Supplemental Methods

Collection of wastewater

Wastewater samples for this study (January 2022 through March 2023) were collected in collaboration with experienced wastewater engineers from the city wastewater utility. The Wisconsin State Laboratory of Hygiene (WSLH) determined specific locations in the wastewater collection system to obtain samples for each round of testing, allowing them to gradually narrow down the origin of the Wisconsin Lineage source region. Sewage lift-stations, manholes, and facility sewer line access points were sampled with compositing autosamplers (ISCO 6712 and 6712c). Depending upon manhole depth, the autosampler was either placed on a shelf adjacent to the wastestream or suspended from the manhole opening, with weighted collection lines placed into the wastewater stream. The autosamplers were programmed to collect 24-hr composites, typically on a time-based mode, with wastewater composited into a 10-liter polypropylene container. The composite was kept cool during collection with ice packed around the collection container. Composite samples were transported to the analytical laboratory within a few hours of sample retrieval. While wastewater flows were available from the pump-stations and central municipal wastewater treatment facility, flow measurements were not made in the manhole waste streams.

Isolation of viral RNA from wastewater

Two approaches were used to isolate viral RNA from wastewater.
For samples processed at WSLH, wastewater samples (homogenized and unfiltered) were spiked with 20 µL/250 mL Calf-Guard® (Zoetis, Parsippany, NJ, USA), a cattle vaccine containing Bovine Coronavirus (BCoV) (as a virus recovery control), and briefly stored at 4°C until the viral targets were isolated and concentrated, typically on the day of receipt. A total of 10 mL (2x5mL) of wastewater was concentrated using Nanotrap Magnetic Virus Particles, Microbiome A and Enhancing Reagent 2 (Ceres Nanosciences, Manassas, VA, USA), using a KingFisher Apex automation platform. Total nucleic acids (TNA) were extracted using Maxwell(R) HT Environmental TNA kits (Promega, Madison, WI, USA) and eluted in 200 µL of 25 mM Tris HCl (pH 8.0) buffer. The extraction was automated using a KingFisher Flex (ThermoFisher Scientific, Waltham, MA, USA). KingFisher programs are available on Figshare: https://doi.org/10.6084/m9.figshare.21538143.v4.The long program was used for the concentration.
For samples processed at the University of Missouri, samples were processed as previously described.[4] Briefly, wastewater samples were centrifuged at 3000×g for 10 min and filtered through a 0.22 µM polyethersulfone membrane (Millipore, Burlington, MA, USA). Approximately 37.5 mL of wastewater was mixed with 12.5 mL solution containing 50% (w/vol) polyethylene glycol 8000 and 1.2 M NaCl, mixed, and incubated at 4C for at least 1 h. Samples were then centrifuged at 12,000×g for 2 h at 4C. Supernatant was decanted and RNA was extracted from the remaining pellet (usually not visible) with the QIAamp Viral RNA Mini Kit (Qiagen, Germantown, MD, USA) using the manufacturer's instructions. RNA was extracted in a final volume of 60 µL.

Quantification of viral RNA by RT-dPCR

Quantification of SARS-CoV-2, BCoV (internal control), PMMoV (fecal marker), and BRSV (spiked inhibition control) was achieved using reverse transcriptase digital PCR (RT-dPCR). Master mix was prepared using the One-Step Viral PCR kit (4x) (Qiagen, Germantown, MD, USA) and GT dPCR SARS-CoV-2 Wastewater Surveillance Assay Kit (GT Molecular, Fort Collins, CO, USA) with quantification of the following viral targets: N1, N2, BCoV, and PMMoV included with the GTMolecular dPCR SARS-CoV-2 Wastewater Surveillance Assay Kit, and BRSV primers and probes from IDT.[5] The samples were quantified on a QIAcuity Four Digital PCR System (Qiagen, Germantown, MD, USA). N1, N2, and BCoV were multiplexed on QIAcuity Nanoplate 26k 24-well plates while PMMoV and BRSV were singleplexed on 8.5k 96-well nanoplates. Cycling and exposure conditions are detailed in the table shown below. Analysis of the RT-dPCR results was performed with the QIAcuity Software Suite version 2.1.7.182. Thresholds were manually set to separate negative and positive partitions.

Table. dPCR Thermocycling Conditions:

Thermocycling Conditions:			
Step		**Time**	**Temp ºC**
Reverse Transcription		30 min	50
DNA polymerase activation		2 min	95
45 cycles	Denaturation	10 sec	95
	Anneal/Extend	30 sec	55
Target	Channel	Exposure	Gain
N1	Red (ROX)	500	4
N2	Green (FAM)	300	6
BCoV	Yellow (HEX)	300	6
PMMoV	Green (FAM)	300	6
BRSV	Yellow (HEX)	500	6

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

The primary RBD RT-PCR was performed using the Superscript IV One-Step RT-PCR System (Thermo Fisher Scientific,12594100). Primary RT-PCR amplification was performed as follows: 25 °C (2:00) + 50 °C (20:00) + 95 °C (2:00) + [95 °C (0:15) + 55 °C (0:30) + 72 °C (1:00)] × 25 cycles using the MiSeq primary PCR primers 5'-ATTCTGTCCTATATAATTCCGCAT-3' and 5'-CCCTGATAAAGAACAGCAACCT-3' (the first primer was changed to 5'-TATATAATTCCGCATCATTTTCCAC-3' starting in May, 2022 to adapt to changing Omicron lineages). Secondary PCR (25 µL) was performed on RBD amplifications using 5 µL of the primary PCR as template with MiSeq nested gene specific primers containing 5′ adapter sequences (0.5 µM each) 5'-acactctttccctacacgacgctcttccgatctGTGATGAAGTCAGA-CAAATCGC-3' and 5'-gtgactggagttcagacgtgtgctcttccgatctATGTCAAGAATCTCAAGTGTCTG-3', dNTPs (100 µM each) (New England Biolabs, N0447L) and Q5 DNA polymerase (New England Biolabs, M0541S). Secondary PCR amplification was performed as follows: 95 °C (2:00) + [95 °C (0:15) + 55 °C (0:30) + 72 °C (1:00)] × 20 cycles. A tertiary PCR (50 µL) was performed to add adapter sequences required for Illumina cluster generation with forward and reverse primers (0.2 µM each), dNTPs (200 µM each) (New England Biolabs, N0447L) and Phusion High-Fidelity or (KAPA HiFi for CA samples) DNA Polymerase (1U) (New England Biolabs, M0530L). PCR amplification was performed as follows: 98 °C (3:00) + [98 °C (0:15) + 50 °C (0:30) + 72 °C (0:30)] × 7 cycles +72 °C (7:00). Amplified product (10 µl) from each PCR reaction is combined and thoroughly mixed to make a single pool. Pooled amplicons were purified by the addition of Axygen AxyPrep MagPCR Clean-up beads (Axygen, MAG-PCR-CL-50) or in a 1.0 ratio to purify final amplicons. The final amplicon library pool was evaluated using the Agilent Fragment Analyzer automated electrophoresis system, quantified using the Qubit HS dsDNA assay (Invitrogen), and diluted according to Illumina's standard protocol. The Illumina MiSeq instrument was used to generate paired-end 300 base pair reads. Adapter sequences were trimmed from output sequences using Cutadapt. Sequencing reads were processed as previously described. Briefly, VSEARCH tools were used to merge paired reads and dereplicate sequences.[6] Dereplicated sequences from RBD amplicons were mapped to the reference sequence of SARS-CoV-2 (NC_045512.2) spike ORF using Minimap2.[7] Mapped amplicon sequences were then processed with SAM Refiner using the same spike sequence as a reference and the command line parameters "--Alpha 1.8 --foldab 0.6".[8]
The haplotypes representing at least 25% of the total sequences in at least one sample were rendered into figures using plot-nine (https://plotnine.readthedocs.io/en/stable/index.html).

SARS-CoV-2 whole genome sequencing of wastewater

Sequencing libraries were generated at the WSLH using the QIAseq DIRECT SARS-CoV-2 Enhanced kits with the primer Booster (QIAGEN, Germantown, MD, USA) following manufacturer's instructions. Briefly, 13 uL of total nucleic acid were reverse transcribed into cDNA using hexaprimers. SARS-CoV-2 genome was then specifically enriched using a SARS-CoV-2 primer panel. The panel consists of approximately 550 primers for creating 425 amplicons, covering the entire SARS-CoV-2 viral genome. UDI were 1:5 diluted. The library preparation was fully automated using the Biomek i5 Automated Workstation (Beckman Coulter). Libraries were quantified using a High Sensitivity Qubit 1X dsDNA HS Assay Kit (ThermoFisher

Scientific) and fragment size analyzed by a QIAxcel Advanced and the QX DNA Screening Kit (QIAGEN, Germantown, MD, USA). Libraries were sequenced on an Illumina MiSeq platform using MiSeq Reagent v2 (300 cycles) kits. Isolated RNA from each Facility Line B time point was whole-genome sequenced at least twice in separate Illumina MiSeq runs in anticipation of needing sequence technical replicates for later analysis. The data were analyzed with the nf-core/viralrecon workflow (https://nf-co.re/viralrecon/2.5) using the SARS-CoV-2 Wuhan-Hu-1 reference genome (Genbank accession MN908947.3) and the QIAseq Direct SARS-CoV-2 primer .bed file (https://www.qiagen.com/us/products/next-generation-sequencing/rna-sequencing/qiaseq-direct-sars-cov-2-kits/). After creating a sample sheet as described on the nf-core/viralrecon website (https://nf-co.re/viralrecon/usage), the workflow was initiated as outlined on the project's data portal (https://go.wisc.edu/4134pl). The output "variants_long_table.csv" from iVar was made into a pivot table in Microsoft Excel to make Supplemental Table 2. Because called variant frequencies differ between sequencing replicates from each time point, we decided to display the results from each replicate for the sake of transparency. Codons with variants detected in at least one sequence replicate from each time point were selected from Supplemental Table 2 and sorted by gene and frequency to make Supplemental Table 3. The presence of a particular called variant in one sequence replicate indicates that that variant could be present in the sample. The absence of a called variant in a replicate, on the other hand, does not prove its absence from the sample. Thus, we decided to include variants in Supplemental Table 3 even if they were only present in one sequence replicate for each time point.

Virus culture

To remove debris, samples were centrifuged twice at 3,500 rpm at 4oC for 15 minutes and then passed through a 0.8 µM syringe filter (Agilent) or left unfiltered. Samples (1ml) were incubated on nearly confluent Vero E6-TMPRSS2 (JCRB1819) or Vero E6-TMPRSS2/hACE2 cells (from Barney Graham, NIH) seeded the day prior in TC252 cm flasks for 1 hour at 370C. After the incubation, cells were washed twice and media was added back to the cells. The media contained 2-times the normal concentration of penicillin, streptomycin and amphotericin along with 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 prior in TC1752 cm flasks.

Variant Proportion Assessment

Variant proportions were assessed from WGS data using Freyja v.1.3.11, a tool previously developed to estimate the proportions of SARS-CoV-2 variants in deep sequence data containing mixed populations (10.1038/s41586-022-05049-6). Briefly, BAM files generated using viralrecon were processed by Freyja to create the variant and depth files (Wuhan-Hu-1 reference genome: MN908947.3). Variant proportions were assessed utilizing the median estimates obtained via the Freyja bootstrap *boot* function (nb = 10). The UShER barcode was updated on March 20th, 2023.

Root-to-tip regression

To generate **Figure 4-A**, we first downloaded from GenBank all full consensus genomes for SARS-CoV-2 belonging to Pango lineage B.1.234 (the inferred parent of the Wisconsin Lineage) and collected from specimens in the Midwest region (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin). The accession numbers for this dataset can be found on the GitHub repository accompanying this repository. The dataset is composed of 304 individual genome sequences collected between 2020-05-04 and 2021-05-01, which represents all the available B.1.234 sequences for the Midwest region available on GenBank. The dataset was filtered to exclude incomplete and low-quality sequences and to retain no more than 50 isolates per state. The list of accession numbers for the filtered isolates can also be found on the GitHub repository accompanying this manuscript. A total of 268 sequences were ultimately aligned to the Wuhan-Hu-1 reference sequence MN908947.3 using MAFFT (v7.505). A maximum likelihood phylogenetic tree was inferred using iqtree (v.2.2.0.3) with a molecular clock and distances obtained through treetime (v0.9.3). The analysis was conducted independently for the wastewater samples (WSLH-222, WSLH-223, WSLH-230, and WSLH-231) and root-to-tip distances for all strains were visualized in R (ggplot, dplyr). Phylogeny was visualized and annotated with FigTree (v.1.4.4). Scripts are available in the GitHub repository accompanying this manuscript (https://github.com/tcflab/wisconsin_cryptic_lineages).

Analyses for natural selection

Variants obtained through the nf-core/viralrecon workflows were processed using custom Python scripts (see Data Availability) to generate panels b-d in **Figure 4**. The multiple replicates for each collection date were used to obtain the intersection of variants, that is, variants that were found in all replicates for each collection date. The frequencies and depth of the resulting variants were recalculated. Variants differing from reference sequence Wuhan-Hu-1 (MN908947.3) were classified as non-synonymous (Non-syn), synonymous (Syn), insertions-deletions (indels), or others (including nonsense and frameshift mutations) using SnpEff (v.5.0). Synonymous and non-synonymous point mutations were quantified and compared between

timepoints, and 95% confidence intervals obtained from the relative risk (RR) of every nucleotide substitution against its inverted change (i.e., RR= `A>C` / `C>A`) using SciPy's relative_risk function (v.1.9.3). To obtain the proportion of variants per site, we enumerated synonymous and non-synonymous substitutions across the SARS-CoV-2 genome, and obtained the proportion against the number of synonymous and non-synonymous sites, respectively, using SNPGenie (v.2019.10.31). A binomial probability distribution was implemented to obtain the 95% confidence intervals via SciPy's binomtest function (v.1.9.3). A Mann-Whitney two-sided test was applied to test the difference between πN and πS on 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. To obtain synonymous and nonsynonymous divergence values (panel e), the average Hamming distance between B.1.234 isolates (dataset used in Figure 4a) and the MN908947.3 reference sequence was calculated as has been done previously for other coronaviruses.[9] Divergence was obtained over a sliding window of 36 days by dividing the observed synonymous and non-synonymous differences between the isolate and reference by the total possible number of synonymous and nonsynonymous nucleotide substitutions. Only windows that contained at least 2 sequences were considered for the analysis. Divergence values were independently calculated for each of the wastewater timepoints against the MN908947.3 reference sequence. Plot was visualized using Matplotlib. Scripts are available in the GitHub repository accompanying this manuscript (https://github.com/tcflab/wisconsin_cryptic_lineages).

Ethics statement

This activity was reviewed by CDC and the Wisconsin Department of Health Services, and was conducted consistent with applicable federal law and CDC policy (see, eg, 45 C.F.R. part 46, 21 C.F.R. part 56; 42 U.S.C. §241(d); 5 U.S.C. §552a; 44 U.S.C. §3501 et seq).

Supplemental Methods References

1 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.

2 Grubaugh ND, Gangavarapu K, Quick J, *et al.* An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* 2019; **20**: 8.

3 Gangavarapu K, Latif AA, Mullen JL, *et al.* Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat Methods* 2023; **20**: 512–22.

4 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.

5 Boxus M, Letellier C, Kerkhofs P. Real Time RT-PCR for the detection and quantitation of bovine respiratory syncytial virus. *J Virol Methods* 2005; **125**. DOI:10.1016/j.jviromet.2005.01.008.

6 Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016; **4**. DOI:10.7717/peerj.2584.

7 Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018; **34**: 3094–100.

8 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.

9 Kistler KE, Bedford T. Evidence for adaptive evolution in the receptor-binding domain of seasonal coronaviruses OC43 and 229e. *Elife* 2021; **10**. DOI:10.7554/eLife.64509.