Datensatzdokumentation

SARS-CoV-2-Sequenzdaten aus Deutschland

Robert Koch-Institut | RKI
Nordufer 20
13353 Berlin

Zitieren

Robert Koch-Institut (**2024**): *SARS-CoV-2-Sequenzdaten aus Deutschland*, Berlin: Zenodo. DOI: 10.5281/zenodo.13788841

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 tagesaktuell 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 SEQUENCE.ID, was eine einfache Verknüpfung mit den bereitgestellten Metadaten erlaubt.

- Kopfzeile: "><IGS_ID> version=<version>"
- Nukleinsäuresequenz: IUB/IUPAC Standard

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

```
>IGS-101XX-CVDP-XX version=1
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACCAACCAACTTTCGATCTCTT...
>IGS-101YY-CVDP-YY version=0
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNACCAACTCTCGGCTGCATGCT...
```

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

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

Variable	Typ	Ausprägungen/Beispiel	Beschreibung
igs_id	String	`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	Datetime	`YYYY-MM-DDThh:mm:ss`	Datum der Probeentnahme im ISO 8601 Format ohne Zeitzone
sequencing_platform	String	siehe ena	Die verwendete Sequenzierungs-Plattform auf Basis der von ENA zugelassenen Ontologie
sequencing_reason	String	`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		DEMIS Vokabular
lab_sequence_id	String		Vom Labor genutzte FASTA ID in verschlüsselter Form
date_of_submission	Datetime	`YYYY-MM-DDThh:mm:ss`	Datum des Eingangs des Genoms am RKI im ISO 8601 Format ohne Zeitzone
version	Integer	`1`	Version der Sequenz startend mit 0
prime_diagnostic_lab.demis_lab_id	String	`DEMIS-10099`	Identifikationsnummer des primärdiagnostischen Labors
prime_diagnostic_lab.postal_code	String	`50858`	Postleitzahl des primärdiagnostischen Labors
sequencing_lab.demis_lab_id	String	`DEMIS-10099`	Identifikationsnummer des sequenzierenden Labors
sequencing_lab.postal_code	String	`50858`	Postleitzahl des sequenzierenden Labors
		`[{'method': 'PANGOLIN_LATEST', 'classification_version': 'PUSHER-v1.28.1', 'tool_version': '4.3',`	

lineages	JSON Blob	`'lineage': 'BA.2', '@qc_notes': 'Ambiguous_content:0.02', '@is_designated': False, '@qc_status': 'pass', '@conflict': 0.0, '@note': 'Usher placements: BA.2(1/1)'}]`	Pangolin Zuordnung

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 der 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": "Created",
    "description": "Date when the Dataset was created"
  }
],
```

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

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 Person 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.