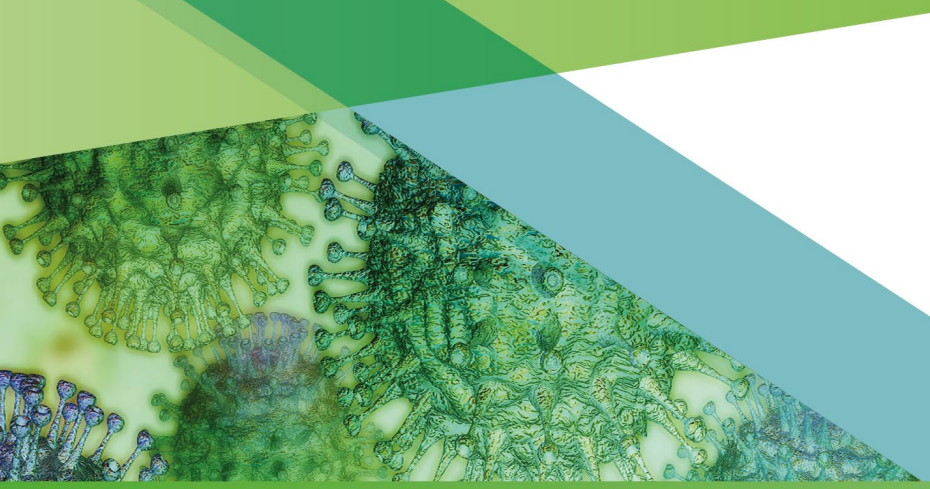

TECHNICAL GUIDANCE

Sequencing of SARS-CoV-2: first update

18 January 2021

Introduction

In January 2020, a previously unknown coronavirus strain was identified as the cause of a respiratory infection and death in humans [1]. The first viral genome was sequenced using high throughput sequencing (HTS) from a sample collected in Wuhan, China. This virus, belonging to the viral species *Severe acute respiratory syndrome–related coronavirus*, has been subsequently named SARS-CoV-2 and the associated disease coronavirus disease 2019 (COVID-19) [2].

Sequencing of (partial) genes and whole genomes (WGS) has been proven as a powerful method to investigate viral pathogen genomes, understand outbreak transmission dynamics and spill-over events and screen for mutations that potentially have an impact on transmissibility, pathogenicity, and/or countermeasures (e.g. diagnostics, antiviral drugs and vaccines). The results are key to informing outbreak control decisions in public health.

Scope

A standardised pipeline to characterise, name and report SARS-CoV-2 sequences has not been established yet, but many countries of the WHO European Region have been sequencing SARS-CoV-2 variants since the beginning of the pandemic and reporting the sequences to the Global Initiative on Sharing All Influenza Data (GISAID) or other publicly accessible databases [3]. Combining information of virus characteristics with clinical and epidemiological data is important. Genetic characterisation of SARS-CoV-2 is used to monitor viral evolution and to timely identify potential markers of increased transmissibility, severity of disease or altered antigenicity. Emerging hypotheses will need to be further investigated in ex vivo, in vitro or animal models. Sequence data will become increasingly important as SARS-CoV-2 vaccines and antivirals become available, in order to monitor the match of the circulating variants with the vaccine and the possible emergence of antiviral resistance.

This technical guidance aims to provide guidelines to laboratories, microbiology experts and relevant stakeholders in making decisions on establishing sequencing capacities and capabilities, in making decisions on which technologies to use and/or in deciding on the role of sequencing for SARS-CoV-2 diagnostics, research, outbreak investigations and surveillance. It addresses the most used sequencing technologies and their applications and proposes a central standardisation process to analyse and report the findings of SARS-CoV-2 genetic characterisations.

Objectives for sequencing SARS-CoV-2

The COVID-19 pandemic is the first pandemic in which WGS capacity has been available to the public health sector from the very beginning. The first sequences were published in January 2020 and the sequence

Suggested citation: European Centre for Disease Prevention and Control. Sequencing of SARS-CoV-2: first update. 18 January 2021. ECDC: Stockholm; 2021.

information was immediately used to set up viral RNA detection systems by nucleic acid amplification techniques (NAAT).

The current objectives of SARS-CoV-2 sequencing are:

Main objectives

- Early detection and characterisation of emerging variant viruses to define if they are of particular concern;
- Assessing the impact of genetic and antigenic variant viruses for the pandemic and monitoring them over time to guide public health action.

Specific objectives

- Investigating virus transmission dynamics and introductions of novel genetic variants;
- Modelling the antigenic properties of the virus to assess the risk of vaccine escape;
- Selecting viruses for vaccine composition;
- Assessing the impact of mutations on the performance of molecular diagnostic, antigen characterisation and serological methods;
- Investigating the relationship between clades/lineages and epidemiological data such as transmissibility and disease severity or risk groups;
- Understanding the impact of response measures on the virus population;
- Assessing relatedness of viral strains within epidemiological clusters and supporting contact tracing and other public health interventions;
- Assessing and confirming reinfections;
- Monitoring emerging lineages within wild/domestic/farmed animal populations that may impact human health;
- Prompting further basic research investigation to confirm the relevance of observed mutations in the pathogenesis of the disease (e.g. infectivity, receptors binding);
- Assessing the impact of mutations on the performance of antiviral drugs;
- Assessing the potential incidence of vaccine-derived virus infections and transmissions should live SARS-CoV-2 vaccines become available.

One major component of sequence-based surveillance of any pathogen is applying meaningful nomenclatures to the sequence data, based on the genetic relatedness of the sequences. This streamlines communication between different actors in the molecular epidemiology field and enables simplified tabulation of the genomic data for integration with standard epidemiological analysis. Several nomenclatures have been implemented for SARS-CoV-2:

- GISAID nomenclature, using the term hCoV-19 for the virus (www.gisaid.org)
- Nextstrain nomenclature (https://nextstrain.org/ncov)
- Lineage nomenclature by Andrew Rambaut et al. (https://cov-lineages.org)

While Nextstrain and GISAID clade nomenclatures aim at providing a broad-brush categorisation of globally circulating diversity, the lineages (https://cov-lineages.org) are intended to correspond to outbreaks in specific geographical regions.

Although WGS was used to detect and identify the novel virus when PCR tests were not yet available, sequencing is currently not widely used for the diagnosis of SARS-CoV-2 infections. The United States' Food and Drug Administration has provided Emergency Use Authorisation to one diagnostic detection system using next generation sequencing [4]. In addition to the public health surveillance objectives of sequencing listed above, sequencing is used widely in research studies, such as prospective genomic studies [5, 6].

Application of sequencing for SARS-CoV-2

Sampling strategy and sample selection

Sample selection will depend on the selected objective and available resources. For surveillance purposes, representative strains of virus from different geographic locations and time points, as well as from patients of varied demographics and across the disease severity spectrum should be selected for sequencing. For targeted monitoring (e.g. of outbreaks), in order to more effectively monitor virus evolution and changes in the virus genome, sampling strategies should include the selection of SARS-CoV-2 vaccine escape variants, viruses causing reinfections, variants emerging in animal populations or variants with increased transmission, especially if they are not explained by other epidemiological factors, for sequencing. For resource-limited settings, an event-/risk-based approach may reduce the need and cost for sequencing. When variants of concern are reported from specific geographical areas, increased focus on sequencing cases with an epidemiological link to these areas, or other evidence suggesting exposure to such a variant, should be considered. If other methods for identifying such variants are available, they should be considered as a complement to sequencing.

To ensure that a representative picture of the distribution of viruses is obtained, it is crucial to continue with surveillance of SARS-CoV-2 viruses in community and hospital settings and use random sampling of COVID-19 cases in those settings in order to assess the prevalence of variant viruses in the population and also to detect variant viruses early.

In addition to representative sampling of the population, targeted sampling should be applied to detect variants of concern. For such targeted sampling, the below criteria of prioritisation for sequencing can be used.

World Health Organization interim guidance on SARS-CoV-2 genomic sequencing for public health goals recommends prioritisation of the following groups for sequencing [7]:

- from individuals vaccinated for SARS-CoV-2 but who later become infected with SARS-CoV-2 despite exhibiting an appropriate immune response to the vaccine;
- in risk settings, such as where there is close human–animal interaction with a large number of animals that are susceptible to SARS-CoV-2 infection, or where there are immunocompromised patients with prolonged shedding, especially when receiving antibody therapy against SARS-CoV-2;
- when there is an unexpected increase or change in SARS-CoV-2 transmissibility and/or virulence;
- when there is suspicion of a change in the performance of diagnostic (antibody, antigen, molecular assays) methods or therapies; and
- during cluster investigations when sequencing can support understanding of transmission events and/or evaluate the efficacy of infection control procedures.

In addition to the above priority list, virus detections related to travel should be monitored for emergence of variant viruses.

Sample size for random and representative sampling

The number of samples to be sequenced should depend on the level of detection and monitoring capabilities that are aimed for and what resources are available. In the Annex, tables A1-A9 describe the required number of samples to be sequenced using a representative or random sampling strategy, depending on the desired capabilities. They show how many sequences are required for detection capability and for being able to quantify the fraction of specific variants within the entire population of viruses, with different levels of precision, for different desired lower variant frequencies. The tables are available for a range of total number of COVID-19 cases, from 500 and up.

Example 1: A country with approximately 50 000 weekly COVID-19 cases wants to monitor the level of circulation of variants at the national level on a weekly basis, above a proportion of 1% variant viruses out of the total COVID-19 cases with a medium precision. The table A3 for 50 000 cases, in the column for medium precision and the row of 1% variant proportion indicates 5 146 samples to be sequenced weekly to reach the desired performance.

Example 2: A country with approximately 12 500 weekly cases wants to detect introductions of new variants above 0.1% in the community on a monthly basis without any assessment of the level of the variant. These 12 500 weekly cases correspond to the 50 000 cases table, as the assessment is going to be carried out on monthly data. The table A3 for 50 000 cases in the column for detection and the row for 0.1% variant proportion indicates 2 748 samples to be sequenced, in this case, on a monthly basis.

Sequence methodologies

There are several methods available for sequencing SARS-CoV-2 from clinical samples [8]. The main method-related parameters that are of importance for the various applications include the library construction approach chosen (untargeted, capture-based, amplicon-based), the read length generated, error rate and error profile, depth of sequencing, and uniformity of coverage across the genome or partially sequenced genome.

For most genomic surveillance objectives, a consensus sequence of the complete or almost complete genome is sufficient. This can be achieved in a cost-effective way by using multiplex amplicon assays, for example the open-source ARTIC protocol (https://artic.network/ncov-2019), commercial kits (available for both Illumina and Ion Torrent platforms), or in-house protocols. For confirmation of direct transmission and/or reinfection, higher sequencing coverage is recommended for the determination of minority variants that can contribute significantly to the evidence for direct transmission or reinfection.

The short amplicon methods do not permit the accurate detection of genome changes that are of similar or larger size than the amplicon length and determine/reconstruct haplotypes. For these reasons, longer amplicons, capture-based, or untargeted libraries combined with long-read sequencing technologies are recommended for these applications. If Sanger sequencing is the preferred method, sequencing of the whole length of the S gene is recommended.

For detection of unknown pathogens using HTS, untargeted sequencing is required. This approach can also be used when SARS-CoV-2 infection is suspected, but rRT-PCR using different primer-probe sets and gene targets have produced negative results. A more cost-effective but less general approach in this situation is to use β-CoV-specific RT-PCR primers and perform Sanger sequencing or HTS of any resulting PCR amplification products.

Table 1. SARS-CoV-2 genome sequencing applications and recommended technologies

Application	Recommended sequencing platforms	Recommended library construction approaches	Recommended read length	Recommended minimum local coverage (approximate)
Transmission patterns, clade/lineage assignment, confirmation of reinfection, phenotypically relevant mutations, data reporting	MiSeq/NextSeq/iSeq/NovaSeq (Illumina), Ion Torrent (Thermo Fisher), MinION (Oxford Nanopore), Sequel System (PacBio)	Amplicon-based (ARTIC, commercial, in-house)	>100 bp	>10x over >95% of genome
Confirmation of reinfection and/or direct transmission (in cases where minority variants are required)	MiSeq/NextSeq/iSeq/NovaSeq (Illumina), Ion S5 series/Genexus (Thermo Fisher)	Amplicon-based (ARTIC, commercial, in-house)	>100 bp	>500x over >95% of genome
In-depth genome analysis (large indels, recombination, rearrangements, quasi-species haplotypes)	MinION (Oxford Nanopore), Sequel System (PacBio)	Amplicon-based (>1000 bp fragments), capture-based, untargeted	>1000 bp	>500x over >95% of genome
Detection of unknown pathogens or highly divergent strains	MiSeq/NextSeq/iSeq/NovaSeq (Illumina), Ion S5 series/Genexus (Thermo Fisher), MinION (Oxford Nanopore)	Untargeted RNA sequencing, β-CoV-specific RT-PCR	>100bp	>5Gbp data per sample

Data-sharing and reporting

Consensus sequences should be shared in the GISAID EpiCov database (www.gisaid.org) to enable global phylogenetic analysis. Raw data can be deposited in the COVID-19 data portal (www.covid19dataportal.org) to make it available for the global community. Both GISAID and ENA/SRA accession numbers can be reported to The European Surveillance System (TESSy). If it is found that mutations in the virus caused a false negative rRT-PCR result, this should be reported promptly to ECDC and the WHO Regional Office for Europe. For recommended metadata for GISAID and TESSy reporting, data submitters should refer to the respective metadata sets.

Data analysis

The ARTIC protocol and the commercial kits for Illumina and Thermo Fischer platforms come with recommended bioinformatics analysis pipelines, which are suitable for majority variant detection for single nucleotide variants (SNVs) and small insertions and deletions of nucleotides as well as generation of consensus sequences for subsequent upload to sequence databases. The functionality of the pipeline should be verified by the laboratory before data are reported. For in-house pipelines, it is important that the bioinformatics pipeline used is fully validated for fitness for purpose by the laboratory. In general, methods based on reference mapping are most suitable for routine analysis.

For other types of analyses, such as minority variant determination or structural genomic variants detection, specific competence in viral genomics is recommended, as these analyses are technically challenging and can often not be fully automated.

Suggested applications at the local and national level

This list includes references to articles and reports that describe methodology for the respective applications:

- Following geographical and temporal trends [3] of clades (GISAID, Nextstrain), lineages (https://cov-lineages.org) and individual mutations to assess the impact of interventions, including vaccinations;
- Tracking community transmission [5] and closed-setting transmission [9];

- Differentiating between reinfection and prolonged carriage by comparing the genomic sequence from different COVID-19 episodes for the same patient [10];
- Assessing the frequency of mutations that can affect the sensitivity of nucleic acid-based detection assays [11, 12] (https://primerscan.ecdc.europa.eu);
- Virological research, e.g. linking nucleotide changes to phenotypic changes;
- Assessing spill-over events from wild or domestic animals to humans and vice-versa.

In order to ensure the sustainable implementation and uptake of HTS methodologies, the associated weaknesses and threats need to be equally considered alongside the strengths and opportunities that such technologies would bring. Countries that have not yet employed HTS technologies should take into account evidence from existing literature on considerations regarding the implementation of such technologies; see suggestions for further reading material below.

Competence in HTS in WHO Reference Laboratories

Table 2. HTS capacities available across WHO Reference Laboratories

WHO Reference Laboratory	Available technology	Contact
National Institute for Infectious Diseases L. Spallanzani, Italy	Illumina MiSeq Ion Torrent GenStudio S5 Prime Oxford Nanopore MinION Acquisition ongoing: Illumina Nexseq550 Ion Torrent Genexus	Antonino Di Caro Antonino.dicaro@inmi.it
Federal Budgetary Research Institution – State Research Center of Virology and Biotechnology VECTOR, Federal Service for Surveillance on Consumer Rights Protection and Human Well-being, Russia	Illumina MiSeq Illumina NexSeq Oxford Nanopore MinION	Sergey Bodnev bodnev@vector.nsc.ru
Institut Pasteur, Molecular genetics of RNA viruses, National Reference Center for Respiratory viruses, France	Illumina MiSeq Illumina NexSeq Oxford Nanopore MinION	Sylvie van der Werf Sylvie.van-der-werf@pasteur.fr
RIVM, the Netherlands	Illumina MiSeq Illumina NextSeq Oxford Nanopore MinION Oxford Nanopore GridION	Harry Vennema harry.vennema@rivm.nl Chantal Reusken chantal.reusken@rivm.nl
ErasmusMC, the Netherlands	Oxford Nanopore Gridion Oxford Nanopore MinION Illumina MiSeq Illumina NovaSeq	Marion Koopmans m.koopmans@erasmusmc.nl
Institute of Virology, Charite - Universitätsmedizin Berlin, Germany	Illumina MiSeq Illumina NexSeq Illumina NovaSeq Oxford Nanopore MinION Oxford Nanopore GridION	Christian Drosten Victor Corman victor.corman@charite.de
Robert Koch Institute (RKI), Germany	Illumina MiSeq Oxford Nanopore MinION	Andreas Nitsche nitschea@rki.de
PHE Colindale, England, the UK	Illumina HiSeq 2500 Illumina NextSeq 550/500 Illumina MiSeq	Maria Zambon maria.zambon@phe.gov.uk

WHO Reference Laboratory	Available technology	Contact
	Acquisition ongoing: Illumina NextSeq 1000 Oxford Nanopore GridION	
Geneva University Hospitals (HUG), Switzerland	Illumina MiSeq Illumina HiSeq (4000)	Isabella Eckerle Isabella.Eckerle@hcuge.ch

Further reading material

Infectious Disease Next Generation Sequencing Based Diagnostic Devices: Microbial Identification and Detection of Antimicrobial Resistance and Virulence Markers

Comprehensive workflow for detecting coronavirus using Illumina benchtop systems, Illumina

Comparison of Ion Torrent and Illumina in viral genome sequencing

Technical guide on next-generation sequencing technologies for the detection of mutations associated with drug resistance in Mycobacterium tuberculosis complex

ECDC contributing experts (in alphabetical order)

Erik Alm, Eeva Broberg, Angeliki Melidou.

Consulted experts

We would like to acknowledge WHO Regional Office for Europe and WHO HQ colleagues for their contributions: Marco Marklewitz, Soudeh Ehsani, Joanna Zwetyenga, Karin von Eije, Lisa Carter, Lorenzo Subissi, Sebastien Cognat, Roger Evans, Céline Barnadas.

Reviewers

We would also like to thank the WHO referral laboratories and the ECOVID19 laboratory network for the review and useful comments: Antonino Di Caro, Chantal Reusken, Harry Vennema, Kim Benschop, Sylvie van der Werf, Sergey Bodnev, Marion Koopmans, Oskar Karlsson Lindsjö, Mia Brytting, Victor Corman.

References

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020 02 20;382(8):727-33.
2. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020 04;5(4):536-44.
3. Alm E, Broberg EK, Connor T, Hodcroft EB, Komissarov AB, Maurer-Stroh S, et al. Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Euro Surveill. 2020;25(32):pii=2001410. Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.32.2001410
4. U.S. Food and Drug Administration (FDA). Emergency Use Authorization. Available from: https://www.fda.gov/emergency-preparedness-and-response/mcm-legal-regulatory-and-policy-framework/emergency-use-authorization
5. Oude Munnink BB, Nieuwenhuijse DF, Stein M, O'Toole Á, Haverkate M, Mollers M, et al. Rapid SARS-CoV-2 whole-genome sequencing and analysis for informed public health decision-making in the Netherlands. Nat Med. 2020 09;26(9):1405-10.
6. Inc. I. Press Release: Illumina Receives First FDA Emergency Use Authorization for a Sequencing-Based COVID-19 Diagnostic Test. Available from: https://www.illumina.com/company/news-center/press-releases/2020/8cd141fb-68d0-4144-8922-45693ac3f453.html
7. WHO. SARS-CoV-2 genomic sequencing for public health goals: Interim guidance, 8 January 2021. Geneva: WHO; 2021. Available from: https://www.who.int/publications/i/item/WHO-2019-nCoV-genomic_sequencing-2021.1
8. Maljkovic Berry I, Melendrez MC, Bishop-Lilly KA, Rutvisuttinunt W, Pollett S, Talundzic E, et al. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. J Infect Dis. 2020 Mar 28;221(Supplement_3):S292-S307.
9. Meredith LW, Hamilton WL, Warne B, Houldcroft CJ, Hosmillo M, Jahun AS, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. Lancet Infect Dis. 2020 11;20(11):1263-72.
10. To KK, Hung IF, Ip JD, Chu AW, Chan WM, Tam AR, et al. COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. Clin Infect Dis. 10.1093/cid/ciaa1275. Available from: https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciaa1275/5897019
11. Ziegler K, Steininger P, Ziegler R, Steinmann J, Korn K, Ensser A. SARS-CoV-2 samples may escape detection because of a single point mutation in the N gene. Euro Surveill. 2020 10;25(39):pii=2001650. Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.39.2001650
12. Artesi M, Bontems S, Göbbels P, Franckh M, Maes P, Boreux R, et al. A Recurrent Mutation at Position 26340 of SARS-CoV-2 Is Associated with Failure of the E Gene Quantitative Reverse Transcription-PCR Utilized in a Commercial Dual-Target Diagnostic Assay. J Clin Microbiol. 2020 09 22;58(10):e01598-20.
13. Wayne DW. Biostatistics: a foundation for analysis in the health sciences. 3rd ed. New York: John Wiley & Sons. 1999.

Annex

Tables A1-A9 show the number of sequences required per time unit and per geographic unit to achieve detection and quantification ability for variant viruses among other circulating viruses. They assume a random or representative sampling scheme. If quantification of the proportion of variants is not a main goal, more efficient approaches targeting specific populations groups can be employed instead. The time unit can, for example, be a month or a week, and the geographic unit can be a country or a region within a country. The sample size for detection is calculated using the binomial method, while the samples sizes for determining the proportion of variants are calculated using the formula [13]:

$$n = \frac{Z^2 P(P-1)}{d^2}$$

Where Z is the Z-statistic for the 95% confidence level, P is the desired expected prevalence and d is the desired absolute precision. The correction for the total number of cases is calculated using the formula:

$$n_{corr} = \frac{n \times M}{n + M}$$

Where M is the total number of cases, it is assumed that sequences are randomly sampled from the population of all positive SARS-CoV-2 tests.

Description of the columns in the table:

1. The expected prevalence for variant frequency detection and/or quantification among all circulating viruses within the chosen unit of time and geography.
2. Number of sequences needed per time unit and geographic unit to be able to detect the virus reliably (95% confidence) at the chosen expected prevalence. However, this number of sequences will not be enough to provide estimates of the prevalence with sufficient precision (refer to points 3-5).
3. Number of sequences needed per time unit and geographic unit to be able to quantify variants with a 95% relative confidence interval of ±50% of the measured value at the expected prevalence. For example, at 1% expected prevalence, the target confidence interval would be ±0.5%, regardless of the actual prevalence.
4. Number of sequences needed per time unit and geographic unit to be able to quantify variants with a 95% relative confidence interval of ±25% of the measured value at the expected prevalence. For example, at 1% expected prevalence, the target confidence interval would be ±0.25%, regardless of the actual prevalence.
5. Number of sequences needed per time unit and geographic unit to be able to quantify variants with a 95% relative confidence interval of ±10% of the measured value at the expected prevalence. For example, at 1% expected prevalence, the target confidence interval would be ±0.1%, regardless of the actual prevalence.

Accurately quantifying the level of a variant at very low levels (<0.5%) is very challenging, as other sources of error likely have a larger impact than the sample size limitation.

Example: To achieve the ability to quantify a variant at and above 2.5% frequency on a weekly basis with a medium precision (CI: ±0.625%) within a country with 5 000 positive SARS-CoV-2 samples per week, 1 270 sequences per week are needed from that country, found in table A6 in the medium precision column on the 2.5% variant proportion row.

Table A1. Number of sequences required, >100,000 total number of cases per time unit and geographic unit (no population size correction)

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	46	184	1 152
10.00%	30	138	553	3 457
5.00%	60	292	1 168	7 299
2.50%	120	599	2 397	14 982
1.00%	300	1 521	6 085	38 032
0.50%	599	3 058	12 232	76 448
0.25%	1 198	6 131	24 525	153 280
0.10%	2 995	15 351	61 404	383 776

Table A2. Number of sequences required, 100 000 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	46	184	1 139
10.00%	30	138	550	3 342
5.00%	60	291	1 154	6 803
2.50%	120	596	2 341	13 030
1.00%	299	1 498	5 736	27 553
0.50%	595	2 967	10 899	43 326
0.25%	1 184	5 777	19 695	60 518
0.10%	2 908	13 308	38 044	79 329

Table A3. Number of sequences required, 50 000 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	46	183	1 114
10.00%	30	138	544	3 133
5.00%	60	289	1 128	5 988
2.50%	119	589	2 236	10 336
1.00%	297	1 455	5 146	17 764
0.50%	588	2 801	8 948	23 212
0.25%	1 156	5 179	14 129	27 379
0.10%	2 748	10 511	21 605	30 669

Table A4. Number of sequences required, 25 000 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	46	182	1 066
10.00%	30	137	533	2 784
5.00%	60	286	1 080	4 831
2.50%	119	575	2 053	7 313
1.00%	293	1 375	4 267	10 385
0.50%	575	2 519	6 590	12 036
0.25%	1 105	4 290	9 027	13 068
0.10%	2 476	7 400	11 589	13 773

Table A5. Number of sequences required, 10 000 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	46	179	964
10.00%	30	135	506	2 178
5.00%	59	278	974	3 257
2.50%	117	544	1 703	4 224
1.00%	285	1 209	2 991	5 094
0.50%	544	2 012	3 972	5 462
0.25%	995	3 002	4 744	5 665
0.10%	1 985	4 253	5 368	5 794

Table A6. Number of sequences required, 5 000 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	45	173	808
10.00%	30	132	459	1 517
5.00%	59	263	815	1 972
2.50%	115	491	1 270	2 290
1.00%	270	973	1 871	2 523
0.50%	490	1 435	2 214	2 610
0.25%	830	1 876	2 434	2 656
0.10%	1 421	2 298	2 589	2 684

Table A7. Number of sequences required, 2 500 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	45	161	611
10.00%	29	125	388	944
5.00%	57	238	615	1 103
2.50%	110	410	842	1 195
1.00%	243	701	1 070	1 256
0.50%	410	912	1 174	1 277
0.25%	623	1 072	1 233	1 288
0.10%	906	1 197	1 272	1 294

Table A8. Number of sequences required, 1 000 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	12	43	139	379
10.00%	28	111	280	486
5.00%	54	192	381	524
2.50%	99	291	457	544
1.00%	196	412	517	557
0.50%	291	477	540	561
0.25%	384	517	552	563
0.10%	475	545	560	564

Table A9. Number of sequences required, 500 total number of cases per time unit and geographic unit

	Number of sequences required per time unit and geographic unit of desired resolution			
Expected prevalence of variant among all circulating viruses	Only detect presence of variant viruses with 95% confidence (no precision)	Determination of proportion of variant		
		Low precision (95% relative CI ±50%)	Medium precision (95% relative CI ±25%)	High precision (95% relative CI ±10%)
25.00%	11	39	109	216
10.00%	27	91	179	246
5.00%	49	139	216	256
2.50%	83	184	239	261
1.00%	141	226	254	263
0.50%	184	244	260	264
0.25%	217	254	262	265
0.10%	244	261	264	265