The NeIC PaRI partners’ guide to SARS-CoV-2 genome data sharing

Produced by the Nordic Pandemic Research Infrastructure project (NeIC PaRI)
https://neic.no/pari/

This document aims to 1) outline what constitutes a good and ultimately (re)usable viral genome data record in a data repository with a Nordic perspective; and 2) provide guidance to projects, labs and other organisations producing or commissioning viral sequencing data.

Introduction

Combining information on virus characteristics with clinical and epidemiological data is desirable when studying various aspects of a pandemic. Viral genomic data provide insight into some important virus characteristics and should be reported to and shared with the global research and clinical communities as early and as openly as possible. The following text will provide guidance on sharing data from viral genome sequencing. The focus will be on the processing and documentation required for sharing the data files containing reads produced by sequencing instruments and consensus sequences produced by analytic workflows for genome assembly—both of which constitute valuable resources for different audiences.

Study design and documentation

National and international recommendations from public health authorities, epidemic surveillance programs and research data communities should be considered when planning a new study or surveillance programme. In particular, you could consult relevant guidance issued by national and international surveillance programs while considering widely adopted guidelines for research documentation, and recommendations provided by data sharing platforms and communities such as INSDC[1], EMBL-EBI[2], and RDA[3].

Adequate documentation must be made available with the data for the data to be usable in research. Both the documentation and the data themselves will enable researchers to query and select data matching characteristics that make the data relevant and reliably usable to test scientific claims. Good practices for annotating sequencing experiments suggest that the documentation, at a minimum, should describe the design of the study or surveillance program, the collected specimens and how the samples were prepared, the experimental setup and protocols, and the analysis workflow.

[1] International Nucleotide Sequence Database Collaboration
[2] European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute (EBI)
[3] Research Data Alliance

Table: Some aspects of the study to document and to share with the data

Study design Common aspects across the study	**Sample acquisition** Describe each sample
● Physical and digital storage ● Identification and retrieval ● Procedures and protocols[4] ● Legal and ethical constraints ● Traceability / quality control ● (Long-term) data access	● Sampling strategy ● Collection event ● Host related ● Isolated virus related ● Sample preparation ● Quality assessment
Sequencing Describe each experiment / run	**Analysis & variant detection** Describe each analysis / result
● Equipment ● Configuration ● Library preparation ● Quality assessment ● Output files	● Software tools and versions ● Databases and versions ● Configurations ● Quality assessment ● Output files

NeIC PaRI Documentation practices

Shared documentation practices should be developed and maintained across Nordic partners. Geographic locations should be granular enough to be useful in studies and surveillance on national and regional level. The structure and the terms used to describe aspects of the virus characteristics and the experiments should be compatible with widely adopted practices for data sharing and across national and European registry services. In addition, efforts should be made to adopt and improve common protocols and conventions for automated processing across Nordic partners.

Aspects of special concern to Nordic research and surveillance efforts include collecting geographic information with sufficient precision using definitions that are openly available and stable over time (e.g., at least to the precision of a municipality and definitions that are commonly used across national registers). Other general concerns include using naming conventions that facilitate citations and tracing derived data to their original sources (e.g., consistent naming of organisations and sample identifiers that can be traced back to an issuing organisation); maintaining references across regional registers and biobanks; that categories used in stratification by age are compatible with national registers; and that reports and derived data maintain a record to indicate how the data was collected (e.g., self-reported, from which register, etc.).

[4] E.g., on how to reuse, collect, process, analyse, and preserve the biological samples and/or data

Aspects related to common workflows for processing SARS-CoV-2 genome data includes keeping references to which protocols and versions were used in preparing and sequencing each individual sample and recording which samples and sequencing libraries were prepared together. This information can be used to identify and address issues related to both workflow/library specific artifacts and sample contamination and it would also be used to choose appropriate configurations and versions in analytic workflows. Some of this information can also be reflected in naming conventions for samples and libraries.

Selected references on study design and documentation

The following resources can be useful when planning a new study or revising practices for documentation or data management:

- WHO's Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health, 8 January 2021, notably chapter 3 on project planning/study design and chapter 4 on data sharing guidance.
- Ten simple rules for annotating sequencing experiments, general guidance on how to prepare documentation for sequencing projects
- ELIXIR's RDMkit, guidance on data management covering topics from data protection and sensitive data to data quality, data publishing and the COVID-19 Data Portal
- The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology, examples of how to design data collection templates and practices for effective data sharing
- ECDC's Methods for the detection and identification of SARS-CoV-2 variants
- WHO's Guidance for surveillance of SARS-CoV-2 variants: Interim guidance, 9 August 2021
- ECDC's Surveillance and study protocols
- ECDC's TESSy reporting protocol for COVID-19

Data sharing platforms and data standards

The most impactful way to share data is to use established data sharing platforms that can reach a wide scientific and clinical audience. A good data sharing platform will provide mechanisms to search, cite/identify and retrieve data and documentation that match any given virus isolate and should have policies in place to ensure that data will be preserved and reusable in the long-term. Several platforms can be used in parallel to reach different audiences and include national and European healthcare surveillance systems administered by public health authorities, such as the ECDC[5]'s TESSy/EpiPulse; international research data exchanges such as INSDC (EMBL-EBI, NCBI[6], DDBJ[7]) and the Federated EGA network; and virus specific initiatives such as GISAID.

[5] European Centre for Disease Prevention and Control
[6] National Center for Biotechnology Information (USA)
[7] DNA Data Bank of Japan

When the host is a human research subject or patient, some data may need to be anonymised (e.g., masked or removed) before it can be shared. The documentation submitted with the data should outline why and how the data was masked and—if applicable—how a researcher can apply for access to the unmasked data.

Note: Consensus sequences are produced by analytic workflows and it is important to share enough documentation to allow researchers to assess the quality of the sequence alignment and variant detection process and the resulting sequence. Sharing the corresponding reads and a description of the analytic workflow enables researchers to do a sufficient assessment to rely on the data in their research and offers an opportunity to improve the analysis process.

Platforms for SARS-CoV-2 genome data sharing

The most prominent platforms for SARS-CoV-2 genome data sharing in Europe are the European COVID-19 Data Portal and GISAID's EpiCoV™ database. Data should be deposited to *both* of these platforms and measures should be taken to ensure that consensus sequences and corresponding instrument reads can be referenced across the two platforms and across surveillance systems, such as the NCOV* reporting subjects in ECDC's TESSy / EpiPulse. Access to instrument reads is important for validation and analysis using alternative or improved workflows.

GISAID focuses on data sharing related to viruses and the EpiCoV™ database is specifically designed for sharing COVID-19 and SARS-CoV-2 genome data. EpiCoV™ is an important resource for many public health authorities and researchers but the database is currently limited to provide access to consensus sequences and enforces a license that places some restrictions on how and by whom the data can be accessed and reused in downstream analysis.

The European COVID-19 Data Portal is also specifically designed for sharing COVID-19 and SARS-CoV-2 data and provides access to a wide range of data including consensus sequences and reads submitted to ENA[8]—or any member of INSDC[9]. Data submitted to ENA is provided under a free and unrestricted access policy that enables further innovations such as public dashboards and analysis pipelines and targeted or local surveillance initiatives.

Contributions and access to national and European surveillance systems is usually administered by national health authorities. Some of the data contained in these systems may not be suitable for widespread or public sharing and the main purpose of the systems is to support policy and public health actions.

Note: Data files with reads produced by sequencing instruments often contain fragments of the host organism's DNA[10]. When the host is a human research subject or patient, these fragments must be masked or removed from the data files before they can be submitted to ENA.

[8] European Nucleotide Archive

[9] International Nucleotide Sequence Database Collaboration

[10] Some kits also include targeted amplification of human host control sequences

Data standards for SARS-CoV-2 genome data sharing

There are several ways to complete a submission to EpiCoV™ and to ENA, ranging from interactive interfaces provided by the platforms to command-line-based software, Galaxy-workflows and API:s for scripts/custom software. This section will focus on the data standards and the documentation that are required to complete a submission.

Both ENA and EpiCoV™ impose restrictions on what file formats they can accept and how the data must be prepared before uploading them to the platforms. In order to complete a submission to both platforms the consensus sequence must be saved in a FASTA format file, while the corresponding reads must be saved as BAM, CRAM or FASTQ format files. The files must also be anonymized by removing human related reads and identifiers followed by compressing (GZIP) and recording checksums (MD5) (for the compressed files).

The minimum documentation that is required by the data sharing platforms consists of the information that is required for the platform to process the data files, to allow its users to query the data in a meaningful way, and to present it in browsable categories or visual displays. Additional information should be provided, to the extent that it is available and would be meaningful to a researcher assessing its potential for reuse. Any information that could be used to query and discover the data in the platform should be entered in the corresponding fields provided by the platform. Supplementary information can also be provided as links and references to publications, protocols, software versions, biobanks, related data records, or instructions on how to request related data with restricted access.

The following subsections provide a summary of the information that is technically required to be able to submit a consensus sequence with corresponding reads to the European COVID-19 data portal via ENA and to GISAID's EpiCoV™. Additional information about the sample must be added to promote future reuse. Please refer to the submission guidelines for ENA, EpiCoV™ and the section below on NeIC PaRI data sharing practices.

Study or data sharing initiative (minimum)

Study title	A short description of the study or data sharing initiative akin to an article title	Free text
Study abstract	A more extensive free-form description of the study or data sharing initiative. This can include additional information about the initiative, how to access related data, and where to find relevant contact details.	Free text, XML-encoded HTML
ENA study accession	Automatically issued by ENA as part of the submission process. A unique identifier designated to this study.	ERP#####

Sample and sample acquisition (minimum)

Collecting institution name	The authoritative international name of the organisation responsible for the clinical specimen or isolate	Free text
Collecting lab address	The international postal address of the organisation above	Free text
Host species	The species from which the specimen was collected, e.g., Human (Homo sapiens, NCBI:txid9606)	Reference the NCBI Taxonomy database
Collection date	The date at which the specimen was collected given as a valid a full date, a month or a year Note that week numbers cannot be used at the time of writing.	YYYY, YYYY-MM, YYYY-MM-DD
Country	The country from where the specimen originates	Reference the INSDC Country List
Virus name/identifier	A nationally unique name given to the virus sequence isolated with this sample, e.g., 03_SE600_000200073983	A-z, numbers, hyphens and underscores
GISAID Accession ID	Automatically issued by GISAID as part of the submission process. A unique identifier designated to this sample/sequence.	
ENA sample accession	Automatically issued by ENA as part of the submission process. A unique identifier designated to this sample.	ERS#####

Note: ENA may list additional fields as required during the submission, but they can be reported to be not applicable or missing according to the INSDC Missing Value Reporting Terms.

Sequencing experiment (minimum)

Sequence lab	The authoritative international name of the organisation responsible for the sequence data	Free text
Sequence lab address	The international postal address of the organisation above	Free text
Sequencing technology	The instrument platform and model	Reference the ENA Training Modules: Instrument

Library strategy	The sequencing technique intended for this library	Reference the ENA Training Modules: Library strategy
Library source	The type of source material that is being sequenced	Reference the ENA Training Modules: Library source
Library layout	Whether to expect single or paired end reads	Single or Paired,
Insert size	Insert size for paired reads	Number
Library selection	Method used to enrich the target in the sequence library preparation	Reference the ENA Training Modules: Library selection
ENA experiment accession	Automatically issued by ENA as part of the submission process. A unique identifier designated to this experiment	ERX######
ENA run accession	Automatically issued by ENA as part of the submission process. A unique identifier designated to files related to running this experiment	ERR######

Genome assembly (minimum)

Assembly program	The software and version used to assemble the sequence (possibly also a doi referring to the specific version used)	Free text
Assembly method	A free form description and/or reference to the assembly procedure/protocol	Free text
Coverage (numeric)	A number representing coverage as defined by the study or surveillance program, e.g., *10x approximate minimum local coverage over 95% of the genome* would be 10.	Numeric
ENA analyses accession	Automatically issued by ENA as part of the submission process. A unique identifier designated to the analysis performed to assemble the consensus sequence	ERZ######

NeIC PaRI Data sharing practices

The following subsections provide a summary of the information that would be recommended to power the PaRI dashboard and data analysis deliverables. Additional information should be provided, to the extent that it is available and would be meaningful to a researcher assessing its potential for reuse. Any information that could be used to query and discover the data in the platform should be entered in the corresponding fields provided by the platform. Supplementary information can also be provided as links and references to publications, protocols, software versions, biobanks, related data records, or instructions on how to request related data with restricted access.

Sample and sample acquisition (recommended)

Note: Where the documentation is available but not shared, the availability of this information can be indicated by assigning the values ”restricted access” or "not provided" as described by INSDC Missing Value Reporting Terms.

Region	The geographic location from where the specimen originates	Reference to the municipality
Host age	The age of the host or age range group that the host belongs to, e.g. 60 years	# unit, #-# unit
Host sex	The biological sex of the host	male, female, not provided
Sample description	A free-form text describing the sample, its origin, and its method of isolation. Reference protocols used by name, version and document number/doi.	Free text
Additional fields	Extend this table with additional rows corresponding to fields from the NeIC PaRI Data dictionary for SARS-C… PaRI partners suggest including: Symptomatic status, Vaccination status, COVID-19 symptoms, Reason for test, In transit/travel, Travel information, Test location, Host isolation source, Sample description	
Custom fields	Extend this table with additional rows corresponding to additional attributes documented for this sample. NCBI’s list of sample attributes can be used as a starting point.	

Sequencing experiment

Library construction protocol	A description of how the sequencing library was constructed. Reference protocols used by name, version (and document number/doi if available), e.g. “amplified with Artic v3 primer set” for ampliconic data	Free text
Library name	The submitter's name for the sequencing library. Naming conventions can be used to indicate which samples and sequencing libraries were prepared together.	Free text
Additional fields	Extend this table with additional rows corresponding to fields from the NeIC PaRI Data dictionary for SARS-Co…	
Custom fields	Extend this table with additional rows corresponding to additional factors documented for this experiment	

Genome assembly

Description	A free-form description of the consensus sequence and assembly method / configuration	Free text
Additional fields	Extend this table with additional rows corresponding to fields from the NeIC PaRI Data dictionary for SARS-Co…	
Custom fields	Extend this table with additional rows corresponding to additional factors documented for this assembly	

Selected references on data sharing

The following resources can be useful when planning a new study or revising practices for documentation or data management:

- PaRI Data Dictionaries and Data Mappings
- ELIXIR’s RDMkit
- EBI: SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2) Submissions — ena-browser-docs latest documentation
- GISAID: Submitting Data to the EpiFlu™ Database
- Galaxy training: Submitting raw sequencing reads to ENA

- The PHA4GE SARS-CoV-2 Contextual Data Specification for Open Genomic Epidemiology.
- SARS-CoV-2 ENA submission workflow + guidance for structuring and releasing metadata.
- SARS-CoV2 GISAID submission protocol.

National support / infrastructures

Estonia

https://covid19dataportal.ee/
https://koroona.ut.ee/

Norway

ELIXIR Norway Research support
ELIXIR Norway COVID-19 resources
ELIXIR Norway SARS-Cov-2 database

Sweden

Swedish COVID-19 Data Portal Support Services
SciLifeLab Data Guidelines

NeIC PaRI case study

ELIXIR Norway has offered services and infrastructure to prepare and submit data to ENA on behalf of NIPH. The process involved setting up legal agreements for data processing, designing data flows to remove human host DNA from reads in a secure environment, and converting data to formats that can be widely shared with the scientific community. The ELIXIR Norway Research Support team can be contacted directly for support to store, analyse and submit data to ELIXIR deposition databases.

ETAIS[11] and ELIXIR Estonia has provided services and infrastructure to analyse and submit data to ENA on behalf of the public health authority of Estonia. The process involved setting up legal agreements, compute infrastructure and establishing a workflow to allow data to be shared on the same day it comes off the sequencer. The Galaxy platform and a set of Galaxy tools curated for COVID-19 initiatives were used to fully automate the process from metadata handling and sequence data processing up to uploading the raw sequence to ENA and presenting the data on local dashboards displaying real-time spread of the virus.

[11] Estonian Scientific Computing Infrastructure

ELIXIR Sweden and SciLifeLab's Data centre has integrated services and infrastructure offerings in the SciLifeLab/KAW[12] National COVID-19 Research Program. The projects were offered support ranging from data design to data sharing and the national COVID-19 data portal functioned as a hub for access to services related to COVID-19 research and to highlight COVID-19 related data.

[12] Knut and Alice Wallenberg Foundation