Dokumentation

SARS-CoV-2 Sequenzdaten aus Deutschland

Robert Koch-Institut

Zitieren
Robert Koch-Institut. (2025). SARS-CoV-2 Sequenzdaten aus Deutschland [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.15182587

--- see English version below ---

Informationen zum Datensatz und Entstehungskontext

Ein zentraler Bestandteil einer erfolgreichen Erregersurveillance ist das Verständnis der Verbreitung eines Erregers sowie seiner pathogenen Eigenschaften. Hierbei stellt das Wissen über das Erregergenom eine wichtige Informationsquelle dar. So erlaubt der Nachweis von Mutationen im Genom eines Erregers, Verwandtschaftsbeziehungen zu rekonstruieren, Übertragungswege aufzudecken und Resistenzen vorherzusagen. Die Integrierte Genomische Surveillance (IGS) von SARS-CoV-2 zielt darauf ab, die Verbreitung des Virus und insbesondere von besorgniserregenden Virusvarianten in der Bevölkerung zu überwachen sowie auftretende Veränderungen des Virus genau zu beobachten. Besondere Bedeutung kommt dabei der öffentlichen Bereitstellung der genomischen Daten zu, um Wissenschaftlern in Deutschland und weltweit die Möglichkeit zu eigenständigen Analysen zu eröffnen.

Im Rahmen der Coronavirus-Surveillanceverordnung wurden bis zum 31.05.2023 SARS-CoV-2 Sequenzdaten aus ganz Deutschland über den Deutschen Elektronischen Sequenzdaten-Hub (DESH) an das RKI übermittelt. Mit Ablauf der Verordnung werden künftig Proben durch das IMSSC2 Labornetzwerk bereitgestellt und am RKI sequenziert, analysiert und hier bereitgestellt. Trotz reduzierter Probenanzahl, wird durch die sorgfältige Auswahl der beteiligten Labore ein repräsentativer Einblick in die Viruspopluation gesichert (Djin Ye Oh *et al.* **2022**). Zusätzlich werden Sequenzen vom NRZ Coronaviren an der Charité beigetragen um das IMSSC2 Netzwerk zu ergänzen.

Administrative und organisatorische Angaben

Der Datensatz "SARS-CoV-2-Sequenzdaten aus Deutschland" wird vom Robert Koch-Institut für Forschungsarbeiten im Zusammenhang mit der SARS-CoV-2-Surveillance im IGS Projekt bereitgestellt.

Die Datenerhebung am RKI erfolgt mit Ablauf der Coronavirus-Surveillanceverordnung über das IMSSC2 Labornetzwerk unter der Leitung von FG 17 | Influenzaviren und weitere Viren des Respirationstraktes und durch das Nationale Referenzzentrum für Coronaviren.

Im Rahmen des IGS Projektes werden die produzierten Daten von MF1 | Genome Competence Centre bioinformatisch analysiert. Fragen bezüglich des Projektes können am besten an IGS@rki.de gerichtet werden.

Die Koordinierung und Meldedatenerfassung wird von FG 36 | Respiratorisch übertragbare Erkrankungen durchgeführt.

Die Veröffentlichung der Daten, die Datenkuration sowie das Qualitätsmanagement der (Meta-)Daten erfolgen durch das Fachgebiet MF 4 | Fach- und Forschungsdatenmanagement des RKI. Fragen zum Datenmanagement können an das Open Data Team des Fachgebiets MF4 gerichtet werden (OpenData@rki.de).

Datenerhebung

Das IMSSC2 Labornetzwerk besteht aus ~20 labormedizinischen Einrichtungen in 13 Bundesländern, die wöchentlich zufällig ausgewähltes SARS-CoV-2-positives Probenmaterial ans RKI senden. Hier erfolgt eine Ganzgenomsequenzierung sowie weiterführende phylogenetische und genombiologische Analysen, die eine Identifizierung der häufigsten in Deutschland zirkulierenden SARS-CoV-2 Linien ermöglicht. Die Ergebnisse werden auf der Webseite des RKI und in Fachzeitschriften zeitnah publiziert und tragen zur Bewertung der aktuellen epidemiologischen Lage von COVID-19 bei. Erweitert werden die IMSSC2 Daten durch Sequenzen, die durch das Nationales Konsiliarlaboratorium für Coronaviren erhoben werden. Die Daten aus beiden Quellen werden über GitHub und andere öffentliche Datenbanken der Öffentlichkeit zur Verfügung gestellt. Ebenfalls im Datensatz enthalten sind SARS-CoV-2 Sequenzdaten aus ganz Deutschland die bis zum 31.05.2023 über den Deutschen Elektronischen Sequenzdaten-Hub (DESH) an das RKI übermittelt wurden.

Zuordnung von Viruslinien basierend auf Pangolin

Die Zuordnung bekannter Viruslinien zu den erhobenen Sequenzen erfolgt mittels Pangolin. Mit Erscheinen einer neuen Version oder aktualisierter Liniendefinitionen von Pangolin erfolgt eine Neuzuordnung der Linieninformation für die gesamte Sequenzkollektion den gesamten Sequenzdatensatz. Die Informationen über die Lineage und die genutzte Pangolin Version befindet sich für jede Sequenz in den Metadaten.

Die bereitgestellten Informationen zu den Viruslinien entsprechen dem aktuellen PANGOLIN Lineage Format. Nur die Spalte "Taxon" wurde zur einfacherer Nachnutzung in SEQUENCE.ID umbenannt. Zentral für die Verknüpfung der Entwicklungslinien mit den weiteren Daten ist die SEQUENCE.ID, die in allen drei Daten enthalten ist. PANGOLIN Lineage Format ist bei Widersprüchen authoritativ.

Qualitätsmanagement

Die Daten, die durch DESH erhoben wurden, durchliefen die Qualitätskontrolle (QC) der IGS am RKI nach veröffentlichten Kriterien (siehe: rki.de - DESH Qualitaetskriterien.pdf). Zusätzlich wird für alle Sequenzen, inklusive IMSSC2 Proben, eine bioinformatische QC der Sequenz mit PRESIDENT: PaiRwisE Sequence IDENtiTy durchgeführt mit einen Identitäts-Grenzwert von 70% und einen N-Grenzwert von 20%. Die Metadaten-QC überprüft die Metadaten auf fehlerhafte Daten und Eingaben, die die weitere Verarbeitung beeinflussen würden.
Bei nicht bestehen der QC für Metadaten oder Sequenzdaten werden diese Daten nicht öffentlich bereitgestellt, um die hohe Qualität des öffentlichen Datensatzes zu gewährleisten.

Aufbau und Inhalt des Datensatzes

Der Datensatz umfasst genomische Sequenzen von SARS-CoV-2-Isolaten aus ganz Deutschland und zugehörige Metadaten. Im Datensatz enthalten sind:

- übermittelte SARS-CoV-2-Genomsequenzen
- Metadaten zu den SARS-CoV-2-Genomsequenzen
- Lizenz mit der Nutzungslizenz des Datensatzes
- Metadaten Datei zum Import in Zenodo
- Informationen über VOCs und VOIs
- Liste von relevanten Lineages

SARS-CoV-2-Sequenzdaten

Die SARS-CoV-2-Sequenzdaten werden im Hauptverzeichnis unter "SARS-CoV-2-Sequenzdaten_Deutschland.fasta.xz" bereitgestellt.

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

Struktur der Sequenzdaten

Die bereitgestellte Datei enthält Sequenzeinträge, die nach dem FASTA-Format strukturiert sind. In diesem Format beginnt jeder Eintrag mit einer kurzen Beschreibung, auch Kopfzeile oder "description line" genannt. Diese Zeile wird durch ein ">"-Zeichen am Zeilenanfang gekennzeichnet. Nach der Kopfzeile folgt die Sequenz selbst, die eine Abfolge von Nukleinsäuren im IUB/IUPAC Format darstellt

Jede Sequenz endet mit dem Beginn eines neuen Sequenzeintrages, gekennzeichnet durch eine neue Kopfzeile, oder, im Falle des letzten Sequenzeintrages, mit dem Ende der Datei.

In den bereitgestellten Sequenzdaten entspricht die Kopfzeile der igs_id, was eine einfache Verknüpfung mit den bereitgestellten Metadaten erlaubt.

- Kopfzeile: "><igs_id> version=<version> id=<genome_id> <contig_index>"
- Nukleinsäuresequenz: IUB/IUPAC Standard

Daraus ergibt sich beispielhaft folgende Struktur einer .fasta-Datei:

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

Komprimierung der Sequenzdaten

Die SARS-CoV-2-Sequenzdaten werden als xz-komprimierte .fasta Datei bereitgestellt. Daraus ergibt sich die Dateiendung .fasta.xz. Es werden Linux Zeilenumbrüche verwendet.

- Zeichensatz: UTF-8
- Komprimierung: .xz
- Enthaltenes Dateiformat: .fasta

- Zeilenumbrüche: Linux Zeilenumbrüche

Die Dateien können auf gängigen Betriebssystemen, beispielsweise mit den Programmen 7zip oder XZ Utils, entpackt werden. Die Komprimierung wird vorgenommen, da insbesondere die .fasta-Dateien mehrere Gigabyte (GB) groß sind.

Sequenzmetadaten

Die Sequenzmetadaten werden in der "SARS-CoV-2-Sequenzdaten_Deutschland.tsv.xz" bereitgestellt. Diese Daten enthalten ebenfalls die zugeordneten Viruslinien.

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

Variablen und Werte

Die Datei SARS-CoV-2-Sequenzdaten_Deutschland.tsv.xz enthält die in der folgenden Tabelle abgebildeten Variablen und deren Ausprägungen. Ein maschinenlesbares Datenschema ist im Data Package Standard in tableschema_SARS-CoV-2-Sequenzdaten_Deutschland.json hinterlegt:

> tableschema_SARS-CoV-2-Sequenzdaten_Deutschland.json

Variable	Typ	Ausprägungen	Beschreibung
igs_id	string	Beispiel: `IGS-10099-CVDP-01A2C74B-54A8-47B1-B7E4-6562C6231234`	Ein eindeutiger Identifikator der Sequenzdaten und Metadaten zusammenführt. Dieser Identifikator wird als Teil der FASTA ID in den Sequenzdaten genutzt.
date_of_sampling	date	Format: `YYYY-MM-DDTHH:MM:SS`	Datum der Probeentnahme im ISO 8601 Format ohne Zeitzone
sequencing_platform	string	Beispiel: `ILLUMINA`	Die verwendete Sequenzierungs-Plattform auf Basis der von ENA zugelassenen Ontologie (siehe ena).
sequencing_reason	string	Werte: `random`, `requested`, `clinical`, `other`	Grund für die Durchführung der Sequenzierung. `random`: Die Probe wurde randomisiert genommen. `requested`: Die Probe wurde aufgrund von Bedenken/Verdacht auf eine neue Variante oder Vergleichbares genommen. `clinical`: Die Probe kommt aus einem klinischem Umfeld. `other`: Der Grund it keiner der oben genannten.
isolation_source	string	Beispiel: `Nasopharyngeal swab (specimen)`	DEMIS Vokabular
lab_sequence_id	string	Beispiel: `873a7cc28d29e3f17b0544ea6e9e8436defe32f6d60649159ee8ac78d4147ac9`	Vom Labor genutzte FASTA ID in verschlüsselter Form
date_of_submission	date	Format: `YYYY-MM-DDTHH:MM:SS`	Datum des Eingangs des Genoms am RKI im ISO 8601

			Format ohne Zeitzone
version	integer	Werte: `≥0`	Version der Sequenz startend mit 0
diagnostic_lab.demis_lab_id	string	Beispiel: `DEMIS-10099`	Identifikationsnummer des primärdiagnostischen Labors
diagnostic_lab.postal_code	string	Beispiel: `50858`	Postleitzahl des primärdiagnostischen Labors
sequencing_lab.demis_lab_id	string	Beispiel: `DEMIS-10099`	Identifikationsnummer des sequenzierenden Labors
sequencing_lab.postal_code	string	Beispiel: `50858`	Postleitzahl des sequenzierenden Labors
genome.gtrs	string	Beispiele: `[{"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"}]`	Genomische Typisierungs Resultate (GTR) im JSON-Format

Die Datei SARS-CoV-2-Entwicklungslinien_berichtet.tsv enthält die in der folgenden Tabelle abgebildeten Variablen und deren Ausprägungen. Ein maschinenlesbares Datenschema ist im Data Package Standard in tableschema_SARS-CoV-2-Entwicklungslinien_berichtet.json hinterlegt:

> tableschema_SARS-CoV-2-Entwicklungslinien_berichtet.json

Variable	Typ	Ausprägungen	Beschreibung
LINEAGE	string	Beispiel: `JN.1`	Zugewiesene Pangolin Lineage
WHO_LABEL	string	Beispiel: `Omikron`	Name der Virusvariante, der von der World Health Organisation vergeben wurde
CONTRIBUTING_LINEAGES	string	Beispiel: `JN.1.1.10`	Pangolin Lineages, die von der Lineage abstammen

Die Datei SARS-CoV-2-Entwicklungslinien_zu_Varianten.tsv enthält die in der folgenden Tabelle abgebildeten Variablen und deren Ausprägungen. Ein maschinenlesbares Datenschema ist im Data Package Standard in tableschema_SARS-CoV-2-Entwicklungslinien_zu_Varianten.json hinterlegt:

> tableschema_SARS-CoV-2-Entwicklungslinien_zu_Varianten.json

Variable	Typ	Ausprägungen	Beschreibung
LINEAGE	string	Beispiel: `BA.2`	Zugewiesene Pangolin Lineage
WHO_LABEL	string	Beispiel: `Omikron`	Name der Virusvariante, der von der World Health Organisation vergeben wurde
CONTRIBUTING_LINEAGES	string	Beispiel: `JN.13.1`	Pangolin Lineages, die von der Lineage abstammen
			Veraltete Variable. Ist nicht mehr relevant und wird persepektivisch

COLOR	any		entfernt.
variant_category	string	Werte: `VOC`, `VOI`	WHO Einstufung der Variante als VOC (variant of concern) oder VOI (variant of interest)

Formatierung der Sequenzmetadaten

Die Sequenzmetadaten werden als xz-komprimierte, kommaseparierte .csv-Datei bereitgestellt. Daraus ergibt sich die Dateiendung .csv.xz. Der verwendete Zeichensatz der .csv-Datei ist UTF-8. Trennzeichen der einzelnen Werte ist ein Komma ",". Datumsangaben sind im ISO-8601-Standard formatiert.

- Zeichensatz: UTF-8
- Datumsformat: ISO 8601
- Komprimierung: .xz
- Enthaltenes Dateiformat: .tsv
- .csv-Trennzeichen: Tab "\t"

Die Dateien können auf gängigen Betriebssystemen, beispielsweise mit den Programmen 7zip oder XZ Utils, entpackt werden. Die Komprimierung wird vorgenommen, da insbesondere die .fasta-Dateien mehrere Gigabyte (GB) groß sind.

Metadaten

Zur Erhöhung der Auffindbarkeit sind die bereitgestellten Daten mit Metadaten beschrieben. Über GitHub Actions werden Metadaten an die entsprechenden Plattformen verteilt. Für jede Plattform existiert eine spezifische Metadatendatei, diese sind im Metadatenordner hinterlegt:

> Metadaten/

Versionierung und DOI-Vergabe erfolgt über Zenodo.org. Die für den Import in Zenodo bereitgestellten Metadaten sind in der zenodo.json hinterlegt. Die Dokumentation der einzelnen Metadatenvariablen ist unter https://developers.zenodo.org/#representation nachlesbar.

> Metadaten/zenodo.json

In der zenodo.json ist neben dem Publikationsdatum (`"publication_date"`) auch der Datenstand in folgendem Format enthalten (Beispiel):

```
"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"
  }
],
```

Zusätzlich beschreiben wir tabellarische Daten mithilfe des Data Package Standards.
Ein Data Package ist eine strukturierte Sammlung von Daten und zugehörigen Metadaten, die den Austausch und die Wiederverwendung von Daten erleichtert. Es besteht aus einer datapackage.json-Datei, die zentrale Informationen wie die enthaltenen Ressourcen, ihre Formate und Schema-Definitionen beschreibt.

Der Data Package Standard wird von der Open Knowledge Foundation bereitgestellt und ist ein offenes Format, das eine einfache, maschinenlesbare Beschreibung von Datensätzen ermöglicht.

Die Liste der in diesem Repository enthaltenen Daten ist in folgender Datei hinterlegt:

> datapackage.json

Für tabellarische Daten definieren wir zusätzlich ein Table Schema, das die Struktur der Tabellen beschreibt, einschließlich Spaltennamen, Datentypen und Validierungsregeln. Diese Schema-Dateien finden sich unter:

> Metadaten/schemas/

Hinweise zur Nachnutzung der Daten

Offene Forschungsdaten des RKI werden auf Zenodo.org, GitHub.com, OpenCoDE und Edoc.rki.de bereitgestellt:

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

Lizenz

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.15182587

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: `873a7cc28d29e3f17b0544ea6e9e8436defe32f6d60649159ee8ac78d4147ac9`	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

		19T11:33:46.427596"}]	

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