

 An official website of the United States government



Log in

SRA

Search

Advanced

Getting Started

Submission Quick Start

Search and Download

SRA in the Cloud

Archive Documentation

NLM GenBank and SRA Data Processing

- Submitting Data
- Processing Submissions
- Public Release and Data Accessibility
- Requesting Data Status Changes
- Data Status Definitions

The National Library of Medicine's (NLM's) National Center for Biotechnology Information (NCBI) manages the GenBank and Sequence Read Archive (SRA) databases of genetic sequence information. On this page and in the accompanying diagram you will find information about:

- how sequence data are submitted, processed, and made available to the public
- responsibilities of the sequence data submitter and NCBI
- key definitions of data status

Getting Started

- Getting Started

Submitting to the S

- Submission Quick

Search and Downlo

- Download Guide

SRA in the Cloud

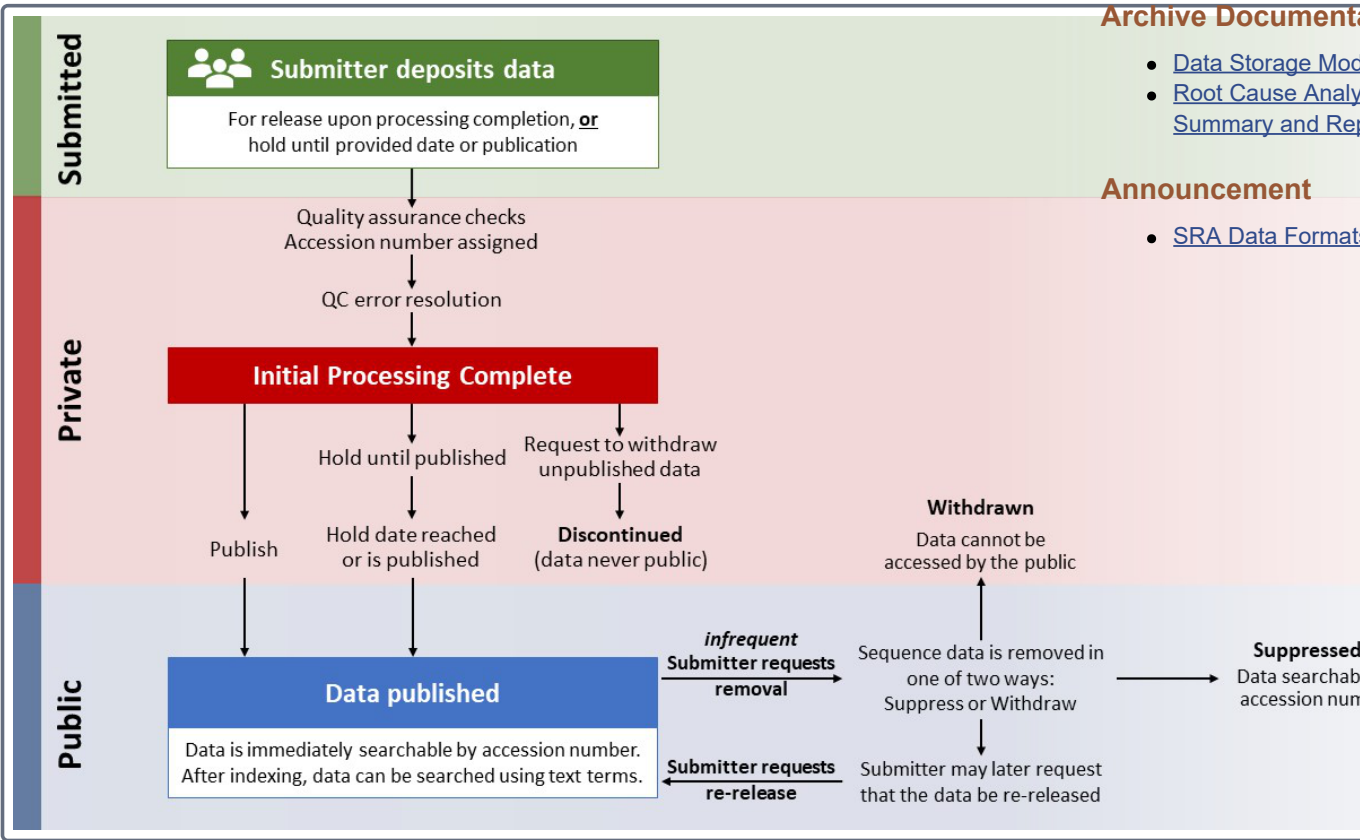
- SRA in the Cloud

Archive Documenti

- Data Storage Moc
- Root Cause Analy
- Summary and Rej

Announcement

- SRA Data Format



Submitting Data

Sequence data submissions may be initiated by a variety of individuals including researchers, staff affiliated with a public health laboratory, sequencing center, data analysis center, and personnel associated with a data coordination center. Submitters deposit sequence data and metadata into GenBank or SRA for many reasons including to:

- comply with data sharing policies established by government authorities, publishers, or funders (e.g., the [NIH Scientific Data Sharing Policies](#))
- support research community established principles such as the [Bermuda Principles](#), or [FAIR](#) (findable, accessible, interoperable, and reusable) data principles
- support open science
- serve the public good

Submitters are responsible for formatting their sequence data for submission, meeting NCBI submission standards, ensuring they have authority to submit the sequence data and using NCBI services to submit the sequence data and metadata. At the time of submission, the submitter may specify a desired public release date for the sequence data (e.g., to align with the anticipated date of publication of a journal article). Submitters may request that private data be made publicly available prior to the scheduled release date or request that the release date be extended (e.g., to align with the anticipated date of publication of a journal article).

Processing Submissions

NCBI is responsible for processing submitted sequence data. Processing includes performing automated and manual checks to ensure data integrity, quality, and assigning accession numbers to submitted sequence data. NCBI holds the sequence data in a [private](#) status during processing, and prior to public release.

In general, NCBI processes submissions in the order received. However, NCBI may prioritize processing of submissions related to a pandemic or public health emergency. Upon the submitter's request, NCBI may also prioritize processing of submissions associated with an upcoming publication release.

NCBI may halt the processing of submitted sequence data at any time prior to public release upon the request of the submitter. In this case, NCBI does not release the sequence data and retains the data in a [discontinued](#) status.

NCBI may determine, based on quality control checks conducted as part of the processing, that the data are not of sufficient quality for public release. In such cases, NCBI will halt data processing and notify the submitter with an explanation. NCBI retains the sequence data in a [discontinued](#) status.

Public Release and Data Accessibility

NCBI is responsible for making sequence data publicly accessible by putting them in a [public](#) status.

NCBI generally makes sequence data publicly accessible upon completion of processing or on the submitter-specified release date. NCBI also makes sequence data that has completed processing publicly accessible prior to the requested release date if NCBI becomes aware that the data or accession numbers have been published in another database, web resource, or publication. NCBI notifies the submitter when sequence data are publicly released.

NCBI posts sequence data to several storage locations and disseminates the data for public access on websites, cloud platforms, ftp sites, tools, and application programming interfaces (APIs). For example, NCBI makes sequence data accessible for accession and text-based search on the NCBI website and for sequence similarity searching using NCBI's [Basic Local Alignment Search Tool](#) (BLAST). GenBank sequence records may also be downloaded from the FTP site or accessed using NCBI's [E-utilities](#) API. SRA sequence records are available using the [SRA Toolkit](#) API or on [Amazon Web Services \(AWS\)](#) and [Google Cloud Platform \(GCP\)](#) clouds. SRA availability on cloud platforms enables rapid access to large datasets.

Sequence data may become publicly available at different times across these NCBI storage locations, websites, APIs, and analysis tools as the newly released data propagates across the system. Upon release, publicly accessible sequence data are searchable by accession number in website interfaces.

NCBI also indexes the data to support text-based searches (e.g., by organism name) in websites and APIs.

In addition, NCBI exchanges sequence data with members of the [International Nucleotide Sequence Database Collaboration \(INSDC\)](#), namely the European Bioinformatics Institute (EBI) at the European Molecular Biology Laboratory (EMBL) and the National Institute of Genetics (NIG) at the Research Organization of Information and Systems in Japan. Such exchanges enable that all INSDC sites provide access to a comprehensive collection of publicly accessible sequence data (INSDC members do not exchange sensitive controlled access human sequence data). As a result, NCBI provides public access to sequence data and metadata that are submitted to and processed by other INSDC organizations, and other INSDC organizations provide public access to sequence data and metadata that are submitted to NCBI.

NCBI also indexes the data to support text-based searches (e.g., by organism name) in websites and APIs.

Public sequence data made available by NCBI may be retrieved and redistributed by other users and presented in other websites, databases, tools, publications, curricula, conference proceedings, or other venues that are not managed by NCBI. These other resources present a snapshot from the time of retrieval and may not contain the most recent updates or changes to status.

Requesting Data Status Changes

Submitters to GenBank and SRA are generally responsible for requesting changes to the status of their sequence data. NCBI does not directly manage the status of sequence data that are submitted to other INSDC members, and submitters to those databases must work directly with those INSDC members to change the status of data.

In certain circumstances, submitters may request that their data be removed after public release. NCBI is responsible for verifying that the request is valid (e.g., it originates from the submitter), determining whether the request meets the criteria for removal described herein, and determining the appropriate removal method.

Sequence data may be removed from public access in one of two ways: *suppress* or *withdraw*.

- Data are [suppressed](#) when the submitter has concerns related to issues such as data quality or changes in the scope or timing of associated publications, and there is a need to maintain data availability via accession number to preserve the integrity of the published scientific record (see examples below). Suppressed data remain publicly accessible by accession number and are removed from indexing for text searches and API or tool retrievals (e.g., BLAST).
- Data are [withdrawn](#) when there are concerns about possible harms resulting from public availability of the data such as those related to national security, privacy, or lack of proper informed consent (see examples below). Withdrawn data are not publicly accessible, even by accession number.

When data are [suppressed](#) or [withdrawn](#), NCBI updates the status of the data and retains the data for archival purposes and to enable possible future re-release. The data status change may take effect at different times across the range of NCBI storage locations, websites, APIs, and analysis tools, including across other INSDC members' resources.

Because public sequence data made available by NCBI may be retrieved and redistributed by other users and presented in other websites, databases, tools, publications, curricula, conference proceedings, or other venues that are not managed by NCBI, data that are suppressed or withdrawn may remain available through other sources that are not managed by NCBI.

Data submitters may request that suppressed data be re-released upon publication of the data or once they confirm or update questionable data.

Examples of valid reasons for a submitter to request removing sequence data include:

- Suppression of public data:
 - Data reported as being from a single organism are discovered, after public release,

- to be contaminated with sequences from another organism.
- o The taxonomic identity of the sequenced organism is determined, after public release, to be unconfirmed. For example, this may occur if there are few or no other sequences available for the organism to carry out an initial validation, and the initial designation is later determined to be incorrect and cannot be updated.
- o Data are found, after public release, to contain errors that cannot be corrected, making the data unsuitable for reuse in future analysis. Errors identified by the submitter may include incorrect assembly, annotation, metadata, sample mix-up, contamination, or low-quality sequence (e.g., the submitted sequence lacks sufficient supporting evidence).
- o Data are later determined to be an unallowable submission type or an erroneous submission.
 - For example, GenBank does not allow submission of another submitter's sequence data without that submitter's collaboration or permission. Absent that, such data can be submitted as a Third-Party Annotation (TPA) if the submitter meets [TPA criteria](#).
 - Submitters may mistakenly submit sequence data (e.g., when carrying out a trial run of the submission process).
- o Data are released upon reaching the submitter-provided public release date, and before the publication or analysis referencing the data is complete.
- o The submitter notifies NCBI of duplicate data in GenBank or SRA (e.g., due to redundant submissions, or an update was provided as a new submission instead of as an update). When possible, the original accession number is added to the newer data as a 'secondary' accession number which results in retrieval of the new accession number for searches for the original accession number. If this tracking is not possible, typically because the submitter does not provide precise mapping from the original accession to the new accession, then the original submitted data is suppressed.
- Withdrawal of public data:
 - o The submitter determines, after public availability, that they did not have proper informed consent to publicly release protected human data.
 - o NCBI is notified (e.g., by the principal investigator, laboratory manager, institution, or journal) that the data should be retracted based on malfeasance or fraud. NCBI will work with the complainant and/or institution to verify the claim.
 - o NCBI is notified that the sequence data was uploaded by a person who was not authorized to submit the data. NCBI will work with the principal investigator, laboratory manager, and/or institution to verify the claim.
 - o NCBI erroneously released sequence data to the public during data processing.

Data Status Definitions

Data submitted to GenBank and SRA are assigned one of the following statuses:

Discontinued: The submitter has elected to halt the submission process for private data or NCBI has detected quality problems prior to public release. NCBI generally keeps the data temporarily to support submitters should they later decide to release the data, but NCBI may not retain data indefinitely from discontinued submissions.

Private: Private data are not available publicly through any means. Data have been submitted and are undergoing processing and/or are scheduled for release at a future date. Private data are pre-decisional and confidential and may or may not become publicly released.

Public: Public data are fully accessible for search and distribution. NCBI has completed processing and publishing the data.

Suppressed: Suppressed data are data that were previously public, have been removed from the NCBI text-based search and comparative analysis results, and may be accessed only by accession number. Suppressed data often have a future date when they will return to public status.

Withdrawn: Withdrawn data are data that were previously public, have been removed from the NCBI text-based search and comparative analysis results, and cannot be accessed by the public even by accession number. NCBI retains the data to preserve the integrity of the scientific record and for disaster recovery with limited exceptions (e.g., national security).

Last updated on 2023-01-04

You are here: [NCBI](#)

Support Center

FOLLOW NCBI

Twitter

Facebook

LinkedIn

```
.cls-
11,
.cls-12
```

```
{ fill:
#737373;
} .cls-
11 {
closure
rule:
evenodd;
}
```

GitHub Last updated: 2023-02-10T12:29:47Z

.cls-
1{fill:#7
NCBI
Insights
Blog

Connect with NLM

National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894

Web Policies
FOIA
HHS Vulnerability Disclosure

Help
Accessibility
Careers

NLM | NIH | HHS | USA.gov