2022).

8 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**. DOI:10.1016/j.onehlt.2020.100192.

9 Hale VL, Dennis PM, McBride DS, *et al.* SARS-CoV-2 infection in free-ranging white-tailed deer. *Nature* 2022; **602**: 481–6.

10 Chandler JC, Bevins SN, Ellis JW, *et al.* SARS-CoV-2 exposure in wild white-tailed deer (Odocoileus virginianus). *Proc Natl Acad Sci U S A* 2021; **118**. DOI:10.1073/pnas.2114828118.

11 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.

12 Rasmussen TB, Fonager J, Jørgensen CS, *et al.* Infection, recovery and re-infection of farmed mink with SARS-CoV-2. *PLoS Pathog* 2021; **17**: e1010068.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

13 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.

14 Bb OM, 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**. DOI:10.1126/science.abe5901.

15 Domańska-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**. DOI:10.2807/1560-7917.ES.2023.28.16.2300188.

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.

17 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**: 562–6.

18 Wilkinson SAJ, Richter A, Casey A, *et al.* Recurrent SARS-CoV-2 mutations in immunodeficient patients. *Virus Evol* 2022; **8**: veac050.

19 Gregory DA, Wieberg CG, Wenzel J, Lin C-H, Johnson MC. Monitoring SARS-CoV-2 Populations in Wastewater by Amplicon Sequencing and Using the Novel Program SAM Refiner. *Viruses* 2021; **13**: 1647.

20 Ewels PA, Peltzer A, Fillinger S, *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 2020; **38**: 276–8.

21 Karthikeyan S, Levy JI, De Hoff P, *et al.* Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. *Nature* 2022; **609**: 101–8.

22 outbreak.info. outbreak.info. https://outbreak.info/ (accessed Sept 8, 2022).

23 Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci Adv* 2020; **6**: eabb5813.

24 Liu X, Liu X, Zhou J, Dong Y, Jiang W, Jiang W. Rampant C-to-U deamination accounts for the intrinsically high mutation rate in SARS-CoV-2 spike gene. *RNA* 2022; **28**: 917–26.

medRxiv preprint doi: https://doi.org/10.1101/2022.10.28.22281553; this version posted May 23, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

25 Oloye FF, Xie Y, Asadi M, *et al.* Rapid transition between SARS-CoV-2 variants of concern Delta and Omicron detected by monitoring municipal wastewater from three Canadian cities. *Sci Total Environ* 2022; **841**: 156741.

26 Schmitz BW, Innes GK, Prasek SM, *et al.* Enumerating asymptomatic COVID-19 cases and estimating SARS-CoV-2 fecal shedding rates via wastewater-based epidemiology. *Sci Total Environ* 2021; **801**: 149794.

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; **0**. DOI:10.3389/fimmu.2021.679841.

28 Heffron AS, McIlwain SJ, Amjadi MF, *et al.* The landscape of antibody binding in SARS-CoV-2 infection. bioRxiv. DOI:10.1101/2020.10.10.334292.

29 Moulana A, Dupic T, Phillips AM, *et al.* Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nat Commun* 2022; **13**: 7011.

30 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**: 1735–7.