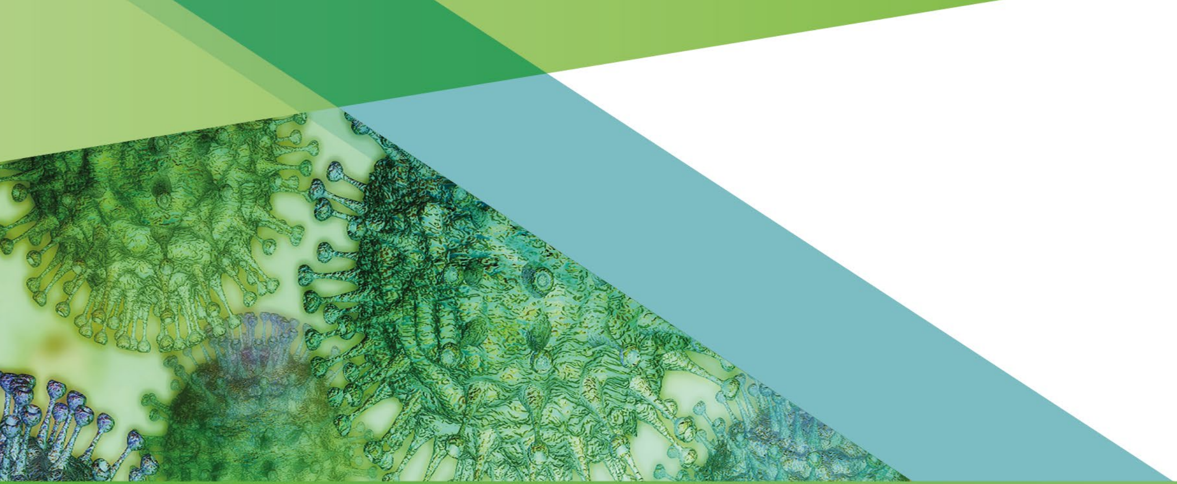

TECHNICAL NOTE

Sequencing of SARS-CoV-2

23 December 2020

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) have been proven as powerful methods 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 note aims to provide guidelines to laboratories 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 information was immediately used to set up viral RNA detection systems by nucleic acid amplification techniques (NAAT).

Suggested citation: European Centre for Disease Prevention and Control. Sequencing of SARS-CoV-2. 23 December 2020. ECDC: Stockholm; 2020.

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

- Investigating virus transmission dynamics and introductions of novel genetic variants;
- Investigating the relationship between clades/lineages and epidemiological data such as transmissibility and disease severity or risk groups to guide public health action;
- Understanding the impact of response measures on the virus population;
- Assessing the impact of mutations on the performance of molecular diagnostic methods;
- Assessing the impact of mutations on the performance of serological methods;
- Assessing the impact of mutations on the performance of antigen detection methods;
- 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 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 impact of mutations and modelling the antigenic properties of the virus to assess the risk of vaccine escape;
- 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 [4]. The United States' Food and Drug Administration has provided Emergency Use Authorisation to one diagnostic detection system using next generation sequencing. 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

Sample selection

Sample selection will depend on the selected objective and available resources and could include reinfections or vaccine failures. 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, as well as SARS-CoV-2 variants emerging in animal populations or causing human outbreaks, especially if they are not explained by epidemiological factors, should be selected for sequencing in order to more effectively monitor virus evolution and changes in the virus genome. 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.

Sequence methodologies

There are several methods available for sequencing SARS-CoV-2 from clinical samples [7]. 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 which 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 research 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 to ECDC and the WHO Regional Office for Europe promptly. 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, especially when vaccines are introduced;
- Tracking community transmission [5] and closed-setting transmission [8];
- Differentiating between reinfection and prolonged carriage by comparing the genomic sequence from different COVID-19 episodes for the same patient [9];
- Assessing the frequency of mutations that can affect the sensitivity of nucleic acid-based detection assays [10,11] (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/Euro 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

WHO Reference Laboratory	Available Technology	Contact
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 Acquisition ongoing: Illumina NextSeq 1000 Oxford Nanopore GridION	Maria Zambon maria.zambon@phe.gov.uk
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 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

Acknowledgements

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.

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 Lindsjo, 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.

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. Nature Microbiology. 2020.

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. Eurosurveillance. 2020.

4. U.S. Food and Drug Administration (FDA). Emergency Use Authorization [Internet]. 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 [Internet]. 2020 September 16;26(9):1405–10. Available from: http://www.nature.com/articles/s41591-020-0997-y

6. Illumina Inc. Press Release: Illumina Receives First FDA Emergency Use Authorization for a Sequencing-Based COVID-19 Diagnostic Test [Internet]. Available from: https://www.illumina.com/company/news-center/press-releases/2020/8cd141fb-68d0-4144-8922-45693ac3f453.html

7. 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 [Internet]. 2019 October 14; Available from: https://academic.oup.com/jid/advance-article/doi/10.1093/infdis/jiz286/5586940

8. 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.

9. To KK-W, Hung IF-N, Ip JD, Chu AW-H, Chan W-M, Tam AR, et al. Coronavirus Disease 2019 (COVID-19) Re-infection by a Phylogenetically Distinct Severe Acute Respiratory Syndrome Coronavirus 2 Strain Confirmed by Whole Genome Sequencing. Clin Infect Dis [Internet]. 2020 August 25; Available from: https://academic.oup.com/cid/advance-article/doi/10.1093/cid/ciaa1275/5897019

10. 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. Caliendo AM, editor. J Clin Microbiol [Internet]. 2020 Jul 20;58(10). Available from: https://jcm.asm.org/content/58/10/e01598-20

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. Eurosurveillance. 2020