Documentation

SARS-CoV-2 Sequence Data from Germany

Robert Koch Institute

Cite
Robert Koch Institute. (2025). SARS-CoV-2 Sequence Data from Germany [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.17421493

Abstract
The dataset ‘SARS-CoV-2 Sequence Data from Germany’ consists of complete virus genome sequences and associated metadata from samples collected nationwide. The samples are sequenced and bioinformatically analysed in collaboration with the IMSSC2 laboratory network, the National Reference Centre for Coronaviruses at Charité and the RKI. The dataset enables robust molecular epidemiological analyses of the spread of SARS-CoV-2 in Germany and represents a central resource for research and public health surveillance.

Table of Content

--- die deutsche Version finden Sie hier ---

Information on the data set and context of origin

A central component of successful pathogen surveillance is understanding the spread of a pathogen and its pathogenic properties. Knowledge of the pathogen genome is an important source of information here. The detection of mutations in the genome of a pathogen makes it possible to reconstruct relationships, uncover transmission routes and predict resistance. The Integrated Genomic Surveillance (IGS) of SARS-CoV-2 aims to monitor the spread of the virus and in particular of virus variants of concern in the population and to closely observe any changes in the virus that occur. The public provision of genomic data is of particular importance in order to enable scientists in Germany and worldwide to carry out their own analyses.

As part of the Coronavirus Surveillance Ordinance, SARS-CoV-2 sequence data from all over Germany were transmitted to the RKI via the German Electronic Sequence Data Hub (DESH) until 31.05.2023. With the expiration of the ordinance, samples will be provided by the IMSSC2 laboratory network in the future and sequenced, analyzed and made available here at the RKI. Despite the reduced number of samples, the careful selection of the participating laboratories ensures a representative insight into the virus population (Djin Ye Oh *et al.* **2022**). In addition, sequences from the NRZ Coronaviruses at the Charité will be contributed to complement the IMSSC2 network.

Administrative and organizational information

The dataset "SARS-CoV-2 sequence data from Germany" is provided by the Robert Koch Institute for research work related to SARS-CoV-2 surveillance in the IGS project.

Data collection at the RKI is carried out with the expiry of the Coronavirus Surveillance Ordinance via the IMSSC2 laboratory network under the direction of FG 17 | Influenza viruses and other viruses of the respiratory tract and by the National Reference Center for Coronaviruses.

As part of the IGS project, the data produced by MF1 | Genome Competence Centre will be analyzed bioinformatically. Questions regarding the project can best be directed to IGS@rki.de.

The coordination and collection of reporting data is carried out by FG 36 | Respiratory communicable diseases.

Publication of the data, data curation and quality management of the (meta-)data are carried out by the RKI's MF 4 | Specialized and Research Data Management department. Questions about data management can be directed to the Open Data Team of the MF4 department (OpenData@rki.de).

Data collection

The IMSSC2 laboratory network consists of ~20 laboratory medical facilities in 13 federal states, which send randomly selected SARS-CoV-2-positive sample material to the RKI on a weekly basis. Here, whole genome sequencing and further phylogenetic and genome biology analyses are carried out to identify the most common SARS-CoV-2 lineages circulating in Germany. The results are published promptly on the RKI website and in scientific journals and contribute to the assessment of the current epidemiological situation of COVID-19. The IMSSC2 data is supplemented by sequences collected by the National Consiliary Laboratory for Coronaviruses. The data from both sources is made available to the public via GitHub and other public databases. Also included in the dataset are SARS-CoV-2 sequence data from all over Germany that were submitted to the RKI via the German Electronic Sequence Data Hub (DESH) by May 31, 2023.

Assignment of virus lines based on pangolin

The assignment of known virus lines to the collected sequences is carried out using Pangolin. When a new version or updated lineage definitions of Pangolin are released, the lineage information for the entire sequence collection is reassigned to the entire sequence dataset. The information about the lineage and the Pangolin version used can be found for each sequence in the metadata.

The information provided on the virus lineages corresponds to the current PANGOLIN Lineage Format. Only the "Taxon" column has been renamed SEQUENCE.ID to facilitate subsequent use. The SEQUENCE.ID, which is contained in all three data, is central for linking the developmental lines with the other data. PANGOLIN Lineage Format is authoritative in case of contradictions.

Quality management

The data collected by DESH passed the quality control (QC) of the IGS at the RKI according to published criteria (see: rki.de - DESH Qualitaetskriterien.pdf). In addition, for all sequences, including IMSSC2 samples, a bioinformatic QC of the sequence is performed with PRESIDENT: PaiRwisE Sequence IDENtiTy with an identity threshold of 70% and an N threshold of 20%. The metadata QC checks the metadata for incorrect data and entries that would influence further processing.
If the QC for metadata or sequence data is not passed, this data is not made publicly available in order to ensure the high quality of the public dataset.

Structure and content of the dataset

The dataset includes genomic sequences of SARS-CoV-2 isolates from all over Germany and associated metadata. The dataset contains:

- Submitted SARS-CoV-2 genome sequences
- Metadata on SARS-CoV-2 genome sequences
- License including the usage license of the dataset
- Metadata file for import into Zenodo
- Information on VOCs and VOIs
- List of relevant lineages

SARS-CoV-2 sequence data

The SARS-CoV-2 sequence data is provided in the root directory under "SARS-CoV-2-Sequenzdaten_Deutschland.fasta.xz".

> SARS-CoV-2-Sequenzdaten_Deutschland.fasta.xz

Structure of the sequence data

The file provided contains sequence entries that are structured according to the FASTA format. In this format, each entry begins with a short description, also known as a header or "description line". This line is identified by a ">" character at the beginning of the line. The header is followed by the sequence itself, which is a sequence of nucleic acids in IUB/IUPAC format

Each sequence ends with the start of a new sequence entry, indicated by a new header, or, in the case of the last sequence entry, with the end of the file.

In the sequence data provided, the header corresponds to the igs_id, which allows a simple link to the metadata provided.

- Header: "><igs_id> version=<version> id=<genome_id> <contig_index>"
- Nucleic acid sequence: IUB/IUPAC standard

This results in the following exemplary structure of a .fasta file:

```
>IGS-101XX-CVDP-XX version=1 id=939421ee-feab-4b79-9f19-6dc248e0ee89 0
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNACCAACCAACTTTCGATCTCTT...
>IGS-101YY-CVDP-YY version=0 id=08f5d734-d135-4d2a-9680-bc5a795b2d34 0
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNACCAACTCTCGGCTGCATGCT...
```

Compression of the sequence data

The SARS-CoV-2 sequence data is provided as an xz-compressed .fasta file. This results in the file extension .fasta.xz. Linux line breaks are used.

- Character set: UTF-8
- Compression: .xz
- Included file format: .fasta
- Line breaks: Linux line breaks

The files can be unpacked on common operating systems, for example with the programs 7zip or XZ Utils. Compression is performed as the .fasta files in particular are several gigabytes (GB) in size.

Sequence metadata

The sequence metadata is provided in "SARS-CoV-2-Sequenzdaten_Deutschland.tsv.xz". This data also contains the assigned virus lines.

> SARS-CoV-2-Sequenzdaten_Deutschland.tsv.xz

Variables and values

The file SARS-CoV-2-Sequenzdaten_Deutschland.tsv.xz contains the variables and their values shown in the following table. A machine-readable data schema is stored in Data Package Format in tableschema_SARS-CoV-2-Sequenzdaten_Deutschland.en.json:

> tableschema_SARS-CoV-2-Sequenzdaten_Deutschland.en.json

Variable	Type	Characteristic	Description
igs_id	string	Example: `IGS-10099-CVDP-01A2C74B-54A8-47B1-B7E4-6562C6231234`	A unique identifier that combines sequence data and metadata. This identifier is used as part of the FASTA ID in the sequence data.
date_of_sampling	date	Format: `YYYY-MM-DDTHH:MM:SS`	Date of sampling in ISO 8601 format without time zone
sequencing_platform	string	Example: `ILLUMINA`	The sequencing platform used based on the ontology approved by ENA

sequencing_reason	string	Values: `random`, `requested`, `clinical`, `other`	Reason for conducting the sequencing. `random`: The sample was taken randomly. `requested`: The sample was taken due to concerns/suspicions about a new variant or something similar. `clinical`: The sample comes from a clinical setting. `other`: The reason is none of the above.
isolation_source	string	Example: `Nasopharyngeal swab (specimen)`	DEMIS Vocabulary
lab_sequence_id	string	Example: `873a7cc28d29e3f17b0544ea6e9e84 36defe32f6d60649159ee8ac78d414 7ac9`	FASTA ID used by the laboratory in encrypted form
date_of_submission	date	Format: `YYYY-MM-DDTHH:MM:SS`	Date of receipt of the genome at the RKI in ISO 8601 format without time zone
version	integer	Values: `≥0`	Version of the sequence starting with 0
diagnostic_lab.demis_lab_id	string	Example: `DEMIS-10099`	Identification number of the primary diagnostic laboratory
diagnostic_lab.postal_code	string	Example: `50858`	Postal code of the primary diagnostic laboratory
sequencing_lab.demis_lab_id	string	Example: `DEMIS-10099`	Identification number of the sequencing laboratory
sequencing_lab.postal_code	string	Example: `50858`	Postal code of the sequencing lab
genome.gtrs	string	Examples: `[{"date_of_creation": "2025-05-19T11:35:46.427598"`, `"method_version": "4.3.1"`, `"database_version": "PUSHER-v1.32"`, `"genomic_typing_result": "BA.2"`, `"date_of_assignment": "2025-01-30T16:14:14.218144"`, `"genomic_method": {"name": "Pangolin Lineage"}`, `"additional_information": "{\"note\": \"Usher placements: BA.2(1/1)\"`, `\"conflict\": 0`, `\"qc_notes\": \"Ambiguous_content:0.02\"`, `\"qc_status\": \"pass\"`, `\"is_designated\": false}"`, `"date_of_modification": "2025-``05-19T11:35:46.427598"}]`	genomic typing results (GTR) in JSON format

The file SARS-CoV-2-Entwicklungslinien_berichtet.tsv contains the variables and their values shown in the following table. A machine-readable data schema is stored in Data Package Format in tableschema_SARS-CoV-2-Entwicklungslinien_berichtet.en.json:

> tableschema_SARS-CoV-2-Entwicklungslinien_berichtet.en.json

Variable	Type	Characteristic	Description
LINEAGE	string	Example: `JN.1`	Assigned Pangolin Lineage
WHO_LABEL	string	Example: `Omikron`	Name of the virus variant assigned by the World Health Organization

CONTRIBUTING_LINEAGES	string	Example: `JN.1.1.10`	Pangolin lineages derived from the lineage

The file SARS-CoV-2-Entwicklungslinien_zu_Varianten.tsv contains the variables and their values shown in the following table. A machine-readable data schema is stored in Data Package Format in tableschema_SARS-CoV-2-Entwicklungslinien_zu_Varianten.en.json:

> tableschema_SARS-CoV-2-Entwicklungslinien_zu_Varianten.en.json

Variable	Type	Characteristic	Description
LINEAGE	string	Example: `BA.2`	Assigned Pangolin Lineage
WHO_LABEL	string	Example: `Omikron`	Name of the virus variant assigned by the World Health Organization
CONTRIBUTING_LINEAGES	string	Example: `JN.13.1`	Pangolin lineages derived from the lineage
COLOR	any		Legacy variable. It is no longer relevant and will be removed perspectively.
variant_category	string	Values: `VOC`, `VOI`	WHO Classification of the variant as VOC (variant of concern) or VOI (variant of interest)

Formatting the sequence metadata

The sequence metadata is provided as an xz-compressed, comma-separated .csv file. This results in the file extension .csv.xz. The character set used in the .csv file is UTF-8. The individual values are separated by a comma ",". Dates are formatted in the ISO 8601 standard.

- Character set: UTF-8
- Date format: ISO 8601
- Compression: .xz
- Included file format: .tsv
- .csv separator: Tab "\t"

The files can be unpacked on common operating systems, for example with the programs 7zip or XZ Utils. Compression is performed as the .fasta files in particular are several gigabytes (GB) in size.

Metadata

To increase findability, the provided data are described with metadata. The Metadata are distributed to the relevant platforms via GitHub Actions. There is a specific metadata file for each platform; these are stored in the metadata folder:

> Metadaten/

Versioning and DOI assignment are performed via Zenodo.org. The metadata prepared for import into Zenodo are stored in the zenodo.json. Documentation of the individual metadata variables can be found at https://developers.zenodo.org/representation.

> Metadaten/zenodo.json

The zenodo.json includes the publication date and the date of the data status in the following format (example):

```
"publication_date": "2024-06-19",
"dates": [
  {
    "start": "2023-09-11T15:00:21+02:00",
    "end": "2023-09-11T15:00:21+02:00",
    "type": "Collected",
    "description": "Date when the dataset was created"
  }
],
```

Additionally, we describe tabular data using the Data Package Standard.
A Data Package is a structured collection of data and associated metadata that facilitates data exchange and reuse. It consists of a `datapackage.json` file that contains key information such as the included resources, their formats, and schema definitions.

The Data Package Standard is provided by the Open Knowledge Foundation and is an open format that enables a simple, machine-readable description of datasets.

The list of data included in this repository can be found in the following file:

> datapackage.json

For tabular data, we additionally define a Table Schema that describes the structure of the tables, including column names, data types, and validation rules. These schema files can be found in:

> Metadaten/schemas/

Guidelines for reuse of the data

Open data from the RKI are available on Zenodo.org, GitHub.com, OpenCoDE, and Edoc.rki.de:

- https://zenodo.org/communities/robertkochinstitut
- https://github.com/robert-koch-institut
- https://gitlab.opencode.de/robert-koch-institut
- https://edoc.rki.de/

License