frontiers in Microbiology

ORIGINAL RESEARCH
published: 13 October 2021
doi: 10.3389/fmicb.2021.747458

Strategy and Performance Evaluation of Low-Frequency Variant Calling for SARS-CoV-2 Using Targeted Deep Illumina Sequencing

Laura A. E. Van Poelvoorde[1,2,3†], Thomas Delcourt[1†], Wim Coucke[4], Philippe Herman[5], Sigrid C. J. De Keersmaecker[1], Xavier Saelens[2,3], Nancy H. C. Roosens[1‡] and Kevin Vanneste[1‡]*

[1] Transversal Activities in Applied Genomics, Sciensano, Brussels, Belgium, [2] Department of Biochemistry and Microbiology, Ghent University, Ghent, Belgium, [3] VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium, [4] Quality of Laboratories, Sciensano, Brussels, Belgium, [5] Expertise and Service Provision, Sciensano, Brussels, Belgium

The ongoing COVID-19 pandemic, caused by SARS-CoV-2, constitutes a tremendous global health issue. Continuous monitoring of the virus has become a cornerstone to make rational decisions on implementing societal and sanitary measures to curtail the virus spread. Additionally, emerging SARS-CoV-2 variants have increased the need for genomic surveillance to detect particular strains because of their potentially increased transmissibility, pathogenicity and immune escape. Targeted SARS-CoV-2 sequencing of diagnostic and wastewater samples has been explored as an epidemiological surveillance method for the competent authorities. Currently, only the consensus genome sequence of the most abundant strain is taken into consideration for analysis, but multiple variant strains are now circulating in the population. Consequently, in diagnostic samples, potential co-infection(s) by several different variants can occur or quasispecies can develop during an infection in an individual. In wastewater samples, multiple variant strains will often be simultaneously present. Currently, quality criteria are mainly available for constructing the consensus genome sequence, and some guidelines exist for the detection of co-infections and quasispecies in diagnostic samples. The performance of detection and quantification of low-frequency variants using whole genome sequencing (WGS) of SARS-CoV-2 remains largely unknown. Here, we evaluated the detection and quantification of mutations present at low abundances using the mutations defining the SARS-CoV-2 lineage B.1.1.7 (alpha variant) as a case study. Real sequencing data were *in silico* modified by introducing mutations of interest into raw wild-type sequencing data, or by mixing wild-type and mutant raw sequencing data, to construct mixed samples subjected to WGS using a tiling amplicon-based targeted metagenomics approach and Illumina sequencing. As anticipated, higher variation and lower sensitivity were observed at lower coverages and allelic frequencies. We found that detection of all low-frequency variants at an abundance of 10, 5, 3, and 1%, requires at least a sequencing coverage of 250, 500, 1500, and 10,000×, respectively. Although increasing variability of estimated allelic frequencies at

OPEN ACCESS

Edited by:
Antonio Battisti, Experimental Zooprophylactic Institute of the Lazio and Tuscany Regions (IZSLT), Italy

Reviewed by:
Matheus Filgueira Bezerra, Aggeu Magalhães Institute (IAM), Brazil
Patricia Alba, Experimental Zooprophylactic Institute of the Lazio and Tuscany Regions (IZSLT), Italy
William Rawlinson, NSW Health Pathology, Australia
Kenji Sadamasu, Tokyo Metropolitan Institute of Public Health, Japan

***Correspondence:**
Kevin Vanneste
kevin.vanneste@sciensano.be

[†] These authors share first authorship

[‡] These authors share last authorship

Specialty section:
This article was submitted to Infectious Agents and Disease, a section of the journal Frontiers in Microbiology

Received: *26 July 2021*
Accepted: *21 September 2021*
Published: *13 October 2021*

Citation:
Van Poelvoorde LAE, Delcourt T, Coucke W, Herman P, De Keersmaecker SCJ, Saelens X, Roosens NHC and Vanneste K (2021) Strategy and Performance Evaluation of Low-Frequency Variant Calling for SARS-CoV-2 Using Targeted Deep Illumina Sequencing. Front. Microbiol. 12:747458. doi: 10.3389/fmicb.2021.747458

decreasing coverages and lower allelic frequencies was observed, its impact on reliable quantification was limited. This study provides a highly sensitive low-frequency variant detection approach, which is publicly available at https://galaxy.sciensano.be, and specific recommendations for minimum sequencing coverages to detect clade-defining mutations at certain allelic frequencies. This approach will be useful to detect and quantify low-frequency variants in both diagnostic (e.g., co-infections and quasispecies) and wastewater [e.g., multiple variants of concern (VOCs)] samples.

Keywords: wastewater surveillance, SARS-CoV-2, Illumina, NGS, variant of concern, co-infection, quasispecies

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of the ongoing COVID-19 pandemic (Kim et al., 2020). To limit the spread of disease, governments were forced to take drastic measures due to the high potential for human-to-human transmission and the lack of immunity in the population (Leclerc et al., 2020). SARS-CoV-2 spreads very easily during close person-to-person contact (Azuma et al., 2020). Consequently, the individual diagnostic testing for SARS-CoV-2 on respiratory samples using reverse transcription quantitative polymerase chain reaction (RT-qPCR) is essential for the diagnosis of patients presenting COVID-19 symptoms for appropriate clinical treatment and isolation, as well as for tracing potential contact transmissions, including asymptomatic individuals. Systematic individual SARS-CoV-2 diagnostics are also used to test certain population cohorts, such as primary caregivers, to avoid transmission of the virus to vulnerable people, such as the elderly (Bayle et al., 2021).

Data from individual diagnostics are also collected and analyzed for surveillance by National Reference Centres to assist governments to monitor the epidemiological situation. The efficiency of this strategy for epidemiological monitoring depends greatly on the extent of testing the complete population. Additionally, it may be biased by the willingness of individuals, covering all population ages, to get tested, whether individuals are aware of being infected, and visitors to a certain country not always being included in the testing strategy. Moreover, despite having a relatively low per-sample cost, the high volume of required tests incurs substantial costs for public health systems for which testing capacities can be exceeded during periods of intense circulation of the virus (Contreras et al., 2021). The detection of newly emerging SARS-CoV-2 strains may be delayed by the lack of testing during such periods. As SARS-CoV-2 viral particles and mRNA have been isolated from feces of COVID-19 patients (Wu et al., 2020b; Zhang et al., 2020), monitoring of wastewater for SARS-CoV-2 has been explored as a complementary and independent alternative for epidemiological surveillance for the competent authorities (European Commission, 2021). Various studies have observed an association between an increase in reported COVID-19 cases and an increase of SARS-CoV-2 RNA concentrations in wastewater (Ahmed et al., 2020; Medema et al., 2020; Wu et al., 2020a). Wastewater-based monitoring could therefore be a cost-effective, non-invasive, easy to collect, and unbiased approach to track circulating virus strains in a community (Thompson et al., 2020). Compared to clinical surveillance, wastewater surveillance could also provide opportunities to estimate the prevalence of the virus and assess its geographic distribution and genetic diversity (Sinclair et al., 2008; Xagoraraki and O'Brien, 2020), and can be used as a non-invasive early-warning system for alerting public health authorities to the potential (re-)emergence of COVID-19 infections (Panchal et al., 2021). Alternatively, the absence of the virus in wastewater surveillance could indicate that an area can be considered at low risk for SARS-CoV-2 infections (European Commission, 2021).

Although the mutation rate of SARS-CoV-2 is estimated as being low compared to other RNA viruses (Duchene et al., 2020), several new variants carrying multiple mutations have already emerged. Some of these variants are characterized by a potential enhanced transmissibility, and can cause more severe infections and/or potential vaccine escape (SAGE-EMG, SPI-B, Tranmission Group, 2020; Davies et al., 2021; GOV.UK - Scientific Advisory Group for Emergencies, 2021; Greaney et al., 2021; Hoffmann et al., 2021). Consequently, monitoring current and potential future variants is crucial to control the epidemic by taking timely measures because these variants can affect epidemiological dynamics, vaccine effectiveness and disease burden.

To monitor SARS-CoV-2 variants, RT-qPCR methods were designed to detect a selection of the mutations that define specific variants of concern (VOCs). VOCs are, however, defined by a combination of multiple mutations and only few mutations can be targeted by RT-qPCR assays. This approach is not sustainable because it is likely that the ongoing vaccination and increased herd immunity will result in the selection of new mutations and emergence of new VOCs (Gómez et al., 2021), as has been observed with other viruses (Boni, 2008; Shao et al., 2017). Since only a few mutations can be targeted by a RT-qPCR assay, an additional step of whole genome sequencing (WGS) is required to fully confirm the variant's sequence (Bal et al., 2021).

Whole genome sequencing has been used to understand the viral evolution, epidemiology and impact of SARS-CoV-2 resulting in, as of July 2021, more than 2,000,000 publicly available SARS-CoV-2 genome sequences, mainly derived from respiratory samples that are frequently submitted to the Global Initiative on Sharing Avian Influenza Data (GISAID) database (Shu and McCauley, 2017). Most of these sequences were obtained using amplicon sequencing in combination with the Illumina or Nanopore technology, with Illumina still being

the most commonly used method (Shu and McCauley, 2017; Charre et al., 2020). This large amount of genomes allows reliable detection of variants based on the consensus genome sequence in patient samples (van Dorp et al., 2020; Firestone et al., 2021; Hartley et al., 2021; Lin et al., 2021). The European Centre for Disease Prevention and Control (ECDC) has defined several quality criteria for diagnostic samples depending on the application. For most genomic surveillance objectives, a consensus sequence of the (near-)complete genome is sufficient and a minimal read length of 100 bp and minimal coverage of 10× across more than 95% of the genome is recommended. To reliably trace direct transmission and/or reinfection, a higher sequencing coverage of 500× across more than 95% of the genome is recommended for determining low-frequency variants (LFV) that can significantly contribute to the evidence for reinfection or direct transmission. In-depth genome analysis, including recombination, rearrangement, haplotype reconstruction and large insertions and deletions (indel) detection, should be investigated using long-read sequencing technologies with a recommended read length of minimally 1000 bp and a sequencing coverage of 500× across more than 95% of the genome (ECDC, 2021). A few studies evaluated quasispecies in diagnostic samples by only evaluating positions with a minimum depth of 100× (Lythgoe et al., 2021), by employing a minimum AF of 2% and a minimum depth of 500× (Siqueira et al., 2021) or by using Lofreq with a false discovery rate cut-off of 1%, minimum coverage of 10×, dynamic Bonferroni correction for variant quality and strand bias filtering (Karim et al., 2021). Due to the high cost of sequencing large quantities of samples from individual patients, samples that tested positive for a selection of mutations related to VOCs using RT-qPCR and have a sufficiently high viral load are typically sequenced. Consequently, only a subset of all circulating variants is detected during routine clinical surveillance. Since wastewater samples contain both SARS-CoV-2 RNA from symptomatic and asymptomatic individuals, sequencing wastewater samples can provide a more comprehensive picture of the genomic diversity of SARS-CoV-2 circulating in the population compared to individual diagnostic testing and sequencing. Wastewater surveillance of SARS-CoV-2 may therefore be of considerable added value for SARS-CoV-2 genomic surveillance by providing a cost-effective, rapid, and reliable source of information on the spread of SARS-CoV-2 variants in the population.

Sequencing of wastewater samples is, however, currently mainly used to reconstruct the consensus genome sequence of the most prevalent SARS-CoV-2 strain in the sample and LFV are often not investigated (Nemudryi et al., 2020; Bar-Or et al., 2021; Crits-Christoph et al., 2021; Sharif et al., 2021). This consensus sequence can be useful to demonstrate that the detected strain in wastewater corresponds to the dominant strain that circulates in individuals within the same community (Crits-Christoph et al., 2021). However, similarly to diagnostic samples, only limited quality criteria are in place when sequencing wastewater samples and those available often only apply for consensus sequence construction. The EU recommends the generation of one million reads per sample and a read length of more than 100 bp (European Commission, 2021). A few studies evaluated LFV in wastewater samples, by using local haplotype reconstruction with ShoRAH (Jahn et al., 2021) or iVar and setting up a minimum coverage of 50×, Phred score of ≥30 and a minimal allelic frequency (AF) of 10% (Izquierdo-Lara et al., 2021) or a minimum base quality filter of 20 with a minimum coverage of 100× (Rios et al., 2021). However, none of these studies evaluated their approach on well-defined populations nor determined detection thresholds for retaining LFV. Since multiple VOCs may co-circulate in a given population, their relative abundance is expected to vary and potentially be very low in wastewater samples. While genome consensus variant calling workflows can only identify mutations present at high AFs, LFV calling methods have been specifically designed to call mutations at lower-than-consensus AFs, and are required to detect VOCs in wastewater samples that are present at an AF below 50%. Appropriate tools and statistical approaches should be provided to ensure reliable and comparable collection and analysis of data, because the detection of LFV is challenging due to the drop in confidence of called mutations at low AFs and sequencing coverages (Macalalad et al., 2012; Wilm et al., 2012; Isakov et al., 2015). High-quality sequencing reads are required to ensure that single nucleotide variants (SNVs) and indels can be reliably called and quantified. Most LFV calling algorithms therefore consider multiple sequencing characteristics such as strand bias, base quality, mapping quality, sequence context, and AF (McCrone et al., 2016) to delineate true variants from sequencing errors. Although the viral diversity in multiple WGS-based studies has been explored using several variant calling methods (Kundu et al., 2013; Rogers et al., 2015; Simon et al., 2019), they are often not benchmarked against defined viral populations, rendering the feasibility of using these methods for detecting SARS-CoV-2 VOCs in mixed samples for wastewater surveillance largely unknown.

In this study, we evaluate the performance of LFV detection and quantification based on targeted SARS-CoV-2 sequencing for mutations present at low abundances via the Illumina technology. We used mutations that define the B.1.1.7 lineage as a proof-of-concept. Using two real sequencing datasets that were *in silico* modified by either introducing mutations of interest into raw wild-type sequencing datasets or mixing wild-type and mutant raw sequencing data, we provide guidelines for minimum sequencing coverages to detect clade-defining mutations at specific AFs. This approach can be used to detect and quantify LFV in diagnostic samples (e.g., to detect co-infections and quasispecies) and wastewater samples (e.g., to detect multiple strains circulating in the population).

MATERIALS AND METHODS

Employed Sequencing Data and Generation of Consensus Genome Sequences

SARS-CoV-2 raw sequencing data from 316 samples was downloaded from the Sequence Read Archive (SRA) (Leinonen et al., 2011). A random selection of samples was done on the

27th of January 2021 from the COVID-19 Genomics UK (COG-UK) consortium (PRJEB37886) including only samples with a submission date in January 2021, sequenced with Illumina Novaseq 6000 and using an amplicon-based enrichment strategy (**Supplementary File 1**).

To ensure correct pairing of fastq files, all samples were re-paired using BBMap v38.89 repair.sh with default settings (Bushnell and BBMap, 2021) (**Figure 1**: Step 1). The consensus genome sequences were generated for all these samples (**Figure 1**: Step 2). The workflow was built using the Snakemake workflow management system using python 3.6.9 (Mölder et al., 2021). Next, the re-paired paired-end reads were trimmed using Trimmomatic v0.38 (Bolger et al., 2014) setting the following options: "LEADING:10", "TRAILING:10", "SLIDINGWINDOW:4:20" and "MINLEN:40". As reference genome for read mapping of the SRA samples, the sequence with GISAID accession number EPI_ISL_837246 was used for the wild-type samples, while EPI_ISL_747518 was used for the mutant samples. Both references were chosen based on the fact that they should have a complete genome according to GISAID. Additionally, these were chosen to be as close to the SRA data as possible based on their location of sampling (i.e., United Kingdom), sampling date that was in the same period as the data obtained from SRA (i.e., December 2020–January 2021), and whether or not it was classified as belonging to the B.1.1.7 lineage. These reference genomes were indexed using Bowtie2-build v2.3.4.3 (Langmead and Salzberg, 2012). Trimmed reads were aligned to their respective reference genomes using Bowtie2 v2.3.4.3 using default parameters. The resulting SAM files were converted to BAM files using Samtools view v1.9 (Danecek et al., 2021) and sorted and indexed using the default settings of respectively Samtools sort and Samtools index v1.9. Using the sorted BAM file, a pileup file was generated with Samtools mpileup v1.9 using the options "--count-orphans" and "--VCF." Next, the variants were called with bcftools call v1.9 using the options "-O z", "--consensus-caller", "--variants-only" and "--ploidy 1", and converted and indexed to uncompressed VCF files with respectively bcftools view v1.9 using the options "--output-type v" and bcftools index v1.9 using the option "--force." Lastly, a temporary consensus sequence was generated using bcftools consensus v1.9 with default settings, providing the reference genome and produced VCF file as inputs. Afterward, the previous steps were repeated once with the same options using the generated temporary consensus sequence as fasta reference to generate the final consensus sequence. These sequences were used to confirm either the presence or absence of the clade-defining mutations of the B.1.1.7 mutant for both the mutant and wild-type samples respectively (**Table 1**). To extract the sequencing coverage for each position and subsequently calculate the median coverage for each sample, Samtools depth v1.9 was used on the BAM files. Additionally, bamreadcount v0.8.0[1] was run on all samples using the BAM files to determine the coverage at each position.

From the initial 316 samples, ten mutant samples were selected that presented similar coverage depths at the positions of interest after normalization (see below). These samples contained the mutations assigned to the B.1.1.7 variant. Ten wild-type samples were also chosen that did not contain any of these mutations (**Tables 1**, **2**) and also presented similar coverage depth at the positions of interest after normalization. Lineage B.1.1.7, termed Variant of Concern (VOC) 202012/01 by Public Health England (PHE) (Public Health England, 2021), 20I/501Y.V1 by Nextstrain (Centers for Disease Control and Prevention [CDC], 2021) and alpha variant by the World Health Organization (WHO, 2021), was first reported in the United Kingdom but became the dominant strain in many European countries until the emergence of the delta variant since mid-April 2021 (Mishra et al., 2021). The B.1.1.7 variant was found to be more transmissible (Davies et al., 2021) and may cause more severe infections (SAGE-EMG, SPI-B, Tranmission Group, 2020; GOV.UK - Scientific Advisory Group for Emergencies, 2021). Lineage B.1.1.7 is defined by multiple spike protein changes, including deletion 69-70 and deletion 144 in the N-terminal domain, amino changes N501Y in the receptor-binding domain, and amino acid changes A570D, P681H, T716I, S982A, D1118H, as well as mutations in other genomic regions (Rambaut et al., 2021). More recently PHE has reported B.1.1.7 cases with an additional mutation, E484K (Public Health England, 2021). Median coverages of the selected samples were consistently high (minimum 13,848×; maximum 36,255×) and median read lengths were always 221 and 201 for the forward and reverse reads respectively (**Table 2**). Additionally, as suggested by ECDC, more than 95% of the genome was covered by reads with a minimal coverage of 500× (ECDC, 2021).

Low-Frequency Variants Detection

The absence of pre-existing wild-type and mutant LFV at the positions defining lineage B.1.1.7 (**Table 1**) was verified in both the mutant and wild-type samples (**Figure 1**: Step 3), respectively, by calling all LFV in these samples and subsequently checking the positions of interest. Python 3.6.9 was used with the packages pysam 0.16.0.1 (Li et al., 2009) and numpy 1.19.5 (Harris et al., 2020). Each generated (final) consensus FASTA file for each sample coming from SRA was used as reference for its respective sample and indexed using Samtools faidx v1.9 and Bowtie2-build v2.3.4.3. Bowtie2 v2.3.4.3 was then used to align the reads of each sample to its reference sequence, producing a SAM file that was converted into BAM using Samtools view v1.9. Next, reads were sorted using Picard SortSam v2.18.14[2] with the option "SORT_ORDER = coordinate" and Picard CreateSequenceDictionary v2.18.14 (Broad Institute)[3] was used to generate a dictionary of the reference FASTA file. Picard AddOrReplaceReadGroups v2.18.14 (Broad Institute) (see footnote 3) was afterward run on the reads with the flags "LB", "PL", "PU", and "SM" set to the arbitrary placeholder value "test." The resulting BAM files were indexed using Samtools index v1.9 and used as input for GATK RealignerTargetCreator 3.7 (McKenna et al., 2010), which was followed by indel realignment using GATK IndelRealigner v3.7

[1]https://github.com/genome/bam-readcount

[2]https://github.com/broadinstitute/picard

[3]Broad Institute Picard (2021). Available online at: http://broadinstitute.github.io/picard/ [Accessed March 26, 2021]

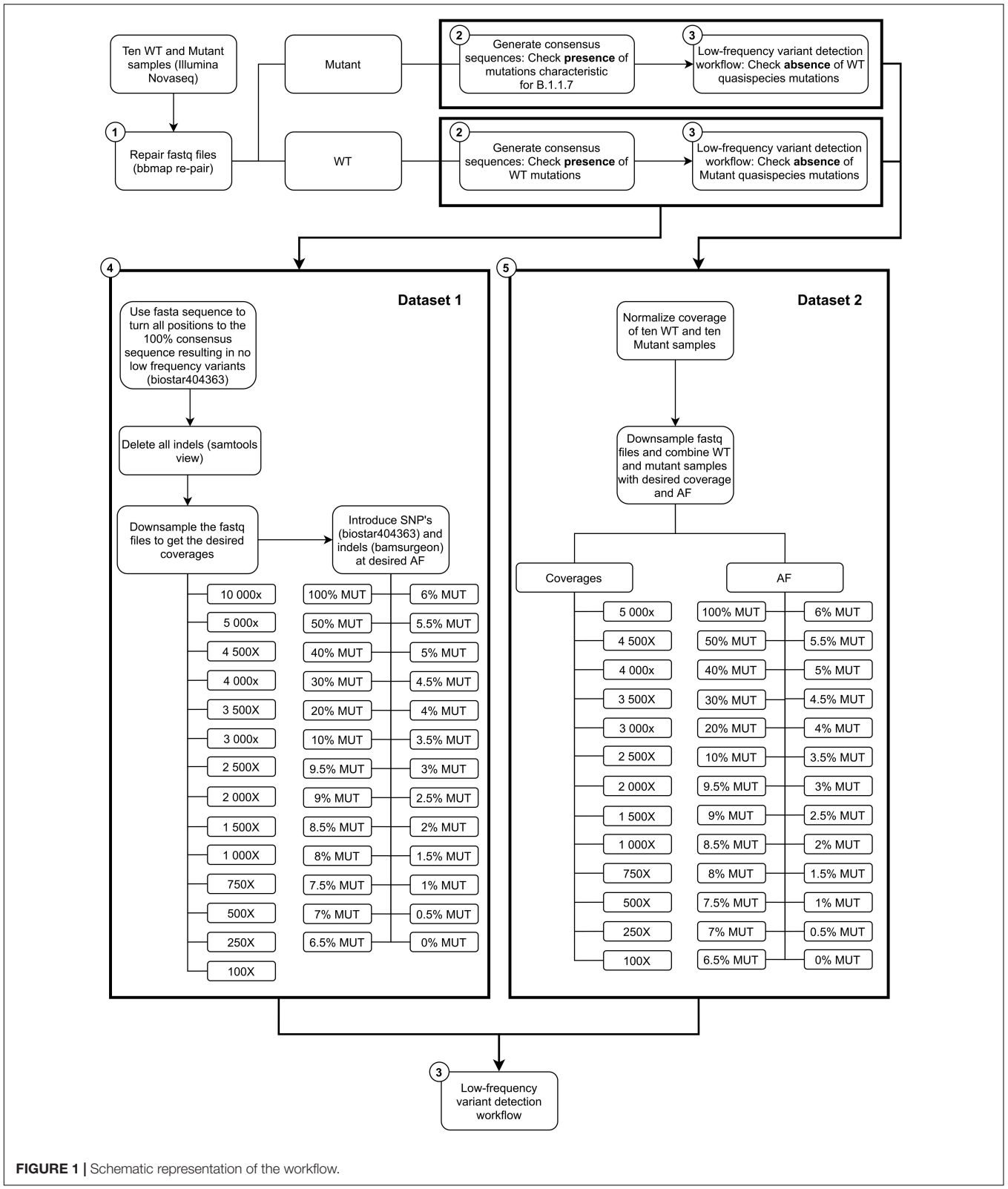

FIGURE 1 | Schematic representation of the workflow.

(McKenna et al., 2010). Next, generated BAM files were indexed using Samtools index v1.9. The call function of the LoFreq v2.1.3.1 package (Wilm et al., 2012) was used to call LFV in the BAM files and generate a VCF file using the options "--call-indels" and "--no-default-filter" and using the consensus sequence as reference to call LFV. Next, the unfiltered VCF

TABLE 1 | Mutations linked to SARS-CoV-2 lineage B.1.1.7 (Rambaut et al., 2021).

Gene	Nucleotide-level mutation	Amino Acid-level mutation	Number of amplicons covering the position?
ORF1ab	C913T	Synonymous	1
	C3267T	T1001I	1
	C5388A	A1708D	1
	C5986T	Synonymous	1
	T6954C	I2230T	1
	11288–11296 deletion	SGF 3675–3677 deletion	1
	C14676T	Synonymous	1
	C15279T	Synonymous	1
	C16176T	Synonymous	2
S	21765–21770 deletion	HV 69-70 deletion	1
	21991–21993 deletion	Y144 deletion	2
	A23063T	N501Y	1
	C23271A	A570D	1
	C23604A	P681H	1
	C23709T	T716I	1
	T24506G	S982A	1
	G24914C	D1118H	2
M	G26801C*	Synonymous	1
Orf8	C27972T	Q27stop	WT: 2; B.1.1.7: 1**
	G28048T	R52I	1
	A28111G	Y73C	2
N	G28280C, A28281T, T28282A	D3L	2
	C28977T	S235F	1

*The first, second, and third columns present respectively the gene name, cDNA-level mutation and protein-level mutation. The last column describes whether the position is covered by one or two amplicons from the enrichment panel (**Supplementary Table 1**). (*) One adaptation was observed for position 26 801. In the wild-type strains a G was observed in contrast to Rambaut et al. (2021) where a T was observed. (**) Due to the tiled amplicon approach used to amplify the samples prior to sequencing, the regions where amplicons overlapped resulted in a double coverage. Mutation C27972T was positioned in such an overlap in the wild-type, but not in the mutant. WT, wild-type.*

file was filtered using the filter function of the LoFreq v2.1.3.1 package, setting the strand bias threshold for reporting a variant to the maximum allowed value by using the option "--sb-thresh 2147483647" to allow highly strand-biased variants to be retained, to account for the non-random distribution of reads due to the design of the amplification panel. All employed scripts are available in **Supplementary File 2**. Additionally, the workflow is also available at the public Galaxy instance of our institute at https://galaxy.sciensano.be as a free resource for academic and non-profit usage. The presence of the nucleotides assigned to the B.1.1.7 lineage or the wild-type (**Table 1**) was verified for the mutant and wild-type samples, respectively. Additionally, it was checked that there were no LFV at these positions, so that the wild-type nucleotide or mutant nucleotide was always present at 100% for the retained 10 WT and 10 mutant samples.

Dataset 1: *In silico* Insertion of Mutations of Interest Into Raw Sequencing Datasets

For the first dataset (**Figure 1**: Step 4), all low-frequency single nucleotide polymorphisms (SNPs) were removed from the raw sequencing data of all samples. SNPs were removed using Jvarkit employing biostar404363 (Lindenbaum, 2015) by converting all nucleotides to the consensus fasta sequence. Next, all ten WT samples were down-sampled using "seqtk sample" with argument "-s100"[4] to 14 different (median) coverages (100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, and 10,000×). The 22 SNP mutations characteristic for the B.1.1.7 lineage (**Table 1**) were introduced at 26 different AF (mutant: 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 20, 30, 40, 50, and 100%) at the various coverages mentioned above employing biostar404363. This resulted in 10 samples at 364 conditions (i.e., combination of coverage and AF). Next, all reads containing indels were removed from these samples using samtools view v1.9. Finally, the three deletions associated with the B.1.1.7 lineage were introduced at the 26 AF mentioned above using BAMSurgeon 1.2 (Ewing et al., 2015), which was adapted to decrease runtime, with the options "-p 10", "--force", "-d 0", "--ignorepileup", "--mindepth 1", "--minmutreads 1", "--maxdepth 1000000", "--aligner mem", and "--tagreads". A minority of reads that were lacking a mate in the targeted regions were removed by using an in-house script making use of Python 3.6.9 and the package pysam 0.16.0.1. Samples in BAM format were then converted back to FASTQ format using bedtools bamtofastq

[4]https://github.com/lh3/seqtk

TABLE 2 | List of SRA accession numbers used for employed wild-type and lineage B.1.1.7 samples in this study.

Sample	WT/lineage B.1.1.7	Median coverage
ERR5058968	Lineage B.1.1.7	13,848
ERR5059033	Lineage B.1.1.7	21,874
ERR5059072	Lineage B.1.1.7	14,628
ERR5059092	Lineage B.1.1.7	16,106
ERR5059123	Lineage B.1.1.7	17,349
ERR5059204	Lineage B.1.1.7	18,149
ERR5059226	Lineage B.1.1.7	22,194
ERR5059238	Lineage B.1.1.7	27,681
ERR5059260	Lineage B.1.1.7	23,975
ERR5059282	Lineage B.1.1.7	27,349
ERR5039162	WT	20,071
ERR5040499	WT	24,440
ERR5059083	WT	18,220
ERR5059114	WT	14,580
ERR5059133	WT	19,866
ERR5059154	WT	28,295
ERR5059253	WT	23,798
ERR5059257	WT	25,894
ERR5059283	WT	36,255
ERR5059286	WT	29,847

Sample IDs, categorized as WT or mutant and the median coverage calculated using Samtools depth v1.9 (Danecek et al., 2021). WT, wild-type.

v2.27.1 (Quinlan and Hall, 2010). Finally the LFV detection workflow (**Figure 1**: Step 3) described in section "Low-Frequency Variants Detection" was used on these 10 samples for all 364 conditions using the FASTA file that was generated for the wild-type samples from SRA as reference with LoFreq.

Dataset 2: Introduction of Mutations of Interest by Mixing Wild-Type and Mutant Raw Sequencing Read Datasets

For the second dataset (**Figure 1**: Step 5), the coverage of all 20 samples (**Table 2**) was normalized to 5000× using BBMap v38.89 bbnorm.sh (Bushnell and BBMap, 2021) with the options "target = 5000", "mindepth = 5", "fixspikes = f", "passes = 3" and "uselowerdepth = t". However, due to the tiled amplicon approach used to amplify these samples prior to sequencing, regions where amplicons overlapped subsequently had double coverage resulting in two coverages, i.e., 5000 and 10,000×, after normalization (**Supplementary Table 1**). *In silico* datasets were then generated by mixing the appropriate number of reads for every combination of the ten wild-type and ten mutant samples, resulting in a total of 100 mixed samples, which were down-sampled using "seqtk sample" (with option "–s100") to the appropriate fractions for the required combination of 13 final coverages (100, 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000×) and 26 AF (mutant: 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9, 9.5, 10, 20, 30, 40, 50, and 100%). This resulted in 100 mixed samples at 338 conditions (i.e., combination of coverage and AF). Finally, the LFV detection workflow (**Figure 1**: Step 3) described in section "Low-Frequency Variants Detection" was used on these samples for all conditions using the FASTA file that was generated for the wild-type samples from SRA as reference, except for samples with 100% AF for the mutant positions where the FASTA file of the mutant sample was used.

Although the second dataset was normalized for total coverage at every genomic position, the tiled amplicon approach resulted in some genomic positions being covered by two overlapping amplicons. Two groups of mutations were therefore obtained for every coverage (**Table 2**), i.e., for a targeted coverage of 5000×, 17 mutations were present at ∼5000× (C913T, C3267T, C5388A, C5986T, T6954C, 11288-11296 deletion, C14676T, C15279T, 21765-21770 deletion, A23063T, C23271A, C23604A, C23709T, T24056G, G26801C, G28048T, and C28977T) and 7 mutations were present at ∼10,000× (T16176C, 21991-21993 deletion, G24914C, A28111G, G28280C, A28281T, and T28282A). Mutation C27972T was excluded from further analysis, because this position in the wild-type samples was located in a region where amplicons overlapped resulting in a coverage of approximately 10,000×, while in mutant samples it was in a region with no overlap and where a coverage of 5000× was therefore observed (**Supplementary Table 1**). For further analysis, the results were pooled together per theoretical coverage resulting in 24 mutations per coverage but only 17 and 7 mutations at the lowest (i.e., 100×) and highest (i.e., 10,000×) coverage, respectively (**Supplementary Table 2**). The actual median coverage was calculated per theoretical targeted coverage using the output of bamreadcount v0.8.0 of each sample. Using this output, the coverage of each position of interest was extracted (**Supplementary Table 2**).

Qualitative Evaluation of Detection of B.1.1.7 at Different Abundances

Since samples of Dataset 1 were normalized for the total median coverage, different individual positions of interest could exhibit deviating coverages. For the qualitative evaluation of LFV detection (i.e., can mutant positions of interest be correctly detected?), the number of false negatives was counted per condition (i.e., combination of AF and coverage) and divided by the total number of observations [i.e., the number of samples (n = 10) and number of mutations considered for that condition (n = 25)]. A mutant position of interest was considered as correctly detected as soon as it was detected by LoFreq, irrespective of its estimated AF.

Dataset 2 was subjected to the same qualitative evaluation as described for Dataset 1. The number of false negatives per condition was divided by the number of observations (i.e., the number of samples (n = 100) and number of mutations considered for that condition [either n = 7, 17, or 24)].

The visualization of the qualitative evaluation was performed using a contour plot from the R package plotly (RStudio 1.0.153; R3.6.1) (Sievert, 2020). The false negative (FN) proportion in the qualitative evaluation plots ranged from 0 to 1 with a step size of 0.1.

Quantitative Evaluation of Detection of B.1.1.7 at Different Abundances

For the quantitative evaluation of LFV detection (i.e., is the estimated AF of correctly detected mutant positions of interest close to the true AF?) of both datasets, FN values were considered as 'below the quantification limit' with the quantification limit equal to the lowest recorded value for that condition (i.e., combination of AF and coverage). Outliers were identified for each condition using the Grubbs test that was sequentially applied by first searching for two outliers at the same side, followed by a search for exactly one outlier. If the p-value of the Grubbs test was below 0.05, outliers were excluded. The standard deviation (SD) and mean value of AF for every condition were estimated by a maximum likelihood model based on the normal distribution that took the FN into account as censor data. Data were modeled according to a normal distribution. If the percentage of FN results was above 75%, the condition was, however, excluded from quantitative evaluation. Finally, a performance metric describing closeness to the true AF was calculated for each targeted AF individually by dividing each pooled squared SD by the maximal pooled squared SD. This metric will range between 0, relatively the closest to the targeted AF, and 1, relatively the furthest from the targeted AF.

As described for the qualitative evaluation, contour plots from the R package plotly (RStudio 1.0.153; R3.6.1) were used for the visualization of the quantitative evaluation. The performance

metric in the quantitative evaluation plots ranged from 0 to 1 with a step size of 0.1.

RESULTS

Qualitative Evaluation Demonstrates That B.1.1.7 Clade-Defining Mutations Can Be Reliably Detected at Low Allelic Frequency When Sequencing Coverage Is Adequately High

To construct samples using targeted SARS-CoV-2 sequencing with a VOC present at low abundances in the viral population, B.1.1.7 clade-defining mutations were first *in silico* introduced at well-defined AFs and coverages in real sequencing data ("Dataset 1") of ten wild-type samples, without, however, using any coverage normalization so that individual mutations could be present at higher or lower coverages compared to the total median genomic coverage due to unevenness of coverage. To assess whether introduced mutations were correctly detected, or alternatively missed as FN, samples of this dataset were analyzed using a LFV calling workflow based on LoFreq.

Figure 2A depicts the proportion of FN observations, and corresponding values are presented in **Table 3**, for all evaluated coverages and targeted AFs until 20%. Results for all targeted AFs (including higher values) are presented in **Supplementary Figure 1** and **Supplementary Table 3**. All LFV could be detected at an AF of 1% at a median coverage of 10,000×. As the coverage decreased, the AF threshold at which no single FN occurred (i.e., perfect sensitivity) increased to 1.5% at 5000×, 3% at 1000×, 5% at 500×, 9.5% at 250×, and 20% at 100×. When allowing a maximum of 10% FN (i.e., sensitivity of 90%), the AF thresholds decreased substantially to 1% at 5000×, 1.5% at 1000×, 2.5% at 500×, 4% at 250×, and 7.5% at 100×. No false positive mutations related to the mutant and wild-type were observed at, respectively, 0 and 100% AF.

A second approach was also considered for constructing samples using targeted SARS-CoV-2 virus sequencing with a VOC present at low abundances, by *in silico* mixing real raw sequencing reads from ten B.1.1.7 samples into ten wild-type samples ("Dataset 2") for a total of 100 mixes at well-defined AFs and coverages, while applying coverage normalization so that individual mutations were present at approximately similar coverages for all B.1.1.7 clade-defining positions.

Figure 2B depicts the proportion of FN observations, and actual values are presented in **Table 4**, for all evaluated coverages and targeted AF until 20%. Results for higher targeted AF are presented in **Supplementary Figure 2** and **Supplementary Table 4**. All LFV could be detected at an AF of 1% at a median coverage of 9792×. As the coverage decreased, the AF thresholds at which no single FN occurred (i.e., perfect sensitivity) increased to 1.5% at 4851×, 3.5% at 969×, 4% at 482×, 7% at 237×, and 20% at 97×. However, when allowing a maximum of 10% FN (i.e., reducing the sensitivity to 90%), the AF thresholds decreased substantially to 1% at 4851×, 2% at 969×, 3% at 482×, 4% at 237×, and 7% at 97×. No false positive mutations related to the mutant and wild-type were observed at 0 and 100%, respectively. Overall, the results for Dataset 1, using the median coverages, and Dataset 2, using the coverages at the positions of interest, were qualitatively similar.

Quantitative Evaluation Demonstrates That the Resulting Allelic Frequencies for B.1.1.7 Clade-Defining Mutations Are Close to Their Target Values

To evaluate the possibility of quantifying LFV in both datasets, the SDs of available observations were first evaluated for each condition (i.e., combination of AF and coverage). This provisional analysis indicated that for both Dataset 1 (**Supplementary File 3**) and Dataset 2 (**Supplementary File 4**), the SD systematically decreased per target AF as coverage increased. This provisional analysis also indicated that for both datasets, irrespective of coverage, the SD generally increased between a targeted AF of 1 to 10%, after which it plateaued for targeted AFs above 20%. We therefore employed the squared SD per AF divided by the maximal squared SD per target AF to describe closeness of observed AF to the true AF, for which results are presented in **Figure 3A** for Dataset 1. As expected, the variation in AF estimates fluctuates in function of the median coverage and targeted AF, with variation decreasing per target AF as coverage increased, but also variation being generally more pronounced at low AFs irrespective of coverage. Notwithstanding, even for regions in **Figure 3A** exhibiting high variation, the variability overall remained small (**Supplementary File 3**). The interquartile range (IQR) (**Supplementary File 3D**) of the observed AF was still limited at the various targeted AF ranging from 0.62–6.26% at an AF of 50%, 0.36–3.49% at an AF of 10% and 0.27–2.07% at an AF of 5% with the highest IQR observed at lower coverages.

Results for the quantitative evaluation of Dataset 2 are presented in **Figure 3B**, and are in accordance with the trends observed for Dataset 1 with the variation decreasing per target AF as coverage increased, and lower target AFs exhibiting increasing variation irrespective of coverage. Notwithstanding, similarly to Dataset 1, the observed total variation remained small (**Supplementary File 4**). The IQR (**Supplementary File 4D**) of the observed AF was limited at the various targeted AF ranging from 0.73–3.93% at an AF of 50%, 0.41–3.93% at an AF of 10% and 0.29–2.27% at an AF of 5% with the highest IQR observed at lower coverages.

DISCUSSION

Whole genome sequencing is a more powerful approach than RT-qPCR to track both existing and newly emerging SARS-CoV-2 variants. WGS is currently, however, mainly used to construct the consensus genome sequence and determine the most prevalent strain in communities, but interest exists in its potential for detecting LFV both within diagnostic samples to detect co-infections and quasispecies, and wastewater samples to determine all circulating variants in a population

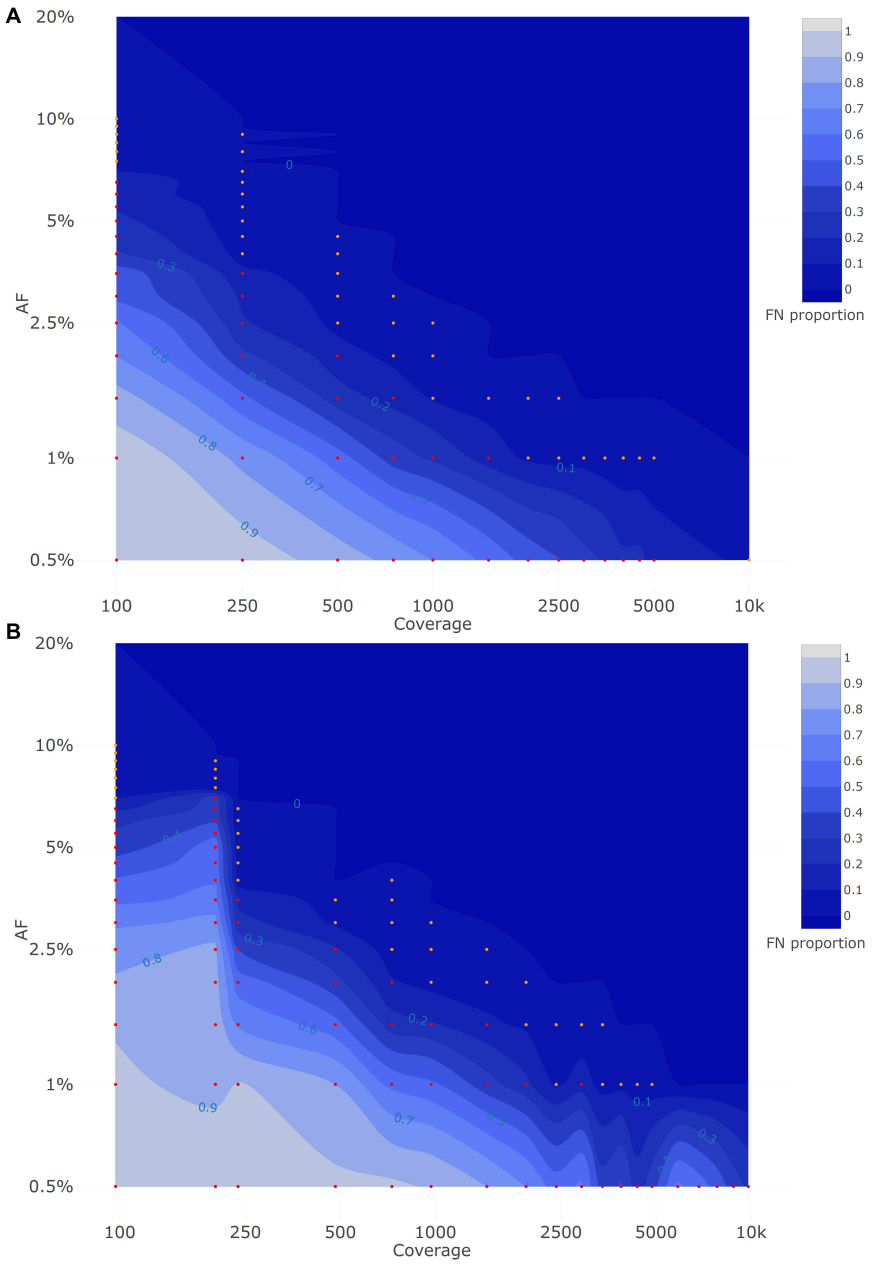

FIGURE 2 | Qualitative evaluation of Dataset 1 **(A)** and Dataset 2 **(B)** based on false negative proportions per condition until a targeted mutant AF of 20%. Orange and red dots represent conditions with a FN proportion between 0 and 0.1, and between 0.1 and 1, respectively. The percentage of FN is colored ranging from 0 (dark) to 1 (light) in intervals of 0.1 as extrapolated using a contour plot in the R package plotly (Sievert, 2020) (actual FN proportions are presented in **Table 3** for Dataset 1 and **Table 4** for Dataset 2). Results for targeted mutant AF values > 20% are presented in **Supplementary Figure 1** for Dataset 1 and **Supplementary Figure 2** for Dataset 2. Both the *X*- and *Y*-axis follow a logarithmic scale.

(European Commission, 2021). To evaluate the potential of targeted amplicon-based SARS-CoV-2 WGS to detect and quantify LFVs at low abundances, we assessed the performance of a workflow designed for LFV detection in WGS data. Mutations defining lineage B.1.1.7 were employed as a proof-of-concept using an approach based on *in silico* modifying real sequencing data to construct two datasets with the Illumina technology. These two datasets comprise in total 35,100 different samples, which results in a thorough *in silico* analysis requiring a considerable amount of computational calculation hours to validate this approach. For the first dataset, lineage B.1.1.7-defining mutations were introduced *in silico* into raw wild-type sequencing datasets. For the second dataset, the same mutations were introduced by mixing wild-type and B.1.1.7 raw

sequencing datasets. In Dataset 1, the coverage profiles of samples corresponded to a typical real dataset including large fluctuations in sequencing coverage at certain positions. In Dataset 2, sequencing coverages were normalized, which allowed evaluating with high precision how reliable AF detection is at specific coverages. Afterward, the ability to both detect and quantify LFV was evaluated. Results demonstrated that WGS enabled detecting LFV with very high performance. As expected, lower coverages and AFs resulted in lower sensitivity and higher variability of estimated AFs. We found, employing the most conservative thresholds from either Datasets 1 or 2, that a sequencing coverage of 250, 500, 1500, and 10,000× is required to detect all LFV at an AF of 10, 5, 3, and 1%, respectively (**Tables 3**, **4**). For quantification of variants, the variability remained overall small for all conditions respecting the thresholds above, resulting in reliable abundance estimations, despite the variability of estimated AF increasing at lower coverages and AF. Of note, it was observed that the profile of the genome coverage differed at some positions between wild-type and mutant samples indicating that the amplicon-based enrichment approach could possibly introduce a bias. Consequently, this should be considered when examining and quantifying the proportion of mutants in samples. Our results can serve as a reference for the scientific community to select appropriate thresholds for the AF and coverage. These could also be context-specific as a smaller or larger degree of false negatives might be warranted for specific applications, and can also be used as a baseline for determining the number of samples that can be multiplexed per run to optimize cost-efficiency of WGS.

With respect to diagnostic samples, this study illustrates it is feasible to use targeted amplicon-based metagenomic approaches to detect co-infections and quasispecies in diagnostic samples. There are currently only limited guidelines available regarding the coverage and AF for such samples and those criteria were not assessed using predefined populations. ECDC has provided limited quality criteria regarding the sequencing coverage, namely 500× across 95% of the genome to detect LFV, but has not indicated the corresponding AF thresholds this corresponds to for reliable LFV detection (ECDC, 2021). Based on the results obtained in this study, a coverage of 500× allowed to detect LFV until an AF of 5% with perfect sensitivity and would therefore be less suited to detect LFV at lower AFs. Lythgoe et al. (2021) recommended a depth of at least 100 reads with an AF of at least 3% to detect the LFV in diagnostic samples with high viral loads (50,000 uniquely mapped reads), while Siqueira et al. (2021) used an AF threshold of 2% and a minimal depth coverage of 500 reads and Karim et al. (2021) adopted an AF of 1% and a minimal depth coverage of 10×. Based on the results in this study, these recommendations appear not sufficiently strict, since we observed that an AF of 1, 2, and 3%, requires at least a sequencing coverage of 10,000, 2500, and 1500× to detect all LFV or 3500, 1000, and 500× to detect 90% of LFV, respectively. However, our study is limited to *in silico* modified data from real diagnostic samples, so these results will need to be validated using real samples with well-established existing LFV in future research.

With respect to wastewater samples, our findings also corroborate the feasibility of using targeted amplicon-based metagenomics approaches for wastewater surveillance, as such samples comprise a collection of different strains, among which the dominant strain will define the consensus sequence of the sample and the detected LFV will represent the circulating strains present at lower frequencies. Only very limited recommendations regarding wastewater sequencing are available by the competent authorities. The EU has recommended the generation of one million reads per sample with a read length of minimum 100 bp which corresponds to a minimum coverage of 3333× using the Lander/Waterman equation (European Commission, 2021). Based on the results obtained in this study, a coverage of 3000 and 3500× allowed to detect LFV until an AF of 2 and 1.5% respectively with perfect sensitivity. Other studies that investigated LFV in wastewater have provided limited quality criteria regarding the coverage and AF. Furthermore, the quality criteria in those studies were not evaluated using a defined population (Izquierdo-Lara et al., 2020; Jahn et al., 2021). Izquierdo-Lara used a minimum depth coverage of 50× and minimum AF of 10% (Izquierdo-Lara et al., 2021), while Rios et al. (2021) adopted a minimum depth coverage of 100× without indicating an AF threshold. Based on the results in this study, these recommendations appear not sufficiently strict as a sequencing coverage of 100× and 250× at an AF of 20 and 10% respectively was required to observe all LFV. Obtaining high-quality sequencing reads for wastewater samples may, however, be challenging under real-world conditions. In contrast to diagnostic samples in which viral loads are typically high, ranging from 10^4 to 10^7 copies/mL (Pan et al., 2020), viral RNA loads in wastewater samples are often low, ranging from 10^{-1} to $10^{3.5}$ copies/mL (Saawarn and Hait, 2021). This renders it more challenging to sequence samples with a low viral load in addition to the RNA degradation that occurs in wastewater samples. Additionally, variants circulating at low frequencies in a community are expected to be present at a low AF in wastewater samples. Nevertheless, employing the most conservative thresholds from either Datasets 1 or 2, 90% of LFV present at an AF of 10, 5, 3, and 1% were still detected at a sequencing coverage of 100, 250, 500, and 2500×, respectively (**Tables 3**, **4**).

This study focused on the sensitivity of LFV detection and did not explore the false positive rates (i.e., specificity). Although our recommendations for AFs and coverages ensure high sensitivity, often an inverse relationship exists between sensitivity and specificity and we can therefore not exclude that false positives occur for AF and coverage combinations considered as providing qualitative results in this study. A false positive detection is, however, typically less problematic compared to a false negative result as the former can still be discovered in follow-up investigation in contrast to the latter. Additionally, false positive observations typically occur randomly over the genome (McCrone et al., 2016) and it is unlikely that all VOC-defining mutations would be simultaneously falsely detected, even at low AFs and coverages. The issue of low viral loads, low expected AF and potential false positives could be mitigated by sequencing samples in duplicate when necessary. Possible false positive results could be investigated using RT-qPCR or RT-ddPCR assays that target those specific positions.

TABLE 3 | Qualitative evaluation of Dataset 1 based on false negative proportions per condition until a targeted mutant AF of 20%.

Coverage→ AF%↓	100×	250×	500×	750×	1000×	1500×	2000×	2500×	3000×	3500×	4000×	4500×	5000×	10,000×
20.00%	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10.00%	5	0	0	0	0	0	0	0	0	0	0	0	0	0
9.50%	4	0	0	0	0	0	0	0	0	0	0	0	0	0
9.00%	7	1	0	0	0	0	0	0	0	0	0	0	0	0
8.50%	4	0	0	0	0	0	0	0	0	0	0	0	0	0
8.00%	9	2	0	0	0	0	0	0	0	0	0	0	0	0
7.50%	8	0	0	0	0	0	0	0	0	0	0	0	0	0
7.00%	10	1	0	0	0	0	0	0	0	0	0	0	0	0
6.50%	15	2	0	0	0	0	0	0	0	0	0	0	0	0
6.00%	15	1	0	0	0	0	0	0	0	0	0	0	0	0
5.50%	19	3	0	0	0	0	0	0	0	0	0	0	0	0
5.00%	22	3	0	0	0	0	0	0	0	0	0	0	0	0
4.50%	26	4	2	0	0	0	0	0	0	0	0	0	0	0
4.00%	31	6	1	0	0	0	0	0	0	0	0	0	0	0
3.50%	45	12	4	0	0	0	0	0	0	0	0	0	0	0
3.00%	47	18	4	1	0	0	0	0	0	0	0	0	0	0
2.50%	62	21	7	2	2	0	0	0	0	0	0	0	0	0
2.00%	70	32	14	7	3	0	0	0	0	0	0	0	0	0
1.50%	84	52	24	16	9	5	1	2	0	0	0	0	0	0
1.00%	96	77	54	35	28	15	8	6	6	3	2	2	2	0
0.50%	99	95	85	77	70	57	46	41	33	29	22	22	16	7
0.00%	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The percentage of FN is colored ranging from 0 (dark) to 1 (light) according to the gradient depicted in ***Figure 2A****.*

TABLE 4 | Qualitative evaluation of Dataset 2 based on false negative proportions per condition until a targeted mutant AF of 20%.

Coverage→ AF (%)↓	97×	201×	237×	482×	728×	969×	1454×	1937×	2413×	2904×	3383×	3872×	4358×	4851×	5855×	6834×	7801×	8790×	9792×
20.00%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10.00%	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.50%	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9.00%	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.50%	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8.00%	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.50%	8	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7.00%	9	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6.50%	18	35	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6.00%	28	38	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5.50%	31	47	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5.00%	35	56	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.50%	43	57	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.00%	51	59	6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.50%	58	63	18	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.00%	68	73	23	8	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2.50%	77	82	40	21	4	3	1	0	0	0	0	0	0	0	0	0	0	0	0
2.00%	81	84	55	33	11	6	4	1	0	0	0	0	0	0	0	0	0	0	0
1.50%	89	86	69	53	24	21	12	8	4	2	1	0	0	0	0	0	0	0	0
1.00%	92	86	91	80	57	52	34	22	8	15	6	7	6	4	0	0	0	0	0
0.50%	100	98	98	92	92	89	80	70	55	62	34	41	24	35	62	55	46	35	28
0.00%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

The percentage of FN is colored ranging from 0 (dark) to 1 (light) according to the gradient depicted in ***Figure 2B****.*

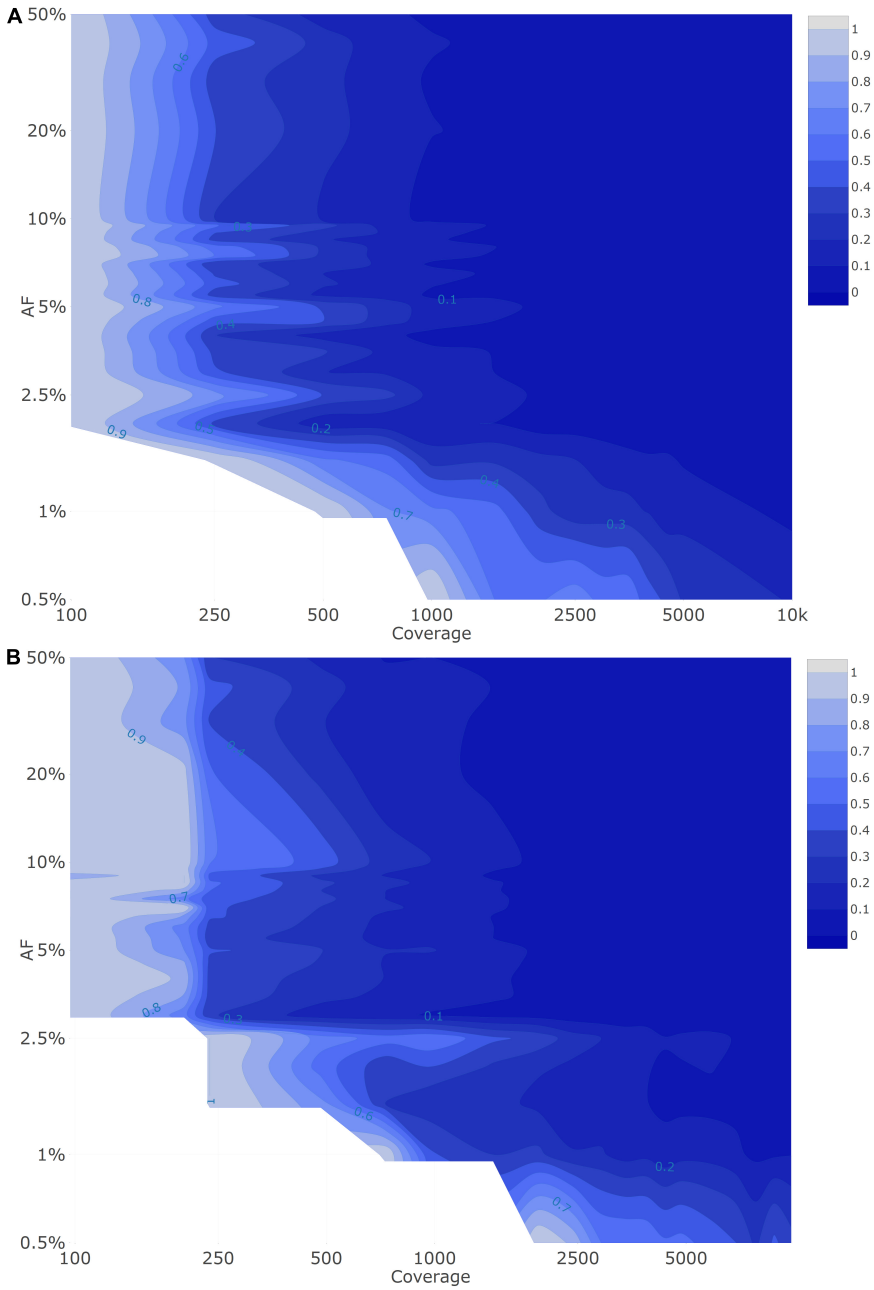

FIGURE 3 | Quantitative evaluation of Dataset 1 **(A)** and Dataset 2 **(B)** using the squared SD divided by the maximal squared SD per targeted AF. The figure is colored ranging from 0 (dark) to 1 (light) in intervals of 0.1 as extrapolated using a contour plot in the R package plotly (Sievert, 2020) (actual values are presented in **Supplementary File 3** for Dataset 1 and **Supplementary File 4** for Dataset 2). Both the *X*- and *Y*-axis follow a logarithmic scale. Conditions with a FN proportion higher than 75% were excluded and correspond to the white plane in the lower left corner.

In this study, the B.1.1.7 variant and a WT (i.e., non-VOC) background of the same time period and location were used as a proof-of-concept, but can be considered to also apply to other combinations (e.g., two VOCs), since additional VOCs in the sample material will translate into more VOC-defining mutations in the background genomic material that will be independently identified by the variant calling engine. In the presence of multiple VOCs, the VOCs can be identified by composing all possibly existing combinations of LFV as a conservative strategy, although multiple VOCs in one sample will also make the estimation of the relative abundance of each VOC more complicated. If multiple VOCs with partially overlapping defining mutations would be present in a wastewater sample, some mutations of interest would consequently be present at different AFs. Haplotype reconstruction methods could be used in such situations to delineate VOCs. However, most haplotype

reconstruction programs perform poorly under higher levels of diversity, and haplotype populations with rare haplotypes are often not recovered (Eliseev et al., 2020). Although haplotype reconstruction has been described for short reads, Nanopore sequencing might offer a substantial advantage for such cases due to its longer reads, despite their higher error rate, to perform haplotype estimation to delineate actual VOCs.

CONCLUSION

There exists a pressing need for recommendations for detecting LFV for both diagnostic samples and wastewater surveillance. Further investigation will be required to investigate the specificity and possibility to detect VOCs instead of just mutations, including for other existing and employed methodologies such as probe-based capture, other amplicon-based methods, and Nanopore sequencing. Nevertheless, using *in silico* modified data derived from WGS of real diagnostic samples, this study demonstrates the feasibility of a targeted metagenomics approach for highly sensitive LFV detection with acceptable relative abundance estimations using a tiled-amplicon enrichment based on the Illumina technology. This approach enables the detection of mutations associated with specific VOCs. Our approach could be used to evaluate the potential occurrence of co-infections with other SARS-CoV-2 variants with different strains in diagnostic samples. It can also be employed to detect multiple strains for wastewater surveillance, although several additional challenges exist for wastewater samples such as low viral load and potential RNA degradation. Since in this study, high-quality data from diagnostic samples was used and modified *in silico* to construct datasets to provide guidelines for sequencing wastewater and diagnostic samples with co-infections, future work will need to consider data coming from samples that are closer to real data from actual diagnostic and wastewater surveillance. In light of the pandemic urgency, and the multiple SARS-CoV-2 wastewater surveillance initiatives that are being established and also being integrated into overarching coordination and preparedness initiatives such as the recently announced European Health Emergency Preparedness and Response Authority (European Commission, 2021), we hope that our results will help establishing guidance and recommendations for wastewater surveillance and other relevant applications.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

NR, KV, and XS: conceptualization. NR: project administration. LV: data curation, investigation, and visualization. LV, TD, WC, SD, NR, and KV: methodology. LV, TD, and WC: software and formal analysis. LV and TD: validation. LV, TD, NR, and KV: writing – original draft preparation. NR and PH: funding acquisition. NR and KV: supervision. All authors: writing – review and editing.

FUNDING

This study was financed by Sciensano through COVID-19 special funding.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.747458/full#supplementary-material

REFERENCES

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J. W., et al. (2020). First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Sci. Total Environ.* 728, 138764. doi: 10.1016/j.scitotenv.2020.138764

Azuma, K., Yanagi, U., Kagi, N., Kim, H., Ogata, M., and Hayashi, M. (2020). Environmental factors involved in SARS-CoV-2 transmission: effect and role of indoor environmental quality in the strategy for COVID-19 infection control. *Environ. Health Preven. Med.* 25:66. doi: 10.1186/s12199-020-00904-2

Bal, A., Destras, G., Gaymard, A., Stefic, K., Marlet, J., Eymieux, S., et al. (2021). Two-step strategy for the identification of SARS-CoV-2 variant of concern 202012/01 and other variants with spike deletion H69–V70, France, August to December 2020. *Eurosurveillance* 26, 1–5. doi: 10.2807/1560-7917.ES.2021.26.3.2100008

Bar-Or, I., Weil, M., Indenbaum, V., Bucris, E., Bar-Ilan, D., Elul, M., et al. (2021). Detection of SARS-CoV-2 variants by genomic analysis of wastewater samples in Israel. *Sci. Total Environ.* 789:148002. doi: 10.1016/j.scitotenv.2021.148002

Bayle, C., Cantin, D., Vidal, J. S., Sourdeau, E., Slama, L., Dumesges, N., et al. (2021). Asymptomatic SARS COV-2 carriers among nursing home staff: A source of contamination for residents? *Infect. Dis. Now* 51, 197–200. doi: 10.1016/j.idnow.2020.11.008

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Boni, M. F. (2008). Vaccination and antigenic drift in influenza. *Vaccine* 26, C8–C14. doi: 10.1016/j.vaccine.2008.04.011

Bushnell, B., and BBMap (2021). Available online at: https://sourceforge.net/projects/bbmap/ [Accessed March 29, 2021].

Centers for Disease Control and Prevention [CDC]. (2021). *SARS-CoV-2 Variants*. Available online at: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html [accessed March 4, 2021]

Charre, C., Ginevra, C., Sabatier, M., Regue, H., Destras, G., Brun, S., et al. (2020). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol.* 6:75. doi: 10.1093/ve/veaa075

Contreras, S., Dehning, J., Loidolt, M., Zierenberg, J., Spitzner, F. P., Urrea-Quintero, J. H., et al. (2021). The challenges of containing SARS-CoV-2 via test-trace-and-isolate. *Nat. Commun.* 12:378. doi: 10.1038/s41467-020-20699-8

Crits-Christoph, A., Kantor, R. S., Olm, M. R., Whitney, O. N., Al-Shayeb, B., Lou, Y. C., et al. (2021). Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 Variants. *mBio* 12:20. doi: 10.1128/mBio.02703-20

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10:8. doi: 10.1093/gigascience/giab008

Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J., et al. (2021). Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. *medRxiv* 2020:20248822. doi: 10.1101/2020.12.24.20248822

Duchene, S., Featherstone, L., HaritopoulouSinanidou, M., Rambaut, A., Lemey, P., and Baele, G. (2020). Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol.* 6, 1–8. doi: 10.1093/ve/veaa061

ECDC. (2021). *Sequencing of SARS-CoV-2: first update (18 January 2021)*.Solna Municipality: ECDC.

Eliseev, A., Gibson, K. M., Avdeyev, P., Novik, D., Bendall, M. L., PerezLosada, M., et al. (2020). Evaluation of haplotype callers for next-generation sequencing of viruses. *Infect. Genet. Evol.* 82:104277. doi: 10.1016/j.meegid.2020.104277

European Commission. (2021). *Commission Recommendation of 17.3.2021 on a common approach to establish a systematic surveillance of SARS-CoV-2 and its variants in wastewaters in the EU*. Brussels: European Commission.

Ewing, A. D., Houlahan, K. E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T. N., et al. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* 12, 623–630. doi: 10.1038/nmeth.3407

Firestone, M. J., Lorentz, A. J., Meyer, S., Wang, X., Como-Sabetti, K., Vetter, S., et al. (2021). First Identified Cases of SARS-CoV-2 Variant P.1 in the United States — Minnesota, January 2021. *MMWR. Morb. Mortal. Week. Rep.* 70, 346–347. doi: 10.15585/mmwr.mm7010e1

Gómez, C. E., Perdiguero, B., and Esteban, M. (2021). Emerging SARS-CoV-2 Variants and Impact in Global Vaccination Programs against SARS-CoV-2/COVID-19. *Vaccines* 9:243. doi: 10.3390/vaccines9030243

GOV.UK - Scientific Advisory Group for Emergencies. (2021). *NERVTAG: Update note on B.1.1.7 severity*. Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/961042/S1095_NERVTAG_update_note_on_B.1.1.7_severity_20210211.pdf [accessed March 4, 2021].

Greaney, A. J., Starr, T. N., Gilchuk, P., Zost, S. J., Binshtein, E., Loes, A. N., et al. (2021). Complete Mapping of Mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *Cell Host Microbe* 29, 44–57. doi: 10.1016/j.chom.2020.11.007

Harris, C. R., Millman, K. J., vanderWalt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Hartley, P. D., Tillett, R. L., AuCoin, D. P., Sevinsky, J. R., Xu, Y., Gorzalski, A., et al. (2021). Genomic surveillance of Nevada patients revealed prevalence of unique SARS-CoV-2 variants bearing mutations in the RdRp gene. *J. Genet. Genomics* 2021:4. doi: 10.1016/j.jgg.2021.01.004

Hoffmann, M., Arora, P., GroB, R., Seidel, A., Hornich, B. F., Hahn, A. S., et al. (2021). SARS-CoV-2 variants B.1.351 and P.1 escape from neutralizing antibodies. *Cell* 2021:36. doi: 10.1016/j.cell.2021.03.036

Isakov, O., Borderia, A. V., Golan, D., Hamenahem, A., Celniker, G., Yoffe, L., et al. (2015). Deep sequencing analysis of viral infection and evolution allows rapid and detailed characterization of viral mutant spectrum. *Bioinformatics* 31, 2141–2150. doi: 10.1093/bioinformatics/btv101

Izquierdo-Lara, R., Elsinga, G., Heijnen, L., Munnink, B. B. O., Schapendonk, C. M. E., Nieuwenhuijse, D., et al. (2021). Monitoring SARS-CoV-2 Circulation and Diversity through Community Wastewater Sequencing, the Netherlands and Belgium. *Emerg. Infect. Dis.* 27, 1405–1415. doi: 10.3201/eid2705.204410

Izquierdo-Lara, R., Elsinga, G., Heijnen, L., Oude Munnink, B. B., Schapendonk, C. M. E., Nieuwenhuijse, D., et al. (2020). Monitoring SARS-CoV-2 circulation and diversity through community wastewater sequencing. *medRxiv* 2020:20198838. doi: 10.1101/2020.09.21.20198838

Jahn, K., Dreifuss, D., Topolsky, I., Kull, A., Ganesanandamoorthy, P., Fernandez-Cassi, X., et al. (2021). Detection of SARS-CoV-2 variants in Switzerland by genomic analysis of wastewater samples. *medRxiv* 2021:21249379. doi: 10.1101/2021.01.08.21249379

Karim, F., Moosa, M. Y. S., Gosnell, B. I., Cele, S., Giandhari, J., Pillay, S., et al. (2021). Persistent SARS-CoV-2 infection and intra-host evolution in association with advanced HIV infection. *medRxiv* 2021:21258228. doi: 10.1101/2021.06.03.21258228

Kim, D., Lee, J. Y., Yang, J. S., Kim, J. W., Kim, V. N., and Chang, H. (2020). The Architecture of SARS-CoV-2 Transcriptome. *Cell* 181, 914–921. doi: 10.1016/j.cell.2020.04.011

Kundu, S., Lockwood, J., Depledge, D. P., Chaudhry, Y., Aston, A., Rao, K., et al. (2013). Next-Generation whole genome sequencing identifies the direction of norovirus transmission in linked patients. *Clin. Infect. Dis.* 57, 407–414. doi: 10.1093/cid/cit287

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leclerc, Q. J., Fuller, N. M., Knight, L. E., Funk, S., and Knight, G. M. (2020). What settings have been linked to SARS-CoV-2 transmission clusters? *Wellcome Open Res.* 5:83. doi: 10.12688/wellcomeopenres.15889.2

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lin, J., Tang, C., Wei, H., Du, B., Chen, C., Wang, M., et al. (2021). Genomic monitoring of SARS-CoV-2 uncovers an Nsp1 deletion variant that modulates type I interferon response. *Cell Host Microbe.* 29, 489–502. doi: 10.1016/j.chom.2021.01.015

Lindenbaum, P. (2015). JVarkit: java-based utilities for Bioinformatics. *Comp. Sci.* 2015:30. doi: 10.6084/M9.FIGSHARE.1425030.V1

Lythgoe, K. A., Hall, M., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., et al. (2021). SARS-CoV-2 within-host diversity and transmission. *Science* 372:eabg0821. doi: 10.1126/science.abg0821

Macalalad, A. R., Zody, M. C., Charlebois, P., Lennon, N. J., Newman, R. M., Malboeuf, C. M., et al. (2012). Highly Sensitive and Specific Detection of Rare Variants in Mixed Viral Populations from Massively Parallel Sequence Data. *PLoS Comput. Biol.* 8:e1002417. doi: 10.1371/journal.pcbi.1002417

McCrone, J. T., Lauring, A. S., and Lauring, S. (2016). Measurements of intrahost viral diversity are extremely sensitive to systematic errors in variant calling. *J. Virol.* 90, 6884–6895. doi: 10.1128/JVI.00667-16

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., and Brouwer, A. (2020). Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 prevalence in the early stage of the epidemic in the netherlands. *Environ. Sci. Technol. Lett.* 7, 511–516. doi: 10.1021/acs.estlett.0c00357

Mishra, S., Mindermann, S., Sharma, M., Whittaker, C., Mellan, T. A., Wilton, T., et al. (2021). Changing composition of SARS-CoV-2 lineages and rise of Delta variant in England. *EClin. Med* 39:101064. doi: 10.1016/j.eclinm.2021.101064

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., et al. (2021). Sustainable data analysis with Snakemake. *F1000 Res.* 10:33. doi: 10.12688/f1000research.29032.1

Nemudryi, A., Nemudraia, A., Wiegand, T., Surya, K., Buyukyoruk, M., Cicha, C., et al. (2020). Temporal Detection and Phylogenetic Assessment of SARS-CoV-2 in Municipal Wastewater. *Cell Rep. Med.* 1, 100098. doi: 10.1016/j.xcrm.2020.100098

Pan, Y., Zhang, D., Yang, P., Poon, L. L. M., and Wang, Q. (2020). Viral load of SARS-CoV-2 in clinical samples. *Lancet Infect. Dis.* 20, 411–412. doi: 10.1016/S1473-3099(20)30113-4

Panchal, D., Prakash, O., Bobde, P., and Pal, S. (2021). SARS-CoV-2: sewage surveillance as an early warning system and challenges in developing countries. *Environ. Sci. Poll. Res.* 2021:8. doi: 10.1007/s11356-021-13170-8

Public Health England. (2021). *Variants of concern or under investigation*. Available online at: https://www.gov.uk/government/publications/covid-19-variants-genomically-confirmed-case-numbers/variants-distribution-of-cases-data [accessed March 4, 2021]

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Rambaut, A., Loman, N., Pybus, O., Barcly, W., Barrett, J., Carabelli, A., et al. (2021). *Preliminary genomic characterisation of an emergent SARS-CoV-2*

lineage in the UK defined by a novel set of spike mutations. Available online at: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 [accessed March 4, 2021].

Rios, G., Lacoux, C., Leclercq, V., Diamant, A., Lebrigand, K., Lazuka, A., et al. (2021). Monitoring SARS-CoV-2 variants alterations in Nice neighborhoods by wastewater nanopore sequencing. *medRxiv* 2021:21257475. doi: 10.1101/2021.07.09.21257475

Rogers, M. B., Song, T., Sebra, R., Greenbaum, B. D., Hamelin, M.-E., Fitch, A., et al. (2015). Intrahost dynamics of antiviral resistance in influenza A virus reflect complex patterns of segment linkage, reassortment, and natural selection. *mBio* 6:14. doi: 10.1128/mBio.02464-14

Saawarn, B., and Hait, S. (2021). Occurrence, fate and removal of SARS-CoV-2 in wastewater: Current knowledge and future perspectives. *J. Environ. Chem. Eng.* 9:104870. doi: 10.1016/j.jece.2020.104870

SAGE-EMG, SPI-B, Tranmission Group. (2020). *Mitigations to Reduce Transmission of the new variant SARS-CoV-2 virus*. Available online at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/948607/s0995-mitigations-to-reduce-transmission-of-the-new-variant.pdf [accessed March 4, 2021].

Shao, W., Li, X., Goraya, M. U., Wang, S., and Chen, J.-L. L. (2017). Evolution of Influenza A Virus by Mutation and Re-Assortment. *Int. J. Mol. Sci.* 18:1650. doi: 10.3390/ijms18081650

Sharif, S., Ikram, A., Khurshid, A., Salman, M., Mehmood, N., Arshad, Y., et al. (2021). Detection of SARs-CoV-2 in wastewater using the existing environmental surveillance network: A potential supplementary system for monitoring COVID-19 transmission. *PLoS One* 16:e0249568. doi: 10.1371/journal.pone.0249568

Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* 22:30494. doi: 10.2807/1560-7917.ES.2017.22.13.30494

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Boca Raton, FL: CRC Press.

Simon, B., Pichon, M., Valette, M., Burfin, G., Richard, M., Lina, B., et al. (2019). Whole Genome Sequencing of A(H3N2) Influenza Viruses Reveals Variants Associated with Severity during the 2016–2017 Season. *Viruses* 11:108. doi: 10.3390/v11020108

Sinclair, R. G., Choi, C. Y., Riley, M. R., and Gerba, C. P. (2008). Pathogen Surveillance Through Monitoring of Sewer Systems. *Adv. Appl. Microbiol.* 65, 249–269. doi: 10.1016/S0065-2164(08)00609-6

Siqueira, J. D., Goes, L. R., Alves, B. M., de Carvalho, P. S., Cicala, C., Arthos, J., et al. (2021). SARS-CoV-2 genomic analyses in cancer patients reveal elevated intrahost genetic diversity. *Virus Evol.* 7:veab013. doi: 10.1093/ve/veab013

Thompson, J. R., Nancharaiah, Y. V., Gu, X., Lee, W. L., Rajal, V. B., Haines, M. B., et al. (2020). Making waves: Wastewater surveillance of SARS-CoV-2 for population-based health management. *Water Res.* 184:116181. doi: 10.1016/j.watres.2020.116181

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* 83:104351. doi: 10.1016/j.meegid.2020.104351

WHO (2021). *Tracking SARS-CoV-2 variants*. Available online at: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/ [accessed June 23, 2021]

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., et al. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 40, 11189–11201. doi: 10.1093/nar/gks918

Wu, F., Zhang, J., Xiao, A., Gu, X., Lee, W. L., Armas, F., et al. (2020a). SARS-CoV-2 Titers in wastewater are higher than expected from clinically confirmed cases. *mSystems* 5:61420. doi: 10.1128/mSystems.00614-20

Wu, Y., Guo, C., Tang, L., Hong, Z., Zhou, J., Dong, X., et al. (2020b). Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *Lancet. Gastroenterol. Hepatol.* 5, 434–435. doi: 10.1016/S2468-1253(20)30083-2

Xagoraraki, I., and O'Brien, E. (2020). "Wastewater-Based Epidemiology for Early Detection of Viral Outbreaks," in *Women in Engineering and Science*, ed. D. J. O'Bannon (New York, NY: Springer International Publishing), 75–97. doi: 10.1007/978-3-030-17819-2_5

Zhang, W., Du, R.-H., Li, B., Zheng, X.-S., Yang, X.-L., Hu, B., et al. (2020). Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes. *Emerg. Microbes Infect.* 9, 386–389. doi: 10.1080/22221751.2020.1729071

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.