Dataset documentation

SARS-CoV-2 Sequence Data from Germany

Robert Koch Institute | RKI
Nordufer 20
13353 Berlin

Robert Koch Institute (2023): SARS-CoV-2 Sequence Data from Germany, Berlin: Zenodo. DOI: 10.5281/zenodo.7735109

Information about the dataset and context of origin.

Accurate knowledge of the properties of SARS-CoV-2 is of central importance for planning measures to contain COVID-19. Mutations of the virus play a special role in this context. For successful containment of the pandemic, it is therefore crucial to obtain a detailed overview of the spread patterns of specific SARS-CoV-2 mutations and also to detect new mutations at an early stage.

For this purpose, the Robert Koch Institute provides the systems for nationwide molecular surveillance. Every laboratory in Germany that sequences SARS-CoV-2 is required by the Verordnung zur molekulargenetischen Surveillance des Coronavirus SARS-CoV-2 to transmit the sequence and associated metadata to the Robert Koch Institute. This transmission is implemented via the German Electronic Sequence Data Hub (DESH).

In the project "OSEDA - Open Sequence Data", the RKI aims at providing the processed and quality-controlled sequence data together with a selection of clinical epidemiological data via the publicly accessible repositories of the European Nucleotide Archive (ENA) and GISAID for further research projects.

Figure: System Structure of the German Electronic Sequence Data Hub (DESH)

> [Kontextmaterialien/2021-01-29_DESH_CorSurV_BAnz_AT_V2.pdf]](https://github.com/robert-koch-institut/SARS-CoV-2-Sequenzdaten_aus_Deutschland/blob/master/Kontextmaterialien/2021-01-29_DESH_CorSurV_BAnz_AT_V2.pdf)

Administrative and organizational data.

The dataset "SARS-CoV-2 Sequence Data from Germany" is provided by the Robert Koch Institute for research related to the SARS-CoV-2 pandemic.

Data transfer to the RKI is done via the system of the German Electronic Sequence Data Hub (DESH). Part of this system is the DESH platform provided by Bundesdruckerei through which sequence data can be transmitted by sequencing laboratories (can only be accessed with an individual certificate). Questions regarding the DESH platform can be directed to the DESH team at desh@rki.de.

Data publication, data curation, and quality management of the (meta-)data are performed by the RKI department MF 4 | Research Data Management. Questions regarding data management can be directed to the Open Data Team of the Department MF4 (OpenData@rki.de).

Submission of sequence data.

The RKI's DESH project website contains instructions for providing sequence data to assist sequencing laboratories in the process of providing metadata and sequence data via https://desh.bdr.de (only accessible with an individual certificate). Sequencing laboratories are required to meet certain quality criteria for sequence data. Compliance with the quality criteria is ensured by the sequencing laboratories. The RKI has no knowledge of the underlying raw data (so-called "reads").

> Kontextmaterialien/2021-02-18_DESH_Anleitung_zur_Bereitstellung_Sequenzdaten.pdf
> Kontextmaterialien/2021-02-08_DESH_Qualitätsvorgaben_für_die_Sequenzdaten.pdf

Publication of sequence data

In the publication of sequence data in ENA and GISAID, there is a delay in publication due to necessary intermediate steps. Therefore, the RKI additionally provides all sequence data received via DESH on a daily basis.

⚠ The data set has not undergone any further quality control by the RKI. It should be noted that data in this dataset for example:

- contain sequence data of low quality
- predict unverified frameshifts
- is present more than once in the dataset
- have already been published by the sequencing laboratory

Therefore, the data published here cannot be compared with the weekly Report on viral variants of SARS-CoV-2 in Germany by the RKI. Furthermore, these data can explicitly not be used as a basis for billing with the KBV.

Structure and content of the data set

The dataset contains data on SARS-CoV-2 sequences in Germany and the contextual materials supporting the data processing. Included in the dataset are:

- Sequence data of the submitted SARS-CoV-2 genome sequences
- Metadata on the SARS-CoV-2 genome sequences
- Information about the viral lineages (PANGOLIN Lineages) of the SARS-CoV-2 genome sequences
- License of the dataset.
- Dataset documentation and context materials in German language
- Metadata file for import into Zenodo

Formatting of the sequence data

The SARS-CoV-2 sequence data is provided as an xz-compressed .fasta file. This results in the .fasta.xz file extension. The lines are wrapped at 80 characters. Linux line breaks are used.

- Character set: UTF-8
- Compression: .xz

- Included file format: .fasta
- Line length: maximum 80 characters
- Line breaks: Linux line breaks

Formatting the metadata

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

- Character set: UTF-8
- Date format: ISO 8601
- Compression: .xz
- Included file format: .csv
- .csv separator: Comma ","

Formatting of the developmental lineage information

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

- Character set: UTF-8
- Date format: ISO 8601
- Compression: .xz
- Included file format: .csv
- .csv separator: Comma ","

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

SARS-CoV-2 sequence data and sequencing metadata.

The SARS-CoV-2 sequence data are provided on a daily basis in the main directory under "SARS-CoV-2-Sequenzdaten-Deutschland.fasta.xz". The same applies to associated metadata contained in the dataset under "SARS-CoV-2-Sequenzdaten-Deutschland.csv.xz" and the lineages contained in the dataset under "SARS-CoV-2-Entwicklungslinien_Deutschland.csv.xz". **Lineage information is not available for all SARS-CoV-2 sequence datasets.**

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

The data is extended every day by the processed sequence data of the current day (cummulation). Sequence data sent in after 20:00 are not processed until the following day. The data status therefore always represents the status of the current day at 19:59.

Structure of the sequence data

The sequence entries of the .fasta file provided start with a one-line description, the header line, also called "Description line". The header line is followed by the nucleic acid sequence of the sequenced SARS-CoV-2 virus.
The header line is marked by a ">", a sequence ends with the end of the file or another sequence tag, start with a new header line.
In the provided sequence data, the header contains the FASTA ID, which in the data corresponds to the IMS_ID of the sample. The IMS_ID allows linkage to the metadata. The coding of the nucleotides of the sequence data follow the IUB/IUPAC standard.

- Header: >IMS_ID
- Nucleic acid sequence: IUB/IUPAC standard

This results in the following structure of a .fasta file as an example:

```
>IMS-101XX-CVDP-XX
NNNNNNNNNNNNNNNNNNNNNNNNACCAACCAACTTTCGATCTCTT
GTTCTCTAAACGAACTTTAAAATCTGTGGCTCTCTCGGCTGCATGCTTAGTGCACT
...
YGACCGGGTGTGACCGAAAGGTAAGATGGAGCCTTGTCCCTGGTTTCAACGAGAA
GGGAGGACTTGAAAGAGCCACCACATTTTCACCGAGGCN
>IMS-101YY-CVDP-YY
NNNNNNNNNNNNNNNNNNNNNNNNACCAACTCTCGGCTGCATGCT
GTTCTCTAAACGAACTTTAAAATCTGTGGCTGTCTTGAAAGAGCCACCACATTTTCA
...
```

Variables and variable characteristics metadata

The metadata file provided as a .csv contains variables listed in the following table as columns. Central to linking the metadata to the genome sequences is the IMS_ID, which is included in all three data files.

Variable	Description	Value Set
IMS_ID	A unique identifier that combines sequence data and metadata. This identifier is used as the FASTA ID in the sequence data	string
DATE_DRAW	Date of sampling	YYYY-MM-DD
SEQ_TYPE	Sequencing platform used	ena
SEQ_REASON	Reason for performing the sequencing	rki
SAMPLE_TYPE	Type of sample	snomed
OWN_FASTA_ID	FASTA ID used by the lab in encrypted form	string
RECEIVE_DATE	Date of data reception at the RKI	YYYY-MM-DD
PROCESSING_DATE	Date of data processing at the RKI (Usually <24 hours after submission by laboratories)	YYYY-MM-DD

Variable	Description	Value Set
SEQUENCING_LAB_PC	Zip code of the sequencing laboratory	string
SENDING_LAB_PC	Zip code of the prime diagnostic laboratory	string
GISAID_ACCESSION	If known, GISAID Accession ID of the sequence	string

For more information on the listed variables, see Instructions for Providing Sequence Data which is also deposited in Kontextmaterialien.

Variables and variable expressions in the developmental lineages

Variable	Description	Value Set
IMS_ID	A unique identifier that combines sequence data and metadata. This identifier is used as the FASTA ID in the sequence data	
lineage	The most likely lineage assigned to a given sequence based on the inference engine used and the SARS-CoV-2 diversity designated.	
conflict	In the pangoLEARN decision tree model, a given sequence gets assigned to the most likely category based on known diversity.	
ambiguity_score	This score is a function of the quantity of missing data in a sequence.	
version	A version number that represents both the pango-designation number and the inference engine used to assign the lineage.	
pangolin_version	The version of pangolin software running.	
pangoLEARN_version	The dated version of the pangoLEARN model installed.	
status	Indicates whether the sequence passed the QC thresholds for minimum length and maximum N content.	
note	If any conflicts from the decision tree, this field will output the alternative assignments.	

The information provided on developmental lineages corresponds to the current PANGOLIN Lineage Format. Only the "Taxon" column has been renamed to IMS_ID for easier re-use. Central to linking the developmental lineages to the rest of the data is the IMS_ID, which is included in all three data. PANGOLIN Lineage Format is authoritative in case of contradictions.

Notes on the subsequent use of the data

⚠ the dataset has not undergone further quality control by the RKI. It should be noted that data in this dataset for example:

- contain sequence data of low quality
- predict unverified frameshifts

- are present more than once in the dataset
- have already been published by the sequencing laboratory

Therefore, the data published here cannot be readily compared with the weekly Report on viral variants of SARS-CoV-2 in Germany by the RKI. Furthermore, these data explicitly cannot be used as a basis for billing laboratories with KBV.

Open research data of the RKI are made available on GitHub.com, Zenodo.org und Edoc.rki.de:

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

Publication metadata.

To increase findability, the provided data is described with metadata. Metadata is distributed to the corresponding platforms via GitHub Actions. A specific metadata file exists for each platform and is stored in the metadata folder:

> Metadaten/

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

> Metadaten/zenodo.json

License