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. Faithb,e

aCenter For Applied Plant Sciences, The Ohio State University, Columbus, Ohio, USA
bInfectious Diseases Institute, The Ohio State University, Columbus, Ohio, USA
cDepartment of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, Ohio, USA
dOhio Water Resources Center, The Ohio State University, Columbus, Ohio, USA
eCenter 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
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
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 ecc5 (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 frequencies 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...
ERR221664 of the *M. tuberculosis* data set. The expected frequencies of these lineages were 5% or less, and neither was detected at the $10^x$ coverage level (Fig. S2 and S3). Specifically, for Delta in mixture 4, 6 of the 22 mutations characteristic of this lineage were detected at $10^x$ 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 $10^x$ 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...
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 exploitation 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 3** *M. tuberculosis* validation analysis with six standards, each consisting of known mixtures of two *M. tuberculosis* lineages (CAS1-Kili and EA16-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.
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

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.
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 Sokolowski et al. (26) were downloaded from SRA. These samples contain different known proportions of the M. tuberculosis lineages EA6-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 EA6-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 $i$ in sample $j$ is inferred by comparing the ratio $n_j/N_j$ to an adjustable threshold value ($\text{default} = 0.5$), where $N_j$ is the total number of mutations associated exclusively with lineage $i$, and $n_j$ is the number of these lineage-specific mutations observed in the sample $j$. 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 $r_m$, where $r_m = \frac{n_m}{r}$, $n_m$ is the number of sequence reads associated with the mutated allele, and $r$ 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 PRJNA627817. 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, Gríkevičius 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/NEJMr1813907.

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...
17. Jahn K, Drefuss DT, Topolsky I, Kull A, Ganesanandamoorthy P, Fernandez-Cassi X, Bänziger C, Devaux AJ, Stachler E, Caduff L, Carti F, Çoröz AT, Fuquía J, Fournier G, Gogol-Dörner 801:149757. https://doi.org/10.1016/j.scitotenv.2021.149757
18. Karthikeyan S, Levy J, De Hoff P, Humphrey G, Birmingham A, Jepsen K, Farmer S, Tubb HM, Valles T, Tribulhorn CE, Tsai R, Aigner S, Sathe S, Moshiro N, Henson B, Earl AM, Hakim A, Baer NA, Barber T, Belda-Ferre P, Chacón M, Cheung W, Cresini ES, Eisinger ER, Lastrell AE, Lawrence ES, Moriatz CA, Ngo TT, Ostrander T, Plascencia A, Salido RA, Seaver P, Smoot EW, McDonald D, Neuhard RM, Scioscia AL, Satterlund AM, Simmons EH, Abelmann DR, Brenner D, Brunner JC, Bucklev A, Elliston M, Gattas J, Goniás 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, Feslint X, Borotina D, Blume A, Meixner M, Grau JH,uire K, Hackenbeck T, Zietschmann G, Griss N, Böckelmann U, Uyår B, Franke V, Barke N, Altmüller J, Rajevsky N, Landthaler M, Aikin A. 2022. SARS-CoV-2 infection dynamics revealed by wastewater sequencing analysis and deconvolution. medRxiv. https://doi.org/10.1101/2021.11.26.2129398.
20. Bajajins JO, Zulli A, Ott IM, Peterson ME, Alpert T, Fauver JR, Kalinchic C, Vogels CB, Briul M, Duvall C, McElroy K, Gheani N, Imaekia M, Mckenzie-Bennett M, Robison K, Plock C, Schilling P, Pierson M, Littlefield R, Spencer M, Simen BB, Hanage WP, Grubaugh ND, Pecci 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/10.1101/2021.03.11.21269592.
21. Danecék P, Bonfeld JK, Liddle J, Marshall J, Ohan V, Pollard MW, Whitham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMTools and BCFtools. Gigascience 10:1042. https://doi.org/10.1093/gigascience/giaa008.
22. DeFristo MA, Banks E, Poplin R, Carinimal KV, Maguire JR, Hart C, Phillipakis AA, del Angel G, Rivas MA, Hannen M, McKenna A, Fennel TJ, Kermutsy MA, Avila-Abarca AY, Cilasot M, Gabriel SB, Wold S, 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, Campriuli-Ferrer D, Marcos MÁ, Villena A, Navarro M, Rodríguez-Elena L, Riera J, Catalá A, Martínez BJ, Blanc J, Zaid MC, Nagesh G, de la Maza M, Casanovas J, Homar M, López J, Güell L, Ibanez M, Pla J, Nolla J, Guerra-Cerda J, Santisteban I, Puccio M, Moyer S, Holm RH, Yeager RA, Sokoloski KJ, Fuqua J, Santos-Silva S, Scola BL. 2021. Monitoring the circulation of SARS-CoV-2 populations in wastewater by amplicon sequencing and de novo assembly. Viruses 13:1647. https://doi.org/10.3390/v13081540.
24. Crampin AC, Clark TG. 2018. Identifying mixed SARS-CoV-2 single-nucleotide variants. Appl Environ Microbiol 84:e01448-21. https://doi.org/10.1128/AEM.00874-22.
25. Farmer S, Tubb HM, Valles T, Tribulhorn CE, Tsai R, Aigner S, Sathe S, Moshiro N, Henson B, Earl AM, Hakim A, Baer NA, Barber T, Belda-Ferre P, Chacón M, Cheung W, Cresini ES, Eisinger ER, Lastrell AE, Lawrence ES, Moriatz CA, Ngo TT, Ostrander T, Plascencia A, Salido RA, Seaver P, Smoot EW, McDonald D, Neuhard RM, Scioscia AL, Satterlund AM, Simmons EH, Abelmann DR, Brenner D, Brunner JC, Bucklev A, Elliston M, Gattas J, Goniás 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.
26. Sobkowiak B, Glynn JR, Houben RMGJ, Mallard K, Phelan JE, Guerra-Assunção J, Banda L, Mzembe T, Vievores M, McNerney R, Parkinson J, Crampin AC, Clark TG. 2018. Identifying mixed SARS-CoV-2 single-nucleotide variants. Appl Environ Microbiol 84:e01448-21. https://doi.org/10.1128/AEM.00874-22.
27. Barke N, Altmüller J, Rajewsky N, Landthaler M, Aikin A, Mckenzie-Bennett M, Robison K, Plock C, Schilling P, Pierson M, Littlefield R, Spencer M, Simen BB, Hanage WP, Grubaugh ND, Pecci J, Baym M, Yale 2022. SARS-CoV-2 Genomic Surveillance Initiative. 2021. Variant abundance estimation for SARS-CoV-2 in wastewater using RNA-seq quantification. medRxiv. https://doi.org/10.1101/2021.03.11.21269592.
28. Danecék P, Bonfeld JK, Liddle J, Marshall J, Ohan V, Pollard MW, Whitham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMTools and BCFTools. Gigascience 10:1042. https://doi.org/10.1093/gigascience/giaa008.
29. DeFristo MA, Banks E, Poplin R, Carinimal KV, Maguire JR, Hart C, Phillipakis AA, del Angel G, Rivas MA, Hannen M, McKenna A, Fennel TJ, Kermutsy MA, Avila-Abarca AY, Cilasot M, Gabriel SB, Wold S, 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.
30. Peiró-Mestres A, Fuertes I, Campriuli-Ferrer D, Marcos MÁ, Villena A, Navarro M, Rodríguez-Elena L, Riera J, Catalá A, Martínez BJ, Blanc J, Zaid MC, Nagesh G, de la Maza M, Casanovas J, Homar M, López J, Güell L, Ibanez M, Pla J, Nolla J, Guerra-Cerda J, Santisteban I, Puccio M, Moyer S, Holm RH, Yeager RA, Sokoloski KJ, Fuqua J, Santos-Silva S, Scola BL. 2021. Monitoring the circulation of SARS-CoV-2 populations in wastewater by amplicon sequencing and de novo assembly. Viruses 13:1647. https://doi.org/10.3390/v13081540.
31. Crampin AC, Clark TG. 2018. Identifying mixed SARS-CoV-2 single-nucleotide variants. Appl Environ Microbiol 84:e01448-21. https://doi.org/10.1128/AEM.00874-22.
32. Farmer S, Tubb HM, Valles T, Tribulhorn CE, Tsai R, Aigner S, Sathe S, Moshiro N, Henson B, Earl AM, Hakim A, Baer NA, Barber T, Belda-Ferre P, Chacón M, Cheung W, Cresini ES, Eisinger ER, Lastrell AE, Lawrence ES, Moriatz CA, Ngo TT, Ostrander T, Plascencia A, Salido RA, Seaver P, Smoot EW, McDonald D, Neuhard RM, Scioscia AL, Satterlund AM, Simmons EH, Abelmann DR, Brenner D, Brunner JC, Bucklev A, Elliston M, Gattas J, Goniás 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.
33. Schumann V-F, Cuadrat RRDC, Wyler E, Wurmus R, Deter A, Quedenau C, Dohmen J, Feslint X, Borotina D, Blume A, Meixner M, Grau JH,uire K, Hackenbeck T, Zietschmann G, Griss N, Böckelmann U, Uyår B, Franke V, Barke N, Altmüller J, Rajevsky N, Landthaler M, Aikin A. 2022. SARS-CoV-2 infection dynamics revealed by wastewater sequencing analysis and deconvolution. medRxiv. https://doi.org/10.1101/2021.11.26.2129398.