AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

PUBLIC AND ENVIRONMENTAL HEALTH MICROBIOLOGY

MixviR: an R Package for Exploring Variation Associated with Genomic Sequence Data from Environmental SARS-CoV-2 and Other Mixed Microbial Samples

Michael G. Sovic,[a] Francesca Savona,[b] Zuzana Bohrerova,[c,d] Seth A. Faith[b,e]

[a]Center For Applied Plant Sciences, The Ohio State University, Columbus, Ohio, USA
[b]Infectious Diseases Institute, The Ohio State University, Columbus, Ohio, USA
[c]Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio, USA
[d]Ohio Water Resources Center, The Ohio State University, Columbus, Ohio, USA
[e]Center of Microbiome Science, The Ohio State University, Columbus, Ohio, USA

ABSTRACT The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)/coronavirus disease 2019 (COVID-19) pandemic has highlighted an important role for efficient surveillance of microbial pathogens. High-throughput sequencing technologies provide valuable surveillance tools, offering opportunities to conduct high-resolution monitoring from diverse sample types, including from environmental sources. However, given their large size and potential to contain mixtures of lineages within samples, such genomic data sets can present challenges for analyzing the data and communicating results with diverse stakeholders. Here, we report MixviR, an R package for exploring, analyzing, and visualizing genomic data from potentially mixed samples of a target microbial group. MixviR characterizes variation at both the nucleotide and amino acid levels and offers the RShiny interactive dashboard for exploring data. We demonstrate MixviR's utility with validation studies using mixtures of known lineages from both SARS-CoV-2 and *Mycobacterium tuberculosis* and with a case study analyzing lineages of SARS-CoV-2 in wastewater samples over time at a sampling location in Ohio, USA.

IMPORTANCE High-throughput sequencing technologies hold great potential for contributing to genomic-based surveillance of microbial diversity from environmental samples. However, the size of the data sets, along with the potential for environmental samples to contain multiple evolutionary lineages of interest, present challenges around analyzing and effectively communicating inferences from these data sets. The software described here provides a novel and valuable tool for exploring such data. Though originally designed and used for monitoring SARS-CoV-2 lineages in wastewater, it can also be applied to analyses of genomic diversity in other microbial groups.

KEYWORDS COVID-19, SARS-CoV-2, bioinformatics, genomics, pathogens, public health, surveillance studies, wastewater

High-throughput DNA sequencing (HTS) technologies represent powerful tools for characterizing the diversity of pathogens. While their application in epidemiological and public health settings has grown in recent years (1, 2), several factors still limit their more widespread usage. Among these are challenges related to efficiently analyzing these large and complex data sets and effectively communicating results with diverse stakeholders (3, 4), many of whom may have little or no experience with genomic data. Overcoming such challenges will help HTS technologies realize their full potential for building stronger understanding around human pathogens and better informing public health decisions.

Rapid and efficient identification of novel mutations and lineages with clinical or other public health relevance is important in the context of trying to best manage and mitigate disease

Editor Nicole R. Buan, University of Nebraska—Lincoln

Address correspondence to Michael G. Sovic, sovic.1@osu.edu.

The authors declare no conflict of interest.

Received 23 May 2022
Accepted 19 September 2022
Published 26 October 2022

impacts during outbreaks, as has been exemplified during the ongoing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic. Analysis of clinical infections offers one avenue for surveillance and represents a traditional usage of HTS in which nucleic acid samples are obtained from a single individual, uniquely barcoded and pooled for sequencing, and subsequently demultiplexed (separated) based on the molecular barcodes prior to analysis. This allows individual sequence reads to be assigned back to a single sample and analyzed separately from other such samples. However, HTS technologies also provide opportunities to expand surveillance efforts beyond patient samples to include those from the broader environment. Indeed, monitoring that targets sources such as wastewater has been implemented as part of SARS-CoV-2 surveillance programs (5–10), and interest in such efforts continues to grow (11, 12).

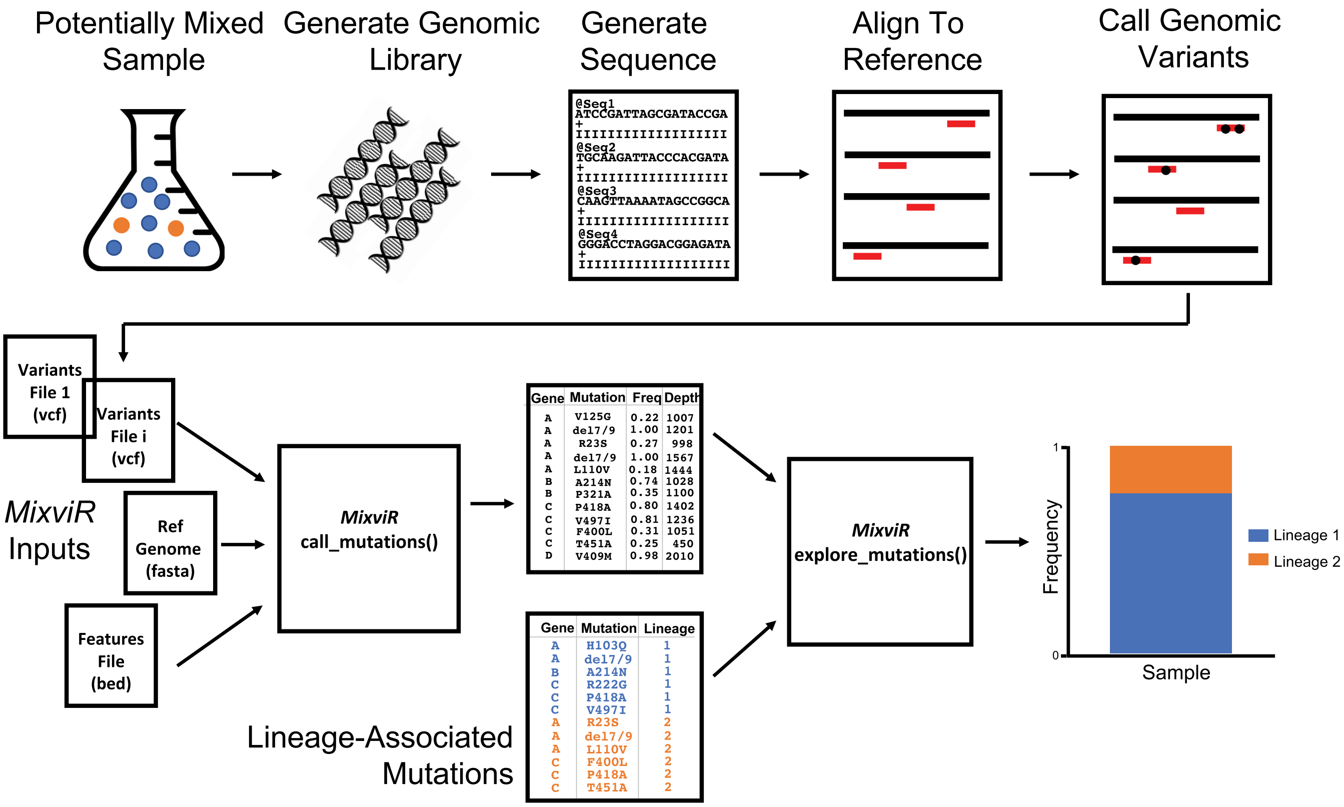
While efforts around monitoring from environmental samples hold many potential advantages, the data generated also present some specific analytical challenges. First, widely used programs for identifying lineages (variants) from a genomic sample, such as UShER (13) and Illumina's DRAGEN COVID Lineage app (San Diego, CA), are built with clinical data in mind and therefore expect samples to contain a single viral lineage. Not only may mixed samples not fit the assumptions of the algorithms implemented in such programs, but metrics uniquely relevant to mixed samples, such as estimates of the relative frequencies of the lineages they contain, are not generated by those tools. In addition, lineages of SARS-CoV-2 and other pathogens of interest are often at least partially characterized at the amino acid level, as amino acid variation is ultimately what gives rise to the most relevant clinical and public health characteristics of the pathogen. In contrast, genomic sequencing produces data exclusively at the nucleic acid level, meaning that translation is often a necessary part of an analysis pipeline. The need for software addressing some of these issues is evidenced by several recent publications describing new bioinformatic tools for lineage deconvolution from mixed genomic samples (14–18), with others reported in preprint form (19, 20). The tools available to date are run in a command line environment, require various levels of computational skills to install and execute, and in some cases, rely on outside software and/or databases that might not be readily available in many research or public health settings.

In this paper, we introduce MixviR, an easy-to-use R package designed to address the challenges above and to allow for efficient exploration, analysis, and visualization of mixed high-throughput sequencing data from a target group. Specific features of the program include characterization of mutations at both the nucleotide and amino acid levels and identification and estimation of the relative frequencies of known lineages in a sample. While the program was written with SARS-CoV-2 mutation and lineage detection in mind, it can be extended to analyses of other microbial groups by simply providing the relevant genomic reference information.

In its most basic usage, MixviR has three required inputs: (i) a FASTA-formatted reference genome file for the group/taxon of interest, (ii) an annotation (bed) file defining genes/open reading frames in the genome, and (iii) one variant call format (VCF) file for each genomic sample to be analyzed for potential mixtures of lineages within a target group. These VCF files can be generated with a number of common workflows, including those utilizing BCFtools (21), the GATK (22), or various applications available through Illumina's BaseSpace environment, such as the DRAGEN COVID Lineage or DRAGEN Somatic apps. For SARS-CoV-2, precompiled genome and annotation data are available and can be specified as an argument when running MixviR, meaning that sample VCF files are the only user-provided inputs required. Additional optional inputs include a file that associates mutations with known lineages of interest, which allows the program to draw inferences about the presence and relative frequencies of specific lineages in each sample, and a file that associates sample dates with individual sample locations, allowing analysis of data in a temporal framework. Outputs include a table of identified mutations with a number of customizable fields, as well as the RShiny interactive dashboard, which can be used for exploring and visualizing the data in a user-friendly way. An overview of a typical workflow involving a MixviR analysis is provided in Fig. 1.

RESULTS

SARS-CoV-2 validation. Sequencing coverage for the full SARS-CoV-2 data set was estimated to be >4,900 reads per base position. The number of lineage-characteristic



```
@Seq1
ATCCGATTAGCGATACCGA
+
IIIIIIIIIIIIIIIIIII
@Seq2
TGCAAGATTACCCACGATA
+
IIIIIIIIIIIIIIIIIII
@Seq3
CAAGTTAAAATAGCCGGCA
+
IIIIIIIIIIIIIIIIIII
@Seq4
GGGACCTAGGACGGAGATA
+
IIIIIIIIIIIIIIIIIII
```

Gene	Mutation	Freq	Depth
A	V125G	0.22	1007
A	del7/9	1.00	1201
A	R23S	0.27	998
A	del7/9	1.00	1567
A	L110V	0.18	1444
B	A214N	0.74	1028
B	P321A	0.35	1100
C	P418A	0.80	1402
C	V497I	0.81	1236
C	F400L	0.31	1051
C	T451A	0.25	450
D	V409M	0.98	2010

Gene	Mutation	Lineage
A	H103Q	1
A	del7/9	1
B	A214N	1
C	R222G	1
C	P418A	1
C	V497I	1
A	R23S	2
A	del7/9	2
A	L110V	2
C	F400L	2
C	P418A	2
C	T451A	2

FIG 1 Overview of a typical workflow involving a MixviR analysis. A sample is collected that potentially contains a mixture of lineages for a taxon of interest. Commonly used approaches for generating whole-genome-scale DNA sequence data and calling genomic variants relative to a reference genome can be used to generate the VCF-formatted files that, along with the relevant reference genome and associated annotation of regions to be translated, make up the necessary inputs for a MixviR run. MixviR then identifies amino acid-level variation and compares the mutations observed in the sample with a user-provided list of mutations associated with lineages of interest to identify the lineages present in the sample and estimate their frequencies.

amino acid mutations considered in the SARS-CoV-2 analyses ranged from 13 (Alpha) to 22 (Delta). The MixviR analyses identified from 90 to 100% of these mutations in samples known to contain their associated lineage (Fig. 2A) and correctly identified the identities of all lineages present in each of the seven mixtures (Fig. 2B), including in cases where the actual frequencies were as little as 3 to 4% (Delta in mixtures 4 and 5). MixviR's estimates of the frequencies of each lineage were also broadly consistent with expectations (see Fig. S1A in the supplemental material; $r^2 = 0.817$), with the notable pattern that estimates for Alpha and Delta tended to be slightly inflated, while those of the two Omicron lineages (BA.1 and BA.2) tended to be slightly lower than expected (Fig. 2B; Fig. S1A).

***Mycobacterium tuberculosis* validation.** The mean sequencing coverage for the full *M. tuberculosis* data set was 182 reads per base position. The numbers of characteristic mutations for the two *M. tuberculosis* lineages tested were 7 (CAS1-Kili) and 10 (EAI6-BGD1). MixviR identified all mutations expected in each sample, with the exception of a single mutation in the data submitted under NCBI SRA accession number ERR221664—a valine to phenylalanine substitution at amino acid position 1319 of the gene *eccC5* (Fig. 3A). The single mutation expected but not detected is associated with lineage EAI, which is expected to occur at a low (5%) frequency in this sample. Consistent with this, MixviR correctly identified and provided relative frequencies very consistent with the expectations for all lineages (Fig. 3B; Fig. S1B; $r^2 = 0.995$).

Evaluation of sequence coverage levels. MixviR was able to correctly identify all lineages present in the four samples at coverages of both 50× and 100× (Fig. S2 and S3), and the relative frequency estimates at these coverage levels were broadly consistent with expectations for all sample/lineage combinations (Fig. S2B and S3B). Most lineages were also correctly identified at the 10× coverage level. The exceptions were the lowest-frequency variants evaluated: Delta in mixture 4 of the SARS-CoV-2 data set and EAI in sample

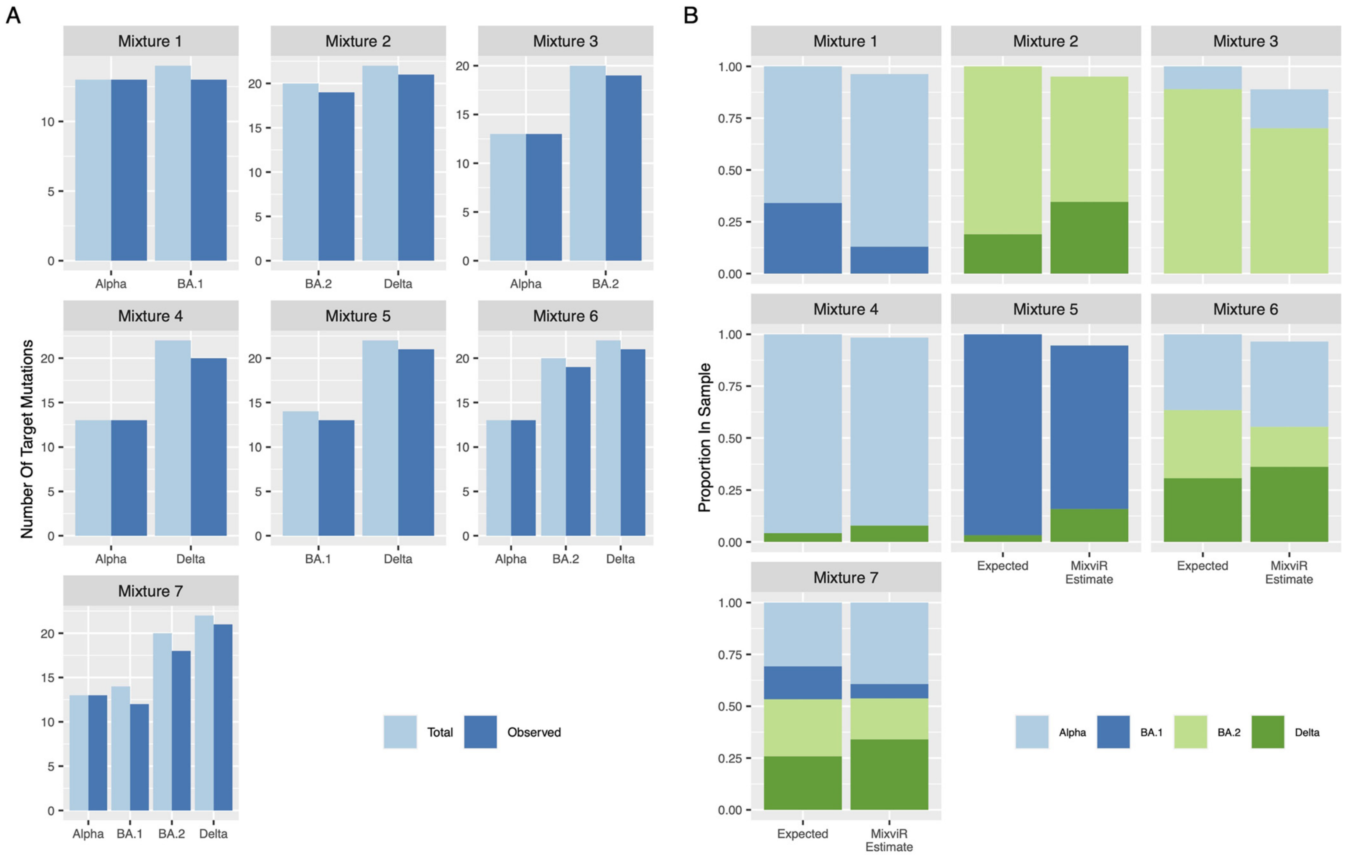


FIG 2 SARS-CoV-2 validation analysis with seven standards, each consisting of known mixtures of two to four of the SARS-CoV-2 lineages Alpha, Delta, Omicron BA.1, and Omicron BA.2. (A) Proportion of characteristic mutations detected for each expected lineage; (B) lineages detected by MixviR, along with their estimated frequencies, in comparison to the expected values for the standards.

ERR221664 of the *M. tuberculosis* data set. The expected frequencies of these lineages were 5% or less, and neither was detected at the 10× coverage level (Fig. S2 and S3). Specifically, for Delta in mixture 4, 6 of the 22 mutations characteristic of this lineage were detected at 10× coverage (less than the 50% threshold used for calling a lineage present), and for EAI in sample ERR221664, none of the 10 target mutations were detected with 10× coverage.

SARS-CoV-2 wastewater case study. MixviR identified 4 different SARS-CoV-2 lineages from the wastewater samples during the sampling period. The primary lineage observed from early April through late June 2021 was Alpha, though the Gamma lineage was also detected in two April samples (Fig. 4). Delta became the only lineage detected in late June and continued until mid-December 2021, when Omicron (BA.1) began to rise in frequency, replacing Delta by early January 2022. All of these patterns are consistent with those observed from clinical samples during the same time period.

DISCUSSION

The importance of monitoring the genetic diversity associated with human pathogens has been recognized for some time, and the value in such efforts has been highlighted during the ongoing SARS-CoV-2 pandemic. While a variety of useful monitoring approaches exist that include clinical screening, contact tracing, culturing of organisms, etc., genomic-based tools continue to be increasingly leveraged. Among the advantages to genomic-based monitoring are the potentially high sensitivity of detection, the ability to obtain samples from a variety of sources, which allows monitoring from environmental samples in addition to clinical samples, and the potentially fine-scale resolution of the evolutionary dynamics of a pathogen in real time. The growing efficiency and availability of high-throughput genetic sequencing

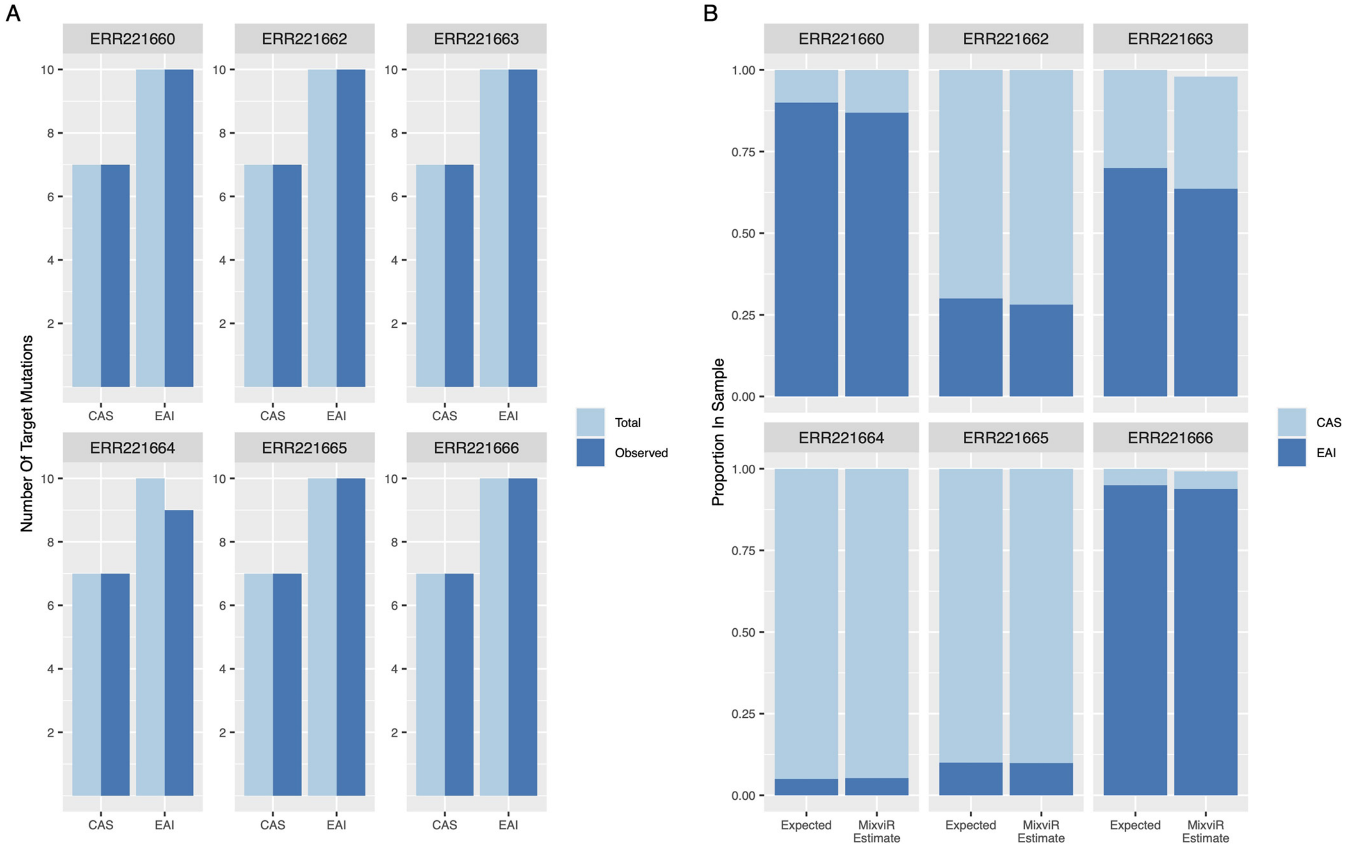


FIG 3 *M. tuberculosis* validation analysis with six standards, each consisting of known mixtures of two *M. tuberculosis* lineages (CAS1-Kili and EAI6-BGD1). (A) Proportion of characteristic mutations detected for each expected lineage; (B) lineages detected by MixviR along with their estimated frequencies in comparison to expected values for the standards.

technologies continues to put the realization of these advantages within practical reach of researchers and public health officials.

Indeed, partly due to the reasons above, genomic sequencing efforts related to the SARS-CoV-2 pandemic have far exceeded those utilized as part of any previous public health event, and while these efforts have provided a wealth of information, they have also raised areas where useful tools are lacking and improvements are needed. As part of our own experimental programs to monitor SARS-CoV-2 diversity from environmental sources, including wastewater and dust, it became apparent that tools were needed to efficiently analyze large HTS data sets with potential mixtures of microbial lineages and to effectively share those inferences with diverse stakeholders.

In response, we developed MixviR, an R package that allows visualization and exploration of genomic sequencing data from input files in formats widely used in genomic studies. Analysis of multiple standards consisting of mixtures of lineages from each of two microbial systems (SARS-CoV-2 and *M. tuberculosis*) and application in a case study with SARS-CoV-2 wastewater samples demonstrated that MixviR is able to accurately identify and estimate the frequencies of lineages present in mixed samples. The sensitivity to detect low frequency variants is a function of sequencing coverage, as demonstrated in the analyses, in which raw data sets were randomly subsampled to specific coverage levels and reanalyzed. Though various factors will influence the sensitivity of detection, including the evenness of coverage across the genome and the threshold used for the proportion of characteristic mutations that must be identified to call a lineage present in the sample, detection of a variant occurring at a 5% frequency would be expected to require approximately 20× coverage. The results from the subsampling analyses were consistent with this expectation, as the two lowest-frequency variants analyzed, which had expected frequencies as low as 3 to 4%, were not detected at 10× coverage but were detected at 50× coverage and above.

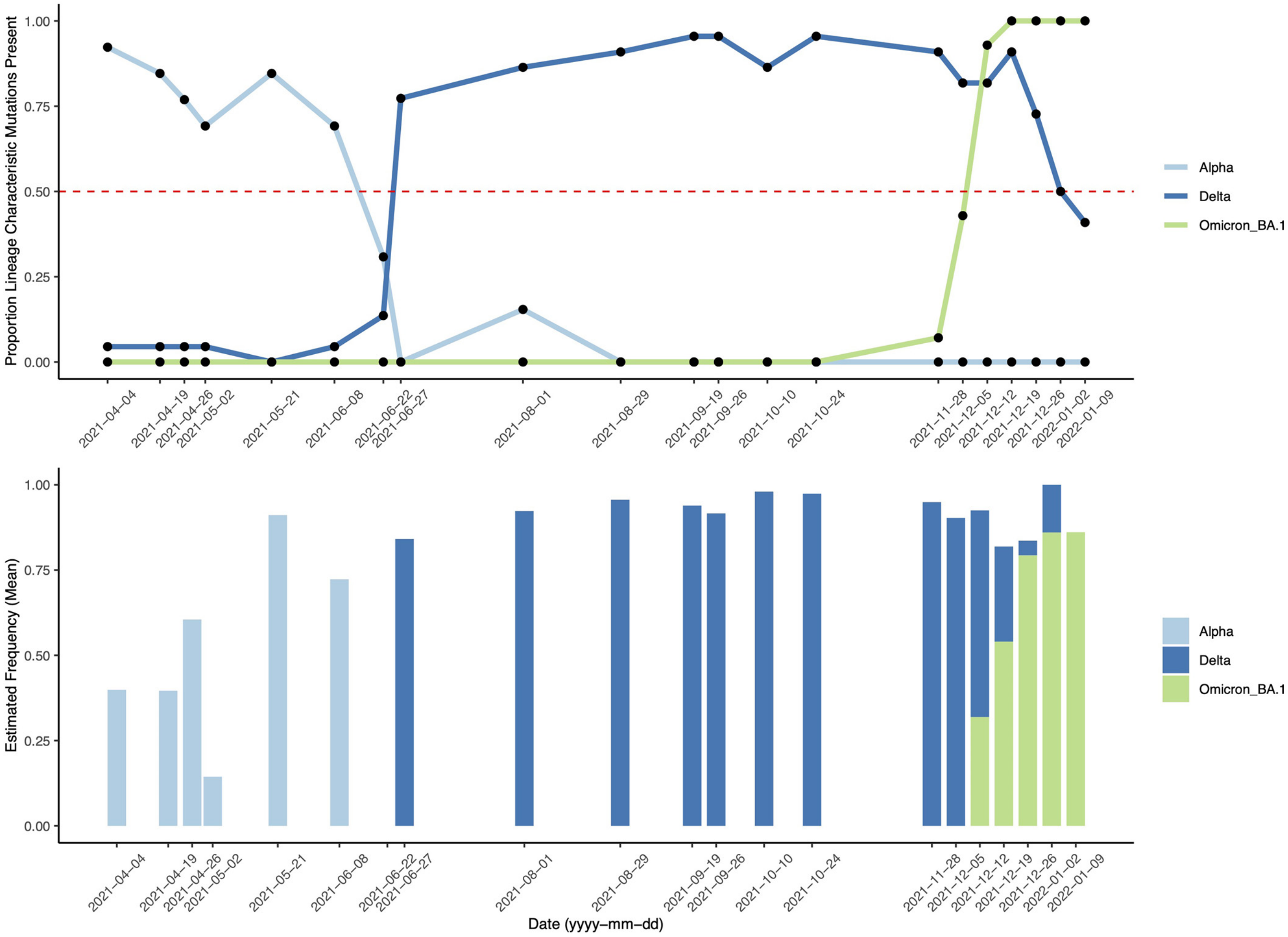


FIG 4 MixviR results for samples collected over time at a wastewater treatment facility in Ohio, USA, and analyzed for SARS-CoV-2 lineages. (Top) lineages present in the sample are represented by points above the red dashed line, which mark the adjustable threshold for the proportion of lineage-characteristic mutations observed in the sample; (bottom) plot of the estimated frequency of each identified lineage. These plots represent one of several types of RShiny outputs available from MixviR for summarizing sample mutation data. Examples of all RShiny outputs are provided in Fig. S4 to S8 in the supplemental material.

By relying on a small number of R functions to generate a mutation list at both the nucleotide and amino acid levels and the associated RShiny interactive dashboard for data exploration (see “RShiny Dashboard Examples” in the supplemental material for details and Fig. S4 to S8 for examples of RShiny output beyond that presented in Fig. 4 above), MixviR is written to allow easy and efficient analysis for users with minimal experience or training in R or other programming languages. The flexibility it provides in allowing analysis of various target microbial groups will make it a valuable tool not just for monitoring SARS-CoV-2 variants but also for environmental monitoring of emerging pathogens, such as those related to monkeypox (23) and polio (24).

MATERIALS AND METHODS

SARS-CoV-2 validation data set. Four clinical SARS-CoV-2 RNA samples were obtained, representing the Alpha, Delta, Omicron BA.1, and Omicron BA.2 lineages (The Ohio State University Applied Microbiology Services Laboratory; institutional review board [IRB] protocol number 2021H0080). Seven mixed SARS-CoV-2 test samples were generated by pooling two to four of the RNA samples in various relative ratios, generating amplicon-based whole-genome libraries with ARTICv4.1 primers and Tagmentation (Illumina), and sequencing each on an Illumina NextSeq 2000 instrument (2 × 101 bp). Raw fastq files were produced and analyzed using Illumina's BaseSpace with the DRAGEN COVID Lineage application v3.5.10 pipeline with the callability and coverage thresholds set to 1. The resulting VCF files served as input for MixviR v3.3.5. The MixviR analysis used the preconfigured SARS-CoV-2 reference (option *reference* = “Wuhan”), a list of mutations characterizing the SARS-CoV-2 lineages Alpha, Beta, Gamma, Delta, Lambda, Mu, Omicron BA.1, and Omicron BA.2 assembled from GSAID data by, and obtained from, the website https://outbreak.info/ (25) on 17 February 2022, and otherwise default settings. The list of mutations used is available from https://github.com/mikesovic/MixviR/tree/main/mutation_files.

***M. tuberculosis* validation data set.** Raw sequencing files for six mixed samples associated with the work of Sobkowiak et al. (26) were downloaded from SRA. These samples contain different known proportions of the *M. tuberculosis* lineages EAI6-BGD1 and CAS1-Kili. Each sample was analyzed using the DRAGEN Somatic application in Illumina's BaseSpace environment. The analyses used the *M. tuberculosis* reference genome H37Rv (build ASM19595v2) with no somatic hot spots, multiallelic filtering disabled, and the vc-enable-unequal-ntd-errors option set to false. The VCF files produced served as input for MixviR v3.3.5. Mutations associated with the *M. tuberculosis* lineages designated L1 and L3 by Merker et al. (27) and corresponding to EAI6-BGD1 and CAS1-Kili, respectively, were obtained and used as part of the MixviR analysis to identify and estimate the frequencies of the two target lineages in each sample. MixviR's genetic.code.num argument was set to 11, and default settings were otherwise used in the analysis.

Analyses of validation data sets. For each mixed sample in the validation data sets, the estimate_lineages() function was used to identify what lineages were present, estimate the frequency of each lineage, and also determine the number of characteristic mutations identified for each lineage. The presence of lineage *l* in sample *i* is inferred by comparing the ratio n_{li}/N_l to an adjustable threshold value (default = 0.5), where N_l is the total number of mutations associated exclusively with lineage *l*, and n_{li} is the number of these lineage-specific mutations observed in the sample *i*. Lineages with a ratio exceeding the threshold are inferred as present, and the proportional contribution of that lineage to the sample is then estimated by averaging over the frequencies for the lineage-specific mutations present in the sample. Frequency estimates for each individual mutation (*m*) are calculated as r_m/r_t, where r_m is the number of sequence reads associated with the mutated allele, and r_t is the total number of reads (sequencing depth) at the genomic position associated with the mutant allele.

Analyses of effects of sequencing coverage. The effects of various levels of sequence coverage were evaluated by selecting two samples from each data set (two from SARS-CoV-2 and two from *M. tuberculosis*) and randomly subsampling the raw fastq reads to generate data sets representing coverages of approximately 10×, 50×, and 100×. One of the two data sets from each group included a lineage expected to be at low frequency (<5%). Each of the 12 subsampled data sets (four samples, each at three coverage levels) were then analyzed using the methods described above for their respective data set. The lineages identified and associated estimated relative frequencies were then evaluated with respect to the various coverage levels.

SARS-CoV-2 wastewater case study. Sequence data were obtained for samples (*N* = 21) collected from one Ohio wastewater treatment plant between April 2021 and January 2022 as part of a program led by the Ohio Water Resources Center at The Ohio State University and the Ohio Department of Health to monitor SARS-CoV-2 occurrence throughout the state. Processing of wastewater samples, genomic library preparation, and sequencing followed methods described by Ai et al. (9) and Hale et al. (28). The sequence data were analyzed using the same methods as for the SARS-CoV-2 validation data set to identify the lineages present and estimate their frequencies.

Data availability. The code and associated files used for all analyses in this study are provided at https://github.com/mikesovic/MixviR/tree/main/mixvir_paper. MixviR can be obtained from the Comprehensive R Archive Network (CRAN) or from GitHub at https://github.com/mikesovic/MixviR. The raw sequence data used for the SARS-CoV-2 validation are available from the NCBI Sequence Read Archive (SRA) under accession number PRJNA827817. Data for the *M. tuberculosis* validation were obtained from the SRA as described above.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 2.1 MB.

ACKNOWLEDGMENTS

Funding was provided from the United States Department of Treasury through Ohio EPA project number OSU-FDCARES20 and from a CDC grant through the Ohio Department of Health, award number 6 NU50CK000543-02-11. We acknowledge the Ohio Water Resources Center for providing administrative support of the project.

The Ohio Department of Health data used in this study were obtained from the Bureau of Environmental Health and Radiation Protection, Ohio Department of Health (ODH). Use of these data does not imply that ODH agrees or disagrees with any presentations, analyses, interpretations or conclusions.

REFERENCES

1. Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, Tóth Á, Griškevičius A, Vatopoulos A, Skoczynska A, Pantosti A, Coignard B, ECDC National Microbiology Focal Points and Experts Group. 2017. Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European national capacities, 2015–2016. Front Public Health 5:347. https://doi.org/10.3389/fpubh.2017.00347.
2. Armstrong GL, MacCannell DR, Taylor J, Carleton HA, Neuhaus EB, Bradbury RS, Posey JE, Gwinn M. 2019. Pathogen genomics in public health. N Engl J Med 381:2569–2580. https://doi.org/10.1056/NEJMsr1813907.
3. Ferdinand AS, Kelaher M, Lane CR, da Silva AG, Sherry NL, Ballard SA, Andersson P, Hoang T, Denholm JT, Easton M, Howden BP, Williamson DA. 2021. An implementation science approach to evaluating pathogen whole

genome sequencing in public health. Genome Med 13:121. https://doi.org/10.1186/s13073-021-00934-7.

4. McClary-Gutierrez JS, Mattioli MC, Marcenac P, Silverman AI, Boehm AB, Bibby K, Balliet M, de Los Reyes FL, Gerrity D, Griffith JF, Holden PA, Katehis D, Kester G, LaCross N, Lipp EK, Meiman J, Noble RT, Brossard D, McLellan SL. 2021. SARS-CoV-2 wastewater surveillance for public health action. Emerg Infect Dis 27:1–8. https://doi.org/10.3201/eid2709.210753.
5. Herold M, d'Hérouël AF, May P, Delogu F, Wienecke-Baldacchino A, Tapp J, Walczak C, Wilmes P, Cauchie HM, Fournier G, Ogorzaly L. 2021. Genome sequencing of SARS-CoV-2 allows monitoring of variants of concern through wastewater. Water 13:3018. https://doi.org/10.3390/w13213018.
6. Izquierdo-Lara R, Elsinga G, Heijnen L, Munnink BBO, Schapendonk CME, Nieuwenhuijse D, Kon M, Lu L, Aarestrup FM, Lycett S, Medema G, Koopmans MPG, de Graaf M. 2021. Monitoring SARS-CoV-2 circulation and diversity through community wastewater sequencing, the Netherlands and Belgium. Emerg Infect Dis 27:1405–1415. https://doi.org/10.3201/eid2705.204410.
7. Wurtz N, Revol O, Jardot P, Giraud-Gatineau A, Houhamdi L, Soumagnac C, Annessi A, Lacoste A, Colson P, Aherfi S, Scola BL. 2021. Monitoring the circulation of SARS-CoV-2 variants by genomic analysis of wastewater in Marseille, south-east France. Pathogens 10:1042. https://doi.org/10.3390/pathogens10081042.
8. Rouchka EC, Chariker JH, Saurabh K, Waigel S, Zacharias W, Zhang M, Talley D, Santisteban I, Puccio M, Moyer S, Holm RH, Yeager RA, Sokoloski KJ, Fuqua J, Bhatnagar A, Smith T. 2021. The rapid assessment of aggregated wastewater samples for genomic surveillance of SARS-CoV-2 on a city-wide scale. Pathogens 10:1271. https://doi.org/10.3390/pathogens10101271.
9. Ai Y, Davis A, Jones D, Lemeshow S, Tu H, He F, Ru P, Pan X, Bohrerova Z, Lee J. 2021. Wastewater SARS-CoV-2 monitoring as a community-level COVID-19 trend tracker and variants in Ohio, United States. Sci Total Environ 801:149757. https://doi.org/10.1016/j.scitotenv.2021.149757.
10. Rothman JA, Loveless TB, Kapcia J, III, Adams ED, Steele JA, Zimmer-Faust AG, Langlois K, Wanless D, Griffith M, Mao L, Chokry J, Griffith JF, Whiteson KL. 2021. RNA viromics of Southern California wastewater and detection of SARS-CoV-2 single-nucleotide variants. Appl Environ Microbiol 87:e01448-21. https://doi.org/10.1128/AEM.01448-21.
11. Farkas K, Hillary LS, Malham SK, McDonald JE, Jones D. 2020. Wastewater and public health: the potential of wastewater surveillance for monitoring COVID-19. Curr Opin Environ Sci Health 17:14–20. https://doi.org/10.1016/j.coesh.2020.06.001.
12. Tereshchenko LG. 2021. Monitoring the spread of SARS-CoV-2 is an important public health task. Am J Public Health 111:1387–1388. https://doi.org/10.2105/AJPH.2021.306392.
13. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, Haussler D, Corbett-Detig R. 2021. Ultrafast sample placement on existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet 53:809–816. https://doi.org/10.1038/s41588-021-00862-7.
14. Gregory DA, Wieberg CG, Wenzel J, Lin C-H, Johnson MC. 2021. Monitoring SARS-CoV-2 populations in wastewater by amplicon sequencing and using the novel program SAM Refiner. Viruses 13:1647. https://doi.org/10.3390/v13081647.
15. Valieris R, Drummond RD, Defelicibus A, Dias-Neto E, Rosales RA, Tojal da Silva I. 2022. A mixture model for determining SARS-CoV-2 variant composition in pooled samples. Bioinformatics 38:1809–1815. https://doi.org/10.1093/bioinformatics/btac047.
16. Amman F, Markt R, Endler L, Hupfauf S, Agerer B, Schedl A, Richter L, Zechmeister M, Bicher M, Heiler G, Triska P, Thornton M, Penz T, Senekowitsch M, Laine J, Keszei Z, Klimek P, Nägele F, Mayr M, Daleiden B, Steinlechner M, Niederstätter H, Heidinger P, Rauch W, Scheffknecht C, Vogl G, Weichlinger G, Wagner AO, Slipko K, Masseron A, Radu E, Allerberger F, Popper N, Bock C, Schmid D, Oberacher H, Kreuzinger N, Insam H, Bergthaler A. 2022. Viral variant-resolved wastewater surveillance of SARS-CoV-2 at national scale. Nat Biotechnol https://doi.org/10.1038/s41587-022-01387-y.
17. Jahn K, Dreifuss D, Topolsky I, Kull A, Ganesanandamoorthy P, Fernandez-Cassi X, Bänziger C, Devaux AJ, Stachler E, Caduff L, Cariti F, Corzón AT, Fuhrmann L, Chen C, Jablonski KP, Nadeau S, Feldkamp M, Beisel C, Aquino C, Stadler T, Ort C, Kohn T, Julian TR, Beerenwinkel N. 2022. Early detection and surveillance of SARS-CoV-2 genomic variants in wastewater using COJAC. Nat Microbiol 7:1151–1160. https://doi.org/10.1038/s41564-022-01185-x.
18. Karthikeyan S, Levy JI, De Hoff P, Humphrey G, Birmingham A, Jepsen K, Farmer S, Tubb HM, Valles T, Tribelhorn CE, Tsai R, Aigner S, Sathe S, Moshiri N, Henson B, Mark AM, Hakim A, Baer NA, Barber T, Belda-Ferre P, Chacón M, Cheung W, Cresini ES, Eisner ER, Lastrella AL, Lawrence ES, Marotz CA, Ngo TT, Ostrander T, Plascencia A, Salido RA, Seaver P, Smoot EW, McDonald D, Neuhard RM, Scioscia AL, Satterlund AM, Simmons EH, Abelman DB, Brenner D, Bruner JC, Buckley A, Ellison M, Gattas J, Gonias SL, Hale M, Hawkins F, Ikeda L, Jhaveri H, Johnson T, et al. 2022. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature 609:101–108. https://doi.org/10.1038/s41586-022-05049-6.
19. Schumann V-F, Cuadrat RRDC, Wyler E, Wurmus R, Deter A, Quedenau C, Dohmen J, Faxel M, Borodina T, Blume A, Meixner M, Grau JH, Liere K, Hackenbeck T, Zietzschmann F, Gnirss R, Böckelmann U, Uyar B, Franke V, Barke N, Altmüller J, Rajewsky N, Landthaler M, Akalin A. 2022. SARS-CoV-2 infection dynamics revealed by wastewater sequencing analysis and deconvolution. medRxiv. https://doi.org/10.1101/2021.11.30.21266952.
20. Baaijens JA, Zulli A, Ott IM, Petrone ME, Alpert T, Fauver JR, Kalinich CC, Vogels CBF, Breban MI, Duvallet C, McElroy K, Ghaeli N, Imakaev M, Mckenzie-Bennett M, Robison K, Plocik A, Schilling R, Pierson M, Littlefield R, Spencer M, Simen BB, Hanage WP, Grubaugh ND, Peccia J, Baym M, Yale SARS-CoV-2 Genomic Surveillance Initiative. 2021. Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-Seq quantification. medRxiv. https://doi.org/10.1101/2021.08.31.21262938.
21. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008.
22. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498. https://doi.org/10.1038/ng.806.
23. Peiró-Mestres A, Fuertes I, Camprubí-Ferrer D, Marcos MÁ, Vilella A, Navarro M, Rodriguez-Elena L, Riera J, Català A, Martínez MJ, Blanco JL, Hospital Clinic de Barcelona Monkeypox Study Group. 2022. Frequent detection of monkeypox virus DNA in saliva, semen, and other clinical samples from 12 patients, Barcelona, Spain, May to June 2022. Euro Surveill 27:2200503. https://doi.org/10.2807/1560-7917.ES.2022.27.28.2200503.
24. Kline A, Dean K, Kossik AL, Harrison JC, Januch JD, Beck NK, Zhou NA, Shirai JH, Boyle DS, Mitchell J, Meschke JS. 2022. Persistence of poliovirus types 2 and 3 in waste-impacted water and sediment. PLoS One 17:e0262761. https://doi.org/10.1371/journal.pone.0262761.
25. Gangavarapu K, Latif AA, Mullen J, Alkuzweny M, Hufbauer E, Tsueng G, Haag E, Zeller M, Aceves CM, Zaiets K, Cano M, Zhou J, Qian A, Sattler R, Matteson NL, Levy JI, Lee RTC, Freitas L, Maurer-Stroh S, Suchard MA, Wu C, Su AI, Andersen KG, Hughes LD, GISAID Core and Curation Team. 2022. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. medRxiv. https://doi.org/10.1101/2022.01.27.22269965.
26. Sobkowiak B, Glynn JR, Houben RMGJ, Mallard K, Phelan JE, Guerra-Assunção JA, Banda L, Mzembe T, Viveiros M, McNerney R, Parkhill J, Crampin AC, Clark TG. 2018. Identifying mixed *Mycobacterium tuberculosis* infections from whole genome sequence data. BMC Genomics 19:613. https://doi.org/10.1186/s12864-018-4988-z.
27. Merker M, Kohl TA, Barilar I, Andres S, Fowler PW, Chryssanthou E, Ängeby K, Jureen P, Moradigaravand D, Parkhill J, Peacock SJ, Schön T, Maurer FP, Walker T, Köser C, Niemann S. 2020. Phylogenetically informative mutations in genes implicated in antibiotic resistance in *Mycobacterium tuberculosis* complex. Genome Med 12:27. https://doi.org/10.1186/s13073-020-00726-5.
28. Hale VL, Dennis PM, McBride DS, Nolting JM, Madden C, Huey D, Ehrlich M, Grieser J, Winston J, Lombardi D, Gibson S, Saif L, Killian ML, Lantz K, Tell RM, Torchetti M, Robbe-Austerman S, Nelson MI, Faith SA, Bowman AS. 2022. SARS-CoV-2 infection in free-ranging white-tailed deer. Nature 602:481–486. https://doi.org/10.1038/s41586-021-04353-x.