Sagol School
of Neuroscience
Tel Aviv University

How to upload your sequencing data using SRA:

Overall introduction

Below, we provide a step-by-step guide for uploading sequencing data using NCBI’s SRA platform. Sequencing data can originate from animals, plants, bacteria, viruses, or environmental samples. Compared to repositories like dbGaP, human data in this context must not pose privacy risks.

There is no strict total dataset size limit for SRA submissions, as they can accommodate extremely large datasets — many laboratories even submit hundreds of terabytes. However, it is recommended to split very large individual files into smaller pieces, ideally below 10–20 GB each. This practice helps prevent transfer errors and makes files more manageable for both storage and download.

An SRA submission typically involves three main components:

- **BioProject**: Describes the goal and scope of your study and includes information about the project and any relevant grants.
- **BioSample**: Provides detailed information about the biological specimen from which the sequence is derived.
- **SRA (metadata and sequence data)**: The metadata describes the sequencing platform and library construction, while the actual sequence data contains the raw reads.

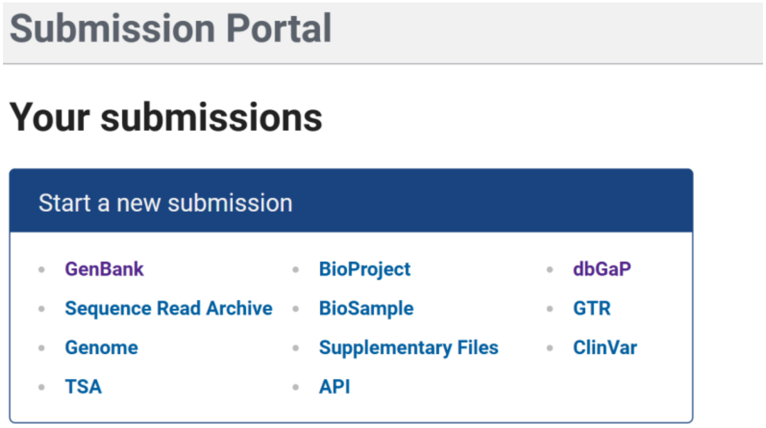
Repository: SRA (Sequence Read Archive)

Make sure you have an NCBI account and have set up your profile information. Only then will you be able to upload to this repository.

Submit your study

Link to explanatory YouTube guide:
https://www.youtube.com/watch?v=PTg9Ru68fc0&ab_channel=NationalLibraryofMedicine

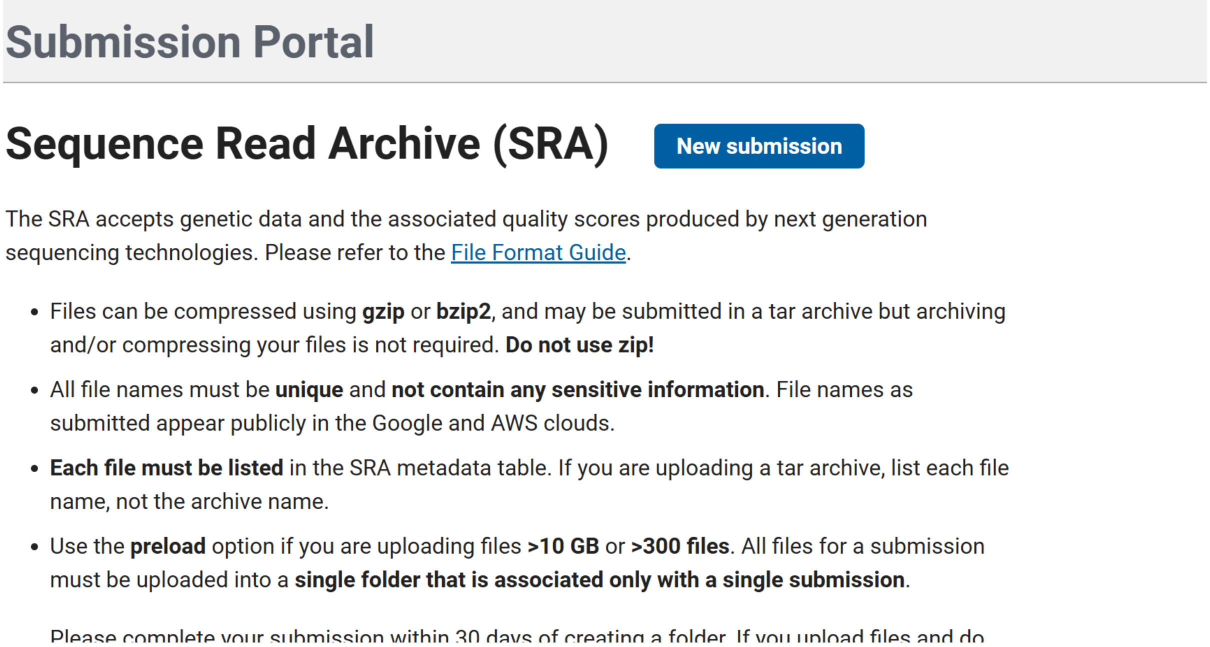
First click this link, and choose Sequence Read Archive:



Submission Portal

Your submissions

Then, press new submission to start your submission:



Submission Portal

Sequence Read Archive (SRA)

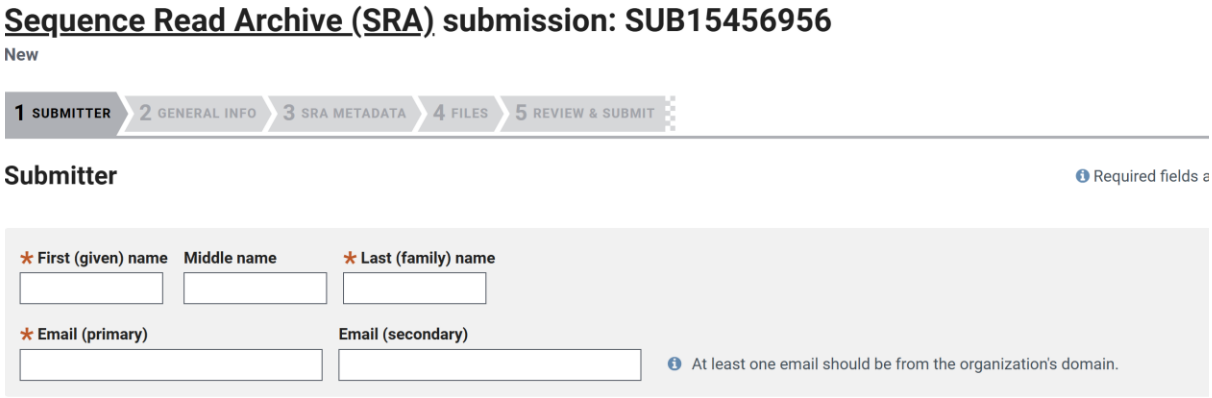
The SRA accepts genetic data and the associated quality scores produced by next generation sequencing technologies. Please refer to the File Format Guide.

- Files can be compressed using **gzip** or **bzip2**, and may be submitted in a tar archive but archiving and/or compressing your files is not required. **Do not use zip!**
- All file names must be **unique** and **not contain any sensitive information**. File names as submitted appear publicly in the Google and AWS clouds.
- **Each file must be listed** in the SRA metadata table. If you are uploading a tar archive, list each file name, not the archive name.
- Use the **preload** option if you are uploading files **>10 GB** or **>300 files**. All files for a submission must be uploaded into a **single folder that is associated only with a single submission**.

Please complete your submission within 30 days of creating a folder. If you upload files and do

Submission steps:

1. First, you will arrive at the **Submitter** page, where you provide information about the person submitting the data.



Sequence Read Archive (SRA) submission: SUB15456956

New

Submitter

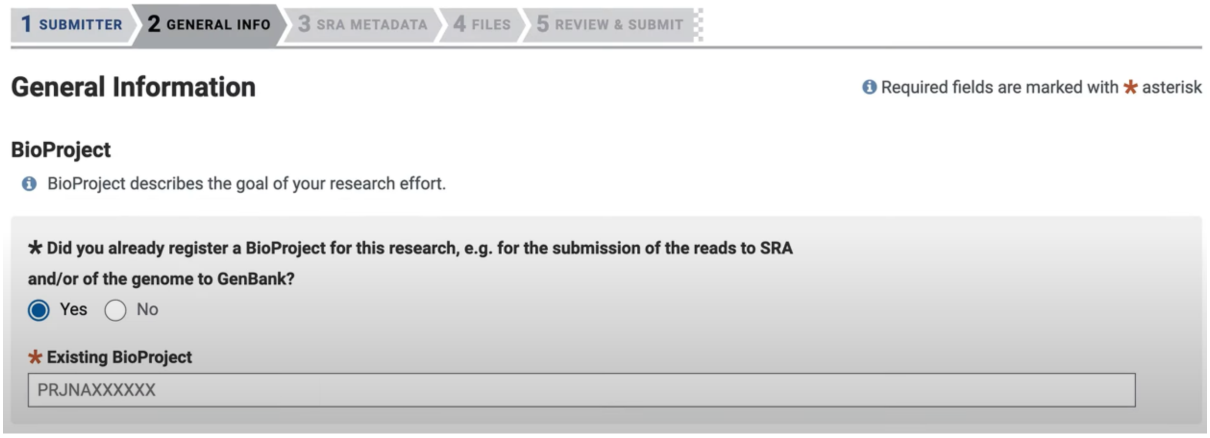
Required fields a

* First (given) name | Middle name | * Last (family) name

* Email (primary) | Email (secondary)

At least one email should be from the organization's domain.

2. After filling in the personal information and clicking **Continue**, you will be prompted to provide the general information about the study.



General Information

Required fields are marked with * asterisk

BioProject

BioProject describes the goal of your research effort.

* Did you already register a BioProject for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?

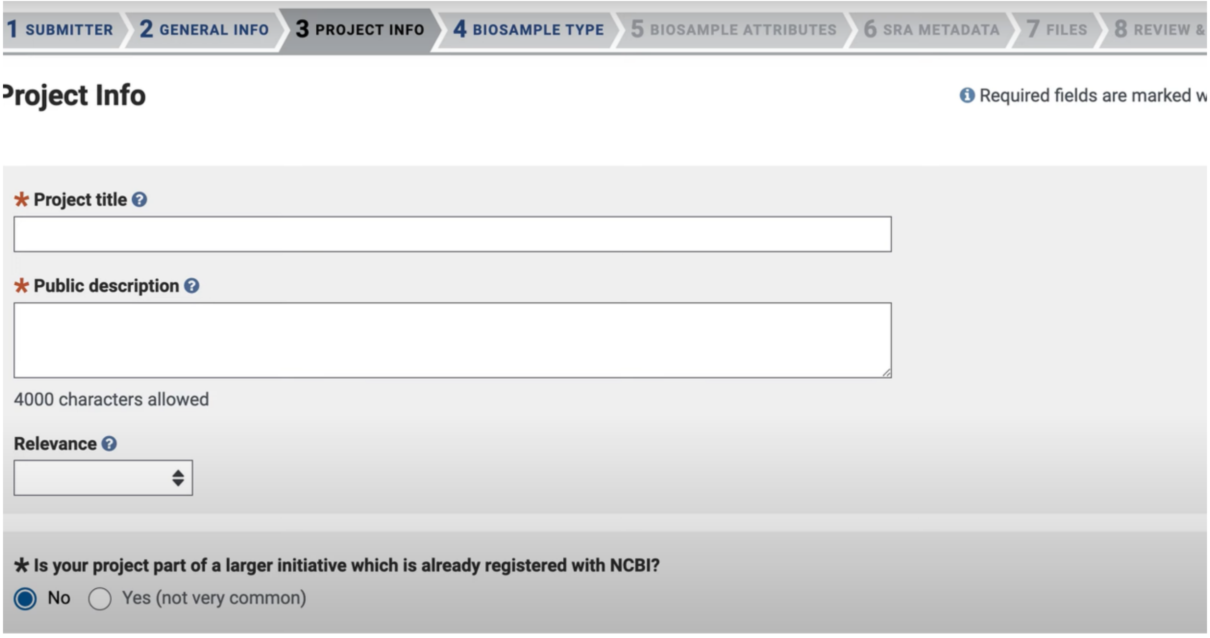
● Yes ○ No

* Existing BioProject

PRJNAXXXXXX

Answer the questions that appear based on your specific intentions, and then click **Continue**. Note that for the question about an existing BioProject, you should select **"No"** if you have not already created one.

3. If you have not already uploaded a BioProject and BioSample beforehand, several additional steps will appear in the top navigation bar.

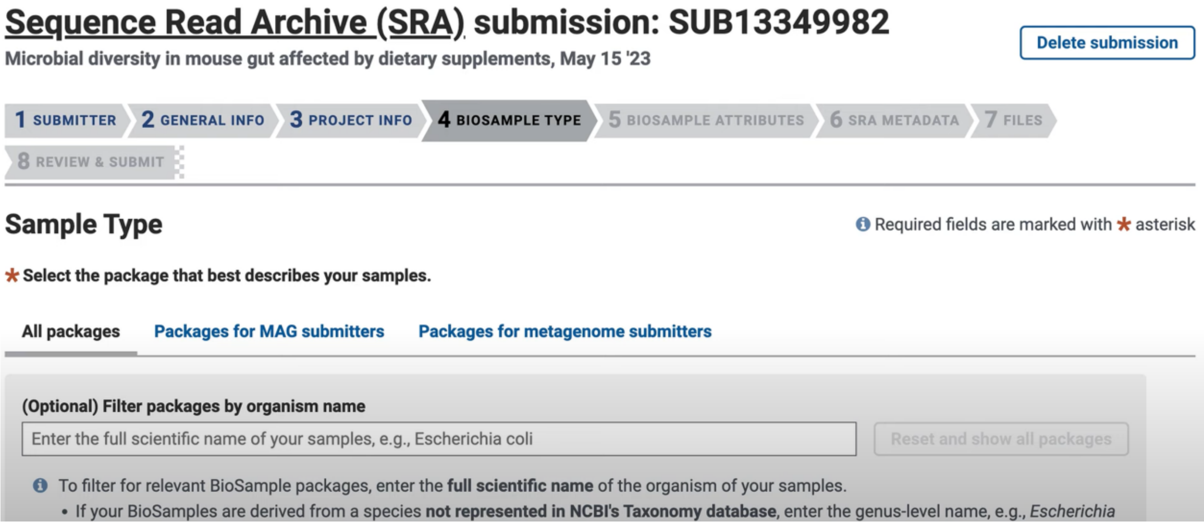


* Is your project part of a larger initiative which is already registered with NCBI?

● No ○ Yes (not very common)

First, fill in the general information about your project, such as the title and description. This is the **BioProject** section of the submission, where you provide details about the project itself as well as any related grants.

4. After clicking **Continue**, you will proceed to the **BioSample** section.



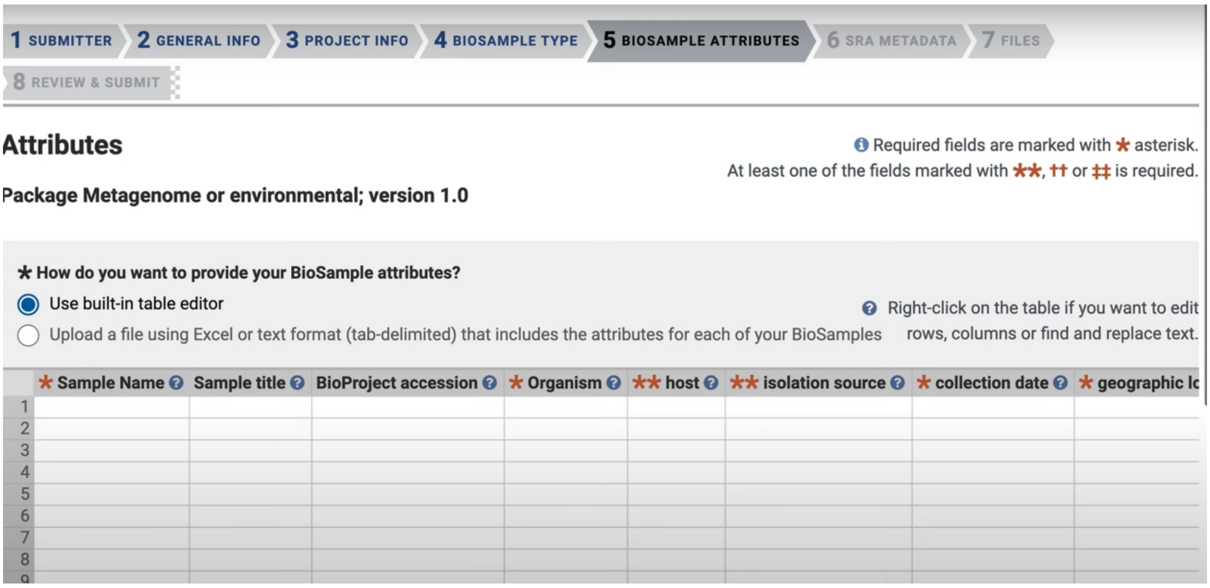
* Select the package that best describes your samples.

To filter for relevant BioSample packages, enter the **full scientific name** of the organism of your samples.

- If your BioSamples are derived from a species **not represented in NCBI's Taxonomy database**, enter the genus-level name, e.g., *Escherichia*

First, you will select the relevant package. A package determines how the BioSample Excel file is formatted by specifying which columns and metadata fields are required for your data.

5. Now, you can either manually enter the relevant information or upload an Excel file. Make sure to fill in the information according to the specified headers, and use the **‘?’** symbol to see detailed guidance for each column. The portal will allow you to continue only if all information has been entered correctly; otherwise, it will display an appropriate error message.



Attributes

Required fields are marked with * asterisk.
At least one of the fields marked with **, †† or ‡‡ is required.

Package Metagenome or environmental; version 1.0

*** How do you want to provide your BioSample attributes?**

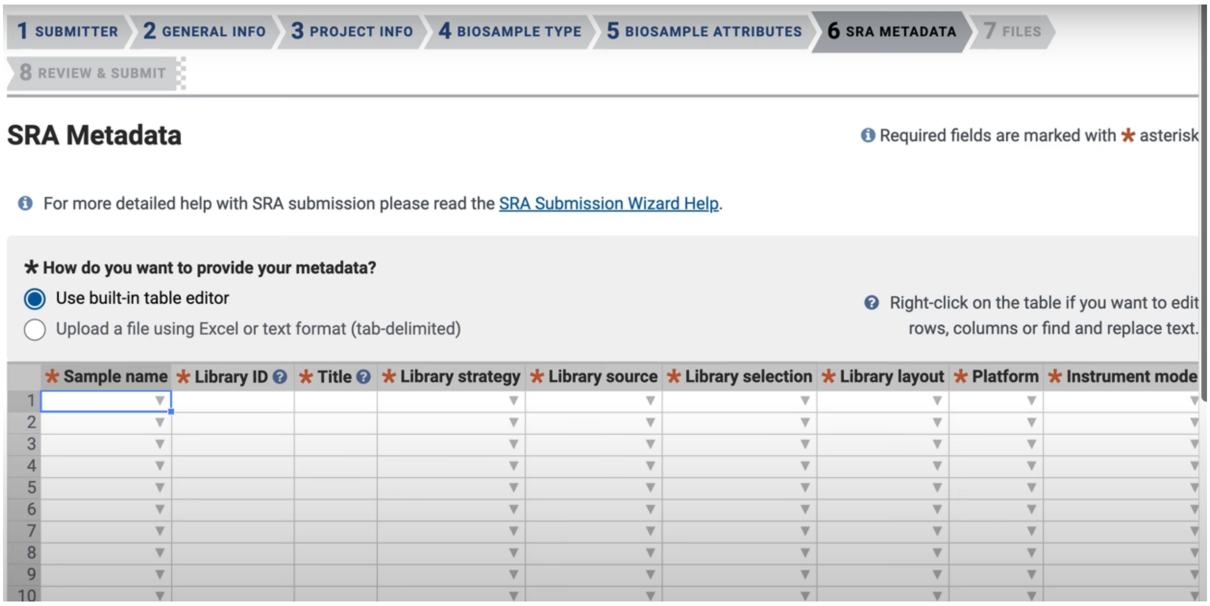
● Use built-in table editor

○ Upload a file using Excel or text format (tab-delimited) that includes the attributes for each of your BioSamples

Right-click on the table if you want to edit rows, columns or find and replace text.

	* Sample Name	Sample title	BioProject accession	* Organism	** host	** isolation source	* collection date	* geographic lc
1								
2								
3								
4								
5								
6								
7								
8								

6. In the next step, you will upload the relevant sequencing metadata. As before, you can either upload an Excel file or manually enter the required information.



SRA Metadata

Required fields are marked with * asterisk

For more detailed help with SRA submission please read the SRA Submission Wizard Help.

*** How do you want to provide your metadata?**

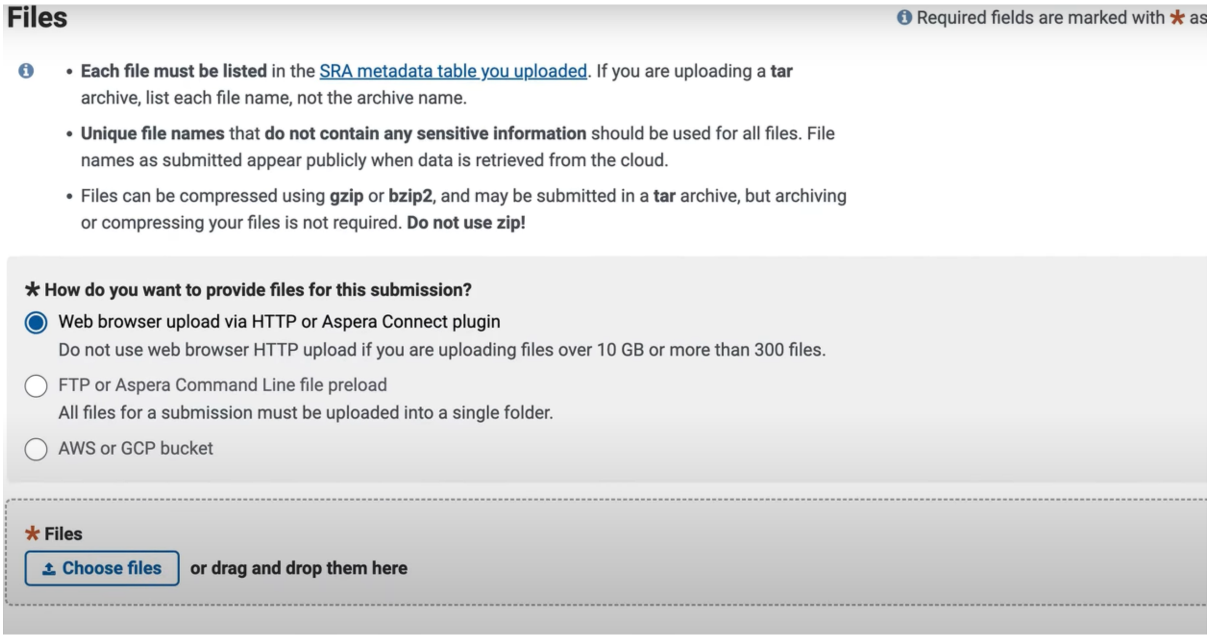
● Use built-in table editor

○ Upload a file using Excel or text format (tab-delimited)

Right-click on the table if you want to edit rows, columns or find and replace text.

	* Sample name	* Library ID	* Title	* Library strategy	* Library source	* Library selection	* Library layout	* Platform	* Instrument mode
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									

7. Finally, you can upload your sequencing files containing the actual data. If you choose the web browser option, you will be able to drag and drop your files directly into the upload section.



Files

Required fields are marked with * as

- **Each file must be listed** in the SRA metadata table you uploaded. If you are uploading a **tar** archive, list each file name, not the archive name.
- **Unique file names** that **do not contain any sensitive information** should be used for all files. File names as submitted appear publicly when data is retrieved from the cloud.
- Files can be compressed using **gzip** or **bzip2**, and may be submitted in a **tar** archive, but archiving or compressing your files is not required. **Do not use zip!**

*** How do you want to provide files for this submission?**

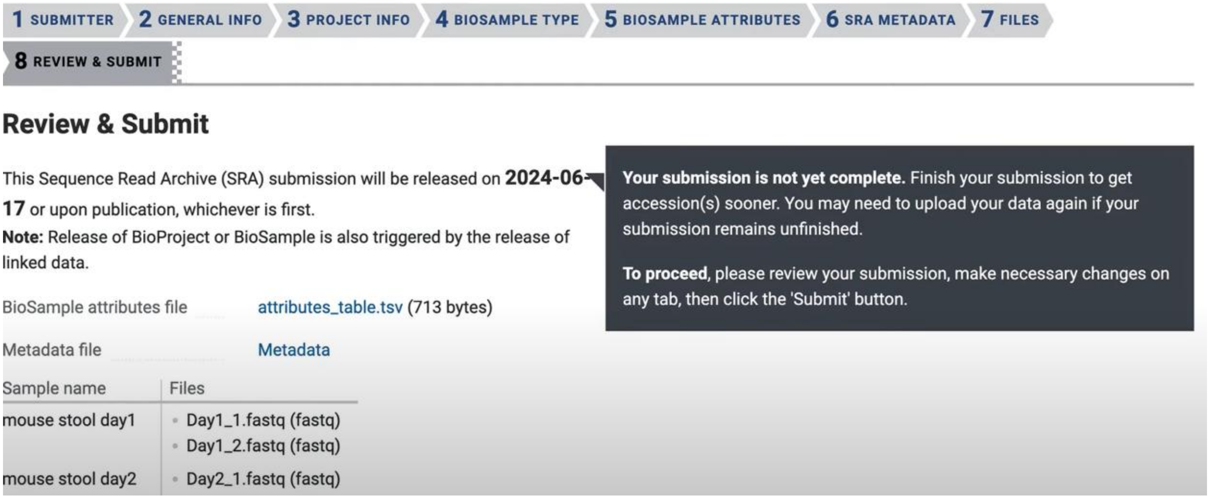
● Web browser upload via HTTP or Aspera Connect plugin
Do not use web browser HTTP upload if you are uploading files over 10 GB or more than 300 files.

○ FTP or Aspera Command Line file preload
All files for a submission must be uploaded into a single folder.

○ AWS or GCP bucket

*** Files**

8. In the final section, you can review all the information you have provided to ensure it is correct. If everything looks good, click **Submit** to complete your submission.



Review & Submit

This Sequence Read Archive (SRA) submission will be released on **2024-06-17** or upon publication, whichever is first.
Note: Release of BioProject or BioSample is also triggered by the release of linked data.

Your submission is not yet complete. Finish your submission to get accession(s) sooner. You may need to upload your data again if your submission remains unfinished.

To proceed, please review your submission, make necessary changes on any tab, then click the 'Submit' button.

BioSample attributes file: attributes_table.tsv (713 bytes)

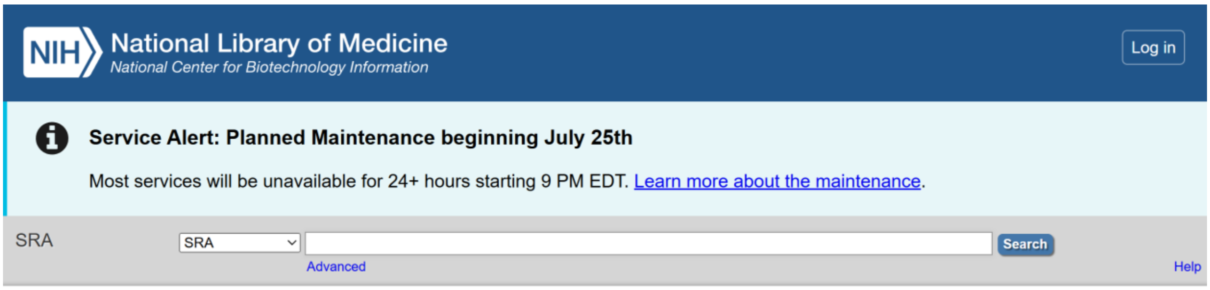
Metadata file: Metadata

Sample name	Files
mouse stool day1	Day1_1.fastq (fastq) Day1_2.fastq (fastq)
mouse stool day2	Day2_1.fastq (fastq)

Once finished, your submission will appear under my submission.

Downloading and viewing SRA datasets:

1. Visit the SRA webpage and search for the data you are interested in:

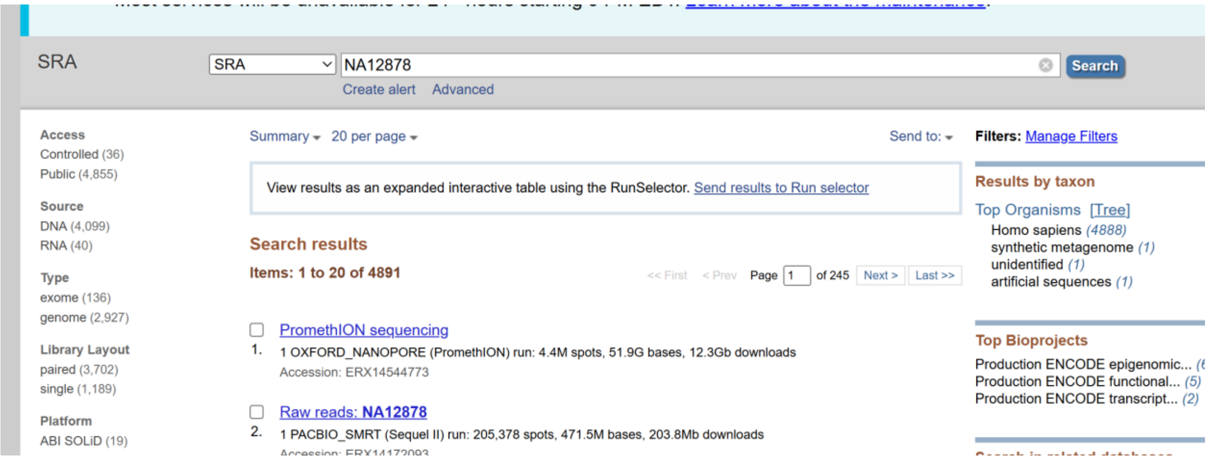


Service Alert: Planned Maintenance beginning July 25th

Most services will be unavailable for 24+ hours starting 9 PM EDT. Learn more about the maintenance.

SRA: SRA

2. You will see several options. Note that you can use the filters on the left to refine your results. For example, select **"Public"** to view only publicly available data.



SRA: SRA NA12878

Access
Controlled (36)
Public (4,855)

Source
DNA (4,099)
RNA (40)

Type
exome (136)
genome (2,927)

Library Layout
paired (3,702)
single (1,189)

Platform
ABI SOLiD (19)

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results

Items: 1 to 20 of 4891

Page 1 of 245

☐ PromethION sequencing
1. 1 OXFORD_NANOPORE (PromethION) run: 4.4M spots, 51.9G bases, 12.3Gb downloads
Accession: ERX14544773

☐ Raw reads: **NA12878**
2. 1 PACBIO_SMRT (Sequel II) run: 205,378 spots, 471.5M bases, 203.8Mb downloads

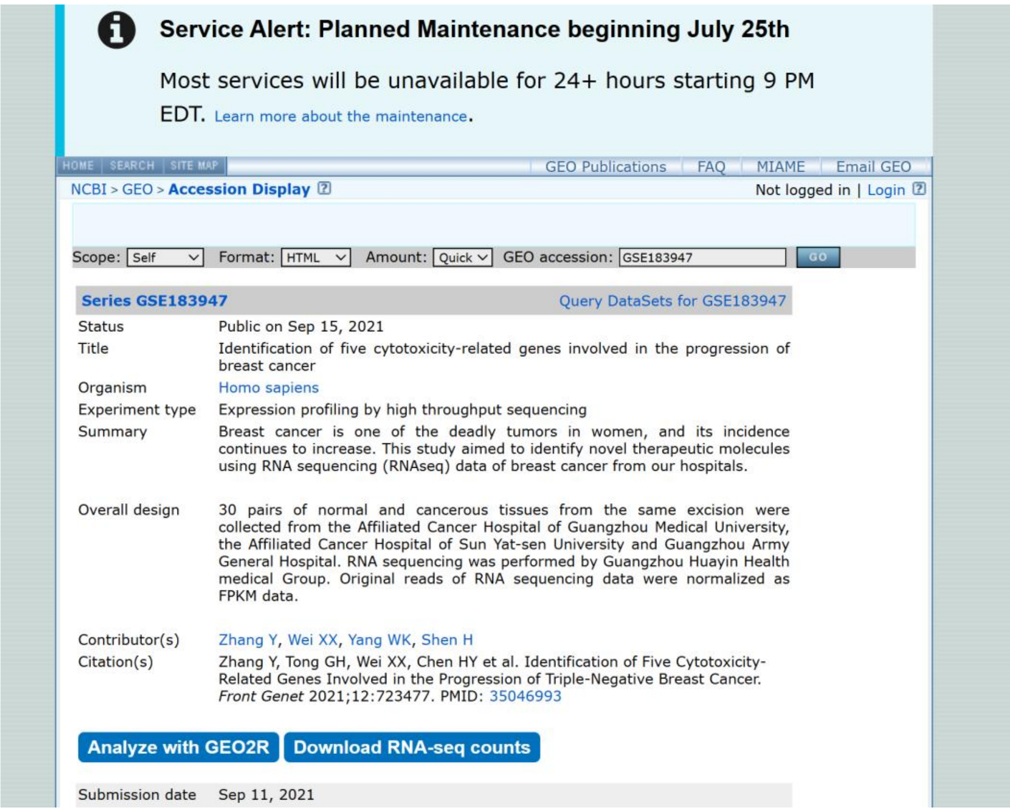
Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(4888)*
synthetic metagenome *(1)*
unidentified *(1)*
artificial sequences *(1)*

Top Bioprojects
Production ENCODE epigenomic...
Production ENCODE functional... *(5)*
Production ENCODE transcript... *(2)*

3. Once you have chosen a dataset (here is a link for the example data), you will see several available options.



Service Alert: Planned Maintenance beginning July 25th

Most services will be unavailable for 24+ hours starting 9 PM EDT. Learn more about the maintenance.

NCBI > GEO > **Accession Display**

Scope: Self Format: HTML Amount: Quick GEO accession: GSE183947

Series GSE183947 Query DataSets for GSE183947

Status	Public on Sep 15, 2021
Title	Identification of five cytotoxicity-related genes involved in the progression of breast cancer
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	Breast cancer is one of the deadly tumors in women, and its incidence continues to increase. This study aimed to identify novel therapeutic molecules using RNA sequencing (RNAseq) data of breast cancer from our hospitals.
Overall design	30 pairs of normal and cancerous tissues from the same excision were collected from the Affiliated Cancer Hospital of Guangzhou Medical University, the Affiliated Cancer Hospital of Sun Yat-sen University and Guangzhou Army General Hospital. RNA sequencing was performed by Guangzhou Huayin Health medical Group. Original reads of RNA sequencing data were normalized as FPKM data.
Contributor(s)	Zhang Y, Wei XX, Yang WK, Shen H
Citation(s)	Zhang Y, Tong GH, Wei XX, Chen HY et al. Identification of Five Cytotoxicity-Related Genes Involved in the Progression of Triple-Negative Breast Cancer. *Front Genet* 2021;12:723477. PMID: 35046993
Submission date	Sep 11, 2021

For example, if you choose **"Analyze with GEO2R,"** you can use an online analysis tool without needing to download the data itself.
If you wish to download the data, please watch the following video for a detailed guide on how to download the entire dataset.