

SARS-CoV-2-Sequenzdaten aus Deutschland

Robert Koch-Institut | RKI

Nordufer 20

13353 Berlin

You can find an english version of the readme [here](#)

Robert Koch-Institut (2022): SARS-CoV-2-Sequenzdaten aus Deutschland, Berlin: Zenodo. DOI: [10.5281/zenodo.7178853](https://doi.org/10.5281/zenodo.7178853)

Informationen zum Datensatz und Entstehungskontext

Für die Planung von Maßnahmen zur Eindämmung von COVID-19 kommt der genauen Kenntnis der Eigenschaften von SARS-CoV-2 eine zentrale Bedeutung zu. Eine besondere Rolle spielen in diesem Zusammenhang Mutationen des Virus. Für eine erfolgreiche Eindämmung der Pandemie ist es daher entscheidend, einen detaillierten Überblick über die Ausbreitungsmuster spezifischer SARS-CoV-2-Mutationen zu erhalten und auch neue Mutation frühzeitig zu entdecken.

Hierfür stellt das Robert Koch-Institut die Systeme zur bundesweiten molekularen Surveillance bereit. Jedes Labor in Deutschland, das SARS-CoV-2 sequenziert, ist laut der [Verordnung zur molekulargenetischen Surveillance des Coronavirus SARS-CoV-2](#) verpflichtet, dem Robert Koch-Institut die Sequenz- und zugehörige Metadaten zu übermitteln. Technisch erfolgt diese Übermittlung über den [Deutschen Elektronischen Sequenzdaten-Hub](#) (DESH).

Im Projekt "OSEDÄ - Offene Sequenzdaten" verpflichtet sich das RKI, die aufgearbeiteten und qualitätskontrollierten Sequenzdaten zusammen mit einer Auswahl von klinisch-epidemiologischen Daten über die öffentlich zugängliche Repositorien des [European Nucleotide Archive](#) (ENA) und [GISAID](#) für weitere Forschungsvorhaben bereitzustellen.

 Abbildung: Systemaufbau des Deutschen Elektronischen Sequenzdaten-Hub (DESH)

[Kontextmaterialien/2021-01-29_DESH_CorSurV_BAnz_AT_V2.pdf](#)

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-Pandemie bereitgestellt.

Die Datenübermittlung an das RKI erfolgt über das System des [Deutschen Elektronischen Sequenzdaten-Hub](#) (DESH). Teil dieses Systems ist die von der Bundesdruckerei bereitgestellte DESH-Plattform über die Sequenzdaten durch sequenzierenden Labore übermittelt werden können (nur mit einem individuellen

Zertifikat anrufbar). Fragen bezüglich der DESH-Plattform können direkt an das DESH Team unter desh@rki.de gerichtet werden.

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

Übermittlung der Sequenzdaten

Auf der [DESH Projektwebseite](#) des RKI befindet sich eine [Anleitung zur Bereitstellung der Sequenzdaten](#), die Sequenzierenden Laboren im Prozess der Bereitstellung der Metadaten und Sequenzdaten über <https://desh.bdr.de> (nur mit einem individuellen Zertifikat aufrufbar) behilflich ist. Für die sequenzierenden Labore werden bestimmte [Qualitätskriterien](#) für die Sequenzdaten gefordert. Die Einhaltung der Qualitätskriterien wird durch die sequenzierenden Labore sichergestellt. Das RKI hat keine Kenntnis über die zugrundeliegenden Rohdaten (sog. "Reads").

[Kontextmaterialien/2021-02-18_DESH_Anleitung_zur_Bereitstellung_Sequenzdaten.pdf](#)
[Kontextmaterialien/2021-02-08_DESH_Qualitätsvorgaben_für_die_Sequenzdaten.pdf](#)

Veröffentlichung der Sequenzdaten

In der Veröffentlichung von Sequenzdaten in [ENA](#) und [GISAID](#) kommt es durch notwendige Zwischenschritte zu einer zeitlichen Verzögerung der Publikation. Daher stellt das RKI zusätzlich alle über DESH empfangenen Sequenzdaten tagesaktuell zu Verfügung.

⚠ Der Datensatz ist keiner weiteren Qualitätskontrolle durch das RKI durchlaufen. Zu beachten ist, dass Daten in diesem Datensatz zum Beispiel:

- Sequenzdaten von niedriger Qualität enthalten
- unverifizierte Frameshifts vorhersagen
- mehrmals im Datensatz vorhanden sind
- bereits vom sequenzierendem Labor veröffentlicht worden sind

Die hier veröffentlichten Daten können daher nicht ohne weiteres mit dem wöchentlichen [Bericht zu Virusvarianten von SARS-CoV-2 in Deutschland des RKIs](#) verglichen werden. Außerdem können diese Daten ausdrücklich nicht als Grundlage für die Abrechnung der Labore mit der KBV verwendet werden.

Aufbau und Inhalt des Datensatzes

Der Datensatz enthält Daten über SARS-CoV-2-Sequenzen in Deutschland und die in der Datenverarbeitung unterstützenden Kontextmaterialien. Im Datensatz enthalten sind:

- [Sequenzdaten der übermittelten SARS-CoV-2-Genomsequenzen](#)
- [Metadaten zu den SARS-CoV-2-Genomsequenzen](#)
- [Informationen zu den Entwicklungslinien \(PANGOLIN Lineages\) der SARS-CoV-2-Genomsequenzen](#)
- Lizenz mit der Nutzungslizenz des Datensatzes
- Datensatzdokumentation und Kontextmaterialien in deutscher Sprache

- Metadaten Datei zum Import in Zenodo

Formatierung der Sequenzdaten

Die SARS-CoV-2-Sequenzdaten werden als [xz-komprimierte .fasta](#) Datei bereitgestellt. Daraus ergibt sich die Dateiendung .fasta.xz. Die Zeilen werden bei 80 Zeichen umgebrochen. Es werden Linux Zeilenumbrüche verwendet.

- Zeichensatz: UTF-8
- Komprimierung: [.xz](#)
- Enthaltenes Dateiformat: [.fasta](#)
- Zeilenlänge: maximal 80 Zeichen
- Zeilenumbrüche: Linux Zeilenumbrüche

Formatierung der Metadaten

Die Metadaten der Sequenzierung 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: .csv
- .csv-Trennzeichen: Komma ","

Formatierung der Entwicklungslinien

Die Entwicklungslinien der Sequenzierung 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: .csv
- .csv-Trennzeichen: Komma ","

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.

SARS-CoV-2-Sequenzdaten und Metadaten der Sequenzierung

Die SARS-CoV-2-Sequenzdaten werden tagesaktuell im Hauptverzeichnis unter "SARS-CoV-2-Sequenzdaten_Deutschland.fasta.xz" bereitgestellt. Gleiches gilt für zugehörigen Metadaten, die unter

"SARS-CoV-2-Sequenzdaten_Deutschland.csv.xz" und die Entwicklungslinien die unter "SARS-CoV-2-Entwicklungslinien_Deutschland.csv.xz" im Datensatz enthalten sind. **Nicht für alle SARS-CoV-2-Sequenzdaten liegen Entwicklungslinien vor.**

- [SARS-CoV-2-Sequenzdaten_Deutschland.fasta.xz](#)
- [SARS-CoV-2-Sequenzdaten_Deutschland.csv.xz](#)
- [SARS-CoV-2-Entwicklungslinien_Deutschland.csv.xz](#)

Die Daten werden jeden Tag um die verarbeiteten Sequenzdaten des aktuellen Tages erweitert (Kummulation). Dabei werden nach 20:00 eingesendete Sequenzdaten erst am Folgetag verarbeitet. Der Datenstand bildet also immer den Stand des aktuellen Tages um 19:59 ab.

Struktur der Sequenzdaten

Die Sequenzeinträge der bereitgestellten .fasta-Datei beginnen mit einer einzeiligen Beschreibung, der Kopfzeile, auch "Description line" genannt. Auf die Kopfzeile folgt die **Nukleinsäuresequenz** des Sequenzierten SARS-CoV-2-Virus.

Die Kopfzeile wird durch ein ">" markiert, eine Sequenz endet mit dem Ende der Datei oder einem weiteren Sequenzeintrag, beginnen mit einer neuen Kopfzeile.

In den bereitgestellten Sequenzdaten enthält die Kopfzeile die FASTA-ID, die in den Daten der IMS_ID der Probe entspricht. Die IMS_ID erlaubt die Verknüpfung mit den Metadaten. Die Kodierung der Nucleotide der Sequenzdaten folgen dem IUB/IUPAC Standard.

- Kopfzeile: >IMS_ID
- Nukleinsäuresequenz: IUB/IUPAC Standard

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

```
>IMS-101XX-CVDP-XX
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACCAACCAACTTTTCGATCTCTT
GTTCTCTAAACGAACCTTAAAATCTGTGTGGCTGTCTCTCGGCTGCATGCTTAGTGCACT
...
YGACCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAA
GGGAGGACTTGAAAGAGCCACCACATTTTCACCGAGGCN
>IMS-101YY-CVDP-YY
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACCAACTCTCGGCTGCATGCT
GTTCTCTAAACGAACCTTAAAATCTGTGTGGCTGTCTTGAAAGAGCCACCACATTTTCA
...
```

Variablen und Variablenausprägungen Metadaten

In den als .csv bereitgestellten Metadaten enthalten in folgender Tabelle aufgeführte Variablen als Spalten. Zentral für die Verknüpfung der Metadaten mit den Genomsequenzen ist die IMS_ID, die in allen drei Daten enthalten ist.

| Variable | Beschreibung | Value Set |
|----------|--------------|-----------|
|----------|--------------|-----------|

| Variable | Beschreibung | Value Set |
|-------------------|--|------------------------|
| IMS_ID | Ein eindeutiger Identifikator der Sequenzdaten und Metadaten zusammenführt. Dieser Identifikator wird als FASTA ID in den Sequenzdaten genutzt | Text |
| DATE_DRAW | Datum der Probeentnahme | JJJJ-MM-TT |
| SEQ_TYPE | Die verwendete Sequenzierungs-Plattform | ena |
| SEQ_REASON | Grund für die Durchführung der Sequenzierung | rki |
| SAMPLE_TYPE | Art der Probe | snomed |
| OWN_FASTA_ID | Vom Labor genutzte FASTA ID in verschlüsselter Form | Text |
| RECEIVE_DATE | Datum der Datenübermittlung an das RKI | JJJJ-MM-TT |
| PROCESSING_DATE | Verarbeitungsdatum im RKI (Üblicherweise <24 Stunden nach Einsendung durch die Labore) | JJJJ-MM-TT |
| SENDING_LAB_PC | Postleitzahl des primärdiagnostischen Labors | Text |
| SEQUENCING_LAB_PC | Postleitzahl des sequenzierenden Labors | Text |
| GISAID_ACCESSION | Falls bekannt, die GISAID Accession ID der Sequenz | Text |

Weitere Informationen zu den aufgeführten Variablen finden sich in der [Anleitung zur Bereitstellung der Sequenzdaten](#) die auch in [Kontextmaterialien](#) hinterlegt ist.

Variablen und Variablenausprägungen Entwicklungslinien

| Variable | Beschreibung | Value Set |
|--------------------|---|-----------|
| IMS_ID | Ein eindeutiger Identifikator der Sequenzdaten und Metadaten zusammenführt. Dieser Identifikator wird als FASTA ID in den Sequenzdaten genutzt | |
| lineage | Die wahrscheinlichste Abstammung, die einer bestimmten Sequenz zugewiesen wird. | |
| conflict | Wenn eine Sequenz in mehr als eine Kategorie passt, ist der Konfliktwert größer als 0 und spiegelt die Anzahl der Kategorien wider, in die die Sequenz passen könnte. | |
| ambiguity_score | Diese Punktzahl ist eine Funktion der Menge der fehlenden Daten in einer Sequenz. | |
| version | Version der "pango-designation" und "inference engine" | |
| pangolin_version | Version der PANGOLIN Software | |
| pangoLEARN_version | Version des pangoLEARN moduls | |

| Variable | Beschreibung | Value Set |
|---------------|---|-----------|
| pango_version | Version der "pango-designation" auf dem zu Zuordnungen basieren | |
| status | Zeigt an, ob die Sequenz die QC-Schwellenwerte für die Mindestlänge und den maximalen N-Gehalt überschritten hat. | |
| note | Bei Konflikten a werden in diesem Feld die alternativen Entwicklungslinien Zuordnungen ausgegeben. | |

Die bereitgestellten Informationen zu den Entwicklungslinien entsprechen dem aktuellen [PANGOLIN Lineage Format](#). Nur die Spalte "Taxon" wurde zur einfacherer Nachnutzung in IMS_ID umbenannt. Zentral für die Verknüpfung der Entwicklungslinien mit den restlichen Daten ist die IMS_ID, die in allen drei Daten enthalten ist. [PANGOLIN Lineage Format](#) ist bei Widersprüchen autoritativ.

Hinweise zur Nachnutzung der Daten

⚠ Der Datensatz ist keiner weitere Qualitätskontrolle durch das RKI durchlaufen. Zu beachten ist, dass Daten in diesem Datensatz zum Beispiel:

- Sequenzdaten von niedriger Qualität enthalten
- unverifizierte Frameshifts vorhersagen
- mehrmals im Datensatz vorhanden sind
- bereits vom sequenzierendem Labor veröffentlicht worden sind

Die hier veröffentlichten Daten können daher nicht ohne weiteres mit dem wöchentlichen [Bericht zu Virusvarianten von SARS-CoV-2 in Deutschland des RKIs](#) verglichen werden. Außerdem können diese Daten ausdrücklich nicht als Grundlage für die Abrechnung der Labore mit der KBV verwendet werden.

Offene Forschungsdaten des RKI werden auf [GitHub.com](#), [Zenodo.org](#) und [Edoc.rki.de](#) bereitgestellt:

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

Metadaten der Publikation

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 Metadaten-Ordner 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](#)

Lizenz

Der Datensatz "SARS-CoV-2-Sequenzdaten_aus_Deutschland" ist lizenziert unter der [Creative Commons Namensnennung 4.0 International Public License | CC-BY 4.0 International](#)

Die im Datensatz bereitgestellten Daten sind, unter Bedingung der Namensnennung des Robert Koch-Instituts als Quelle, frei verfügbar. Das bedeutet, jede_r hat das Recht die Daten zu verarbeiten und zu verändern, Derivate des Datensatzes zu erstellen und sie für kommerzielle und nicht kommerzielle Zwecke zu nutzen. Weitere Informationen zur Lizenz finden sich in der [LICENSE](#) bzw. [LIZENZ](#) Datei des Datensatzes.