SRA Knowledge Base

Last Updated: March 18, 2014

National Center for Biotechnology Information (US)
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

NLM Citation: SRA Knowledge Base [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2011-.

The SRA Knowledge Base provides the answers to common questions asked by Sequence Read Archive users. It is arranged as a topical reference for key word searching.

Table of Contents

Downloading and Accessing Data

Sequence Read Archive Submissions Staff

Created: August 22, 2011; Updated: March 18, 2014.

This section of the Sequence Read Archive (SRA) Knowledge Base contains:

- Instructions for downloading data from the SRA website
- Instructions for downloading data using command line utilities
- Instructions for downloading SRA data from dbGaP**
- Information for dealing with problems that may arise during the download process
- Information on converting .sra files into other formats and direct downloads of fasta and fastq format data

** **Please Note**: The information provided in this section regarding the download of individual-level data from dbGaP assumes that you have completed the dbGaP data access request process, and have been granted authorized access to individual level data.

If you do not have authorized access to dbGaP individual-level data, you will not be able to access individual-level data using the instructions in this section, since you will not have access to the appropriate password-protected dbGaP sites mentioned. For information on how to apply for authorized access, please see the "Data Access Request" section of the dbGaP FAQ Archive.

To begin searching this section of the Sequence Read Archive Quick Start, you can:

- **Enter** your **search word(s)** text in the text box **at the top of the page** and click on the "Go" button

  OR
- **Clickon** any of the "GenBank Submission Resources Quick Start" **sub-categories listed** in the "Contents" section **below** to go to the sub-category of your choice.

Downloading data from the SRA website

General Information

Overview

SRA data may be downloaded using HTTP, FTP, or the Aspera Connect plugin. The easiest way to acquire SRA data is to (1) search for data sets of interest, (2) click on an Experiment of interest from the results to learn more, (3) click on one of the "SRR" accessions that comprise the experiment (this will lead you to the SRA Run Browser), and (4) click on the "Download" tab for this data set. You will be presented with the option to download the individual data set ("Run"), all related datasets within the same Experiment, and all data for the entire Study. HTTP, FTP, and Aspera links can be directly accessed through a web browser on this page. Please note that download via Aspera requires installation of the free Aspera Connect plugin. FTP links may be directly copied into a FTP client for browsing or download, if you prefer.

What is Aspera Connect?

Aspera Connect is a free software package that allows you to transfer data files using a browser plugin or a command line program called ascp (Aspera secure copy) on Linux, Windows, and Macintosh operating systems.

The Aspera Connect Fast And Secure Protocol (fasp) uses a User Datagram Protoccol (UDP) rather than the Transmission Control Protocol (TCP) used by File Transfer Protocol (FTP) — making Aspera much faster than FTP, so a transfer over a long distance is less likely to end in the middle of the transmission.

When you want to systematically download a large number of files, use the ascp command line program that is included in the Aspera Connect download instead of the Aspera Connect web plugin.

Some important Aspera Connect Features:

- The ability to restart if your data transfer is interrupted midstream
- Will run as a lower priority task if there is other data traffic on your network connections to avoid starving other protocols that you have running.
- Effective throughput up to 600 Mbps to a single site.

Which version of Aspera Connect Software do I need? Do all versions work equally well?

Make sure you are running a current version of Aspera Connect. If you experience any transfer problems or find that your transfer speed is below what you expect, update to the most recent version of Aspera Connect when you trouble-shoot your transfer problem.

Notes:

- There is no time limit on your Aspera Connect key; you can use it for downloading SRA data as long as NCBI/SRA contracts with Aspera.
- When you download the Aspera Software, be sure you are downloading the Aspera Connect *Client Software*, rather than the Aspera Connect *Server Software*.

When should I use Aspera Connect to download SRA data, and when should I use the Aspera secure copy (ascp) command line program?

Use Aspera Connect if you want to download files through an internet browser. If you want programmatically download of a large number of files, use the ascp command-line program.

Using the Aspera Connect Web Plugin Software

Downloading and configuring Aspera Connect

- Download Aspera Connect here.
- Instructions for installing and configuring the Aspera Connect web plugin software are in the SRA Aspera Transfer Guide.

Note: It is important to properly configure Aspera Connect, as the default transfer speed is only 10 Mbps. Please see the "Aspera Connect Configuration" section of the Aspera Transfer Guide for more information.

Note: Once you download Aspera Connect, it will run in the background so that when you click to download a file, the Aspera fasp protocol will automatically take precedence above the FTP protocol for your file transfer.

Note: We do **not** recommend running more than 4 Aspera Connect download sessions simultaneously.

Do I need to supply an ID or code key to use Aspera Connect web software to download NCBI data?

No. You do not need to supply an ID or code key. Since the key is already integrated into the software, when you click to download a file, the Aspera fasp protocol will start automatically.

Note: Your Aspera Connect key does not expire, and you can use it to download SRA data as long as NCBI/SRA contracts with Aspera.

Downloading dbGaP SRA Data**

**Please Note: The information provided in this section regarding the download of individual-level data from dbGaP assumes that you have completed the dbGaP data access request process, and have been granted authorized access to individual level data.

If you do not have authorized access to dbGaP individual-level data, accessing individual-level data using the instructions in this section will not be possible, as you will not have access to the appropriate password-protected dbGaP sites mentioned. For information on how to apply for authorized access, please see the "Data Access Request" section of the dbGaP FAQ Archive.

I found some human SRA data that I want to download, but the entry for the data I want doesn't have direct links to download the data —I have to go to dbGaP authorized access. Why can't I just download the data directly?

The sequence data you found is human individual level data —this means that **the data was derived from the tissue of a particular person who signed an informed consent agreement which specified restrictions in the way the data derived from their tissue can be used**.

Because of these restrictions, **the data can only be accessed by investigators who have successfully applied for dbGaP authorized data access**. This application process ensures that the stated research purpose of the applicant follows the restrictions stipulated by the study participant and that the access applicant and his or her institution will abide by the study's data use certification and terms of use.

Once you are granted authorized access:

- You will not have access to all the data housed in the authorized access portion of dbGaP since data access policies are determined on a per-study basis.
- Access will be only to the data you've been approved for in one particular study.
- You can request additional data sets from within the study you have been approved for, but these requests must be approved before the data can be released to you.
- If you need data files from more than one study, you will need to apply for authorized access for each study that generated the data you require.

You will find a brief description of the steps required access and download SRA data contained within dbGaP in this SRA Knowledge Base.

How do I access and download the dbGaP SRA data?

Below is a brief overview of the steps to access and download SRA data contained within dbGaP:

1. Find the dbGaP study and consent group names in any of the SRA report types (SRP, SRR, SRS) where the data you want appears. You will need these names for your controlled access application.
2. Apply for authorized access to the dbGaP study and consent group. An overview of the process for requesting access to a dbGaP study can be found here.
3. Request data sets from the study for which you have been granted access.
4. Obtain the dbGaP repository key file (".ngc" file). Information on .ngc files can be obtained here.
5. Download the SRA toolkit and use the included Java utility to configure a dbGaP workspace.
6. Run the Java configuration tool and import the .ngc file.
7. **IMPORTANT:** Switch, or 'cd' into the configured workspace directory made in (5) you can then download and operate directly on encrypted SRA data ("on-the-fly" download and "as-needed"

decryption) or move a previously-downloaded ".sra.ncbi_enc" encrypted SRA file into your configured workspace and operate directly on it.

Note: **Only a** person classified as **a Principle Investigator or PI** (NIH extramural or intramural researcher) **can apply for dbGaP authorized access**.

Non-PIs cannot independently apply for access to individual-level data, but they can be approved for local access to downloaded data files within the PI's lab if they are listed as collaborating investigators on a PI's application.

An explanation of why these steps are necessary is available elsewhere.

Problems Downloading Data

FTP Download Quits before Complete

I am trying to download files from the SRA, but the FTP process quits before the files are completely downloaded.

Sometimes this can be a temporary problem; wait and try your download again in an hour or two. If you continue to have problems, here are a few suggestions:

- If you are using an internet browser as your transfer client, you may want to consider a dedicated FTP client. There are many free FTP clients like FileZilla, Cyberduck, and NcFTP available.
- Do you regularly use FTP at your location without issue? If you are unable to access and download data from the Mozilla Foundation's FTP site, FTP traffic may be blocked or restricted by your institution.
- Consider using Aspera Connect to conduct your download since:
  - The amount of data for SRA projects can exceed 10 gigabytes and traditional FTP may be too slow to download your data effectively.
  - FTP performance degrades proportionally with the number of hops or switches the data must take to get to you. Aspera performance does not degrade with distance.

See the "Performance comparison of FTP and ascp downloads" section of this Knowledge Base for information on the differences between FTP and Aspera for SRA downloads.

No "Download Data" Link in Entrez Search Results

What if the data download link is missing from either an experiment or a run I found using Entrez?

If the download links are missing, then there is a technical issue with the data. Please contact sra@ncbi.nlm.nih.gov and provide us with the Experiment or Run accession.

No Download Link for Specific File Types

I want to download the data I've found in a particular format, but I only see a download link for .sra files.

The SRA archive format (".sra files") can be converted to several standardized file formats, including fasta, fastq, sam/bam, sff, ABI colorspace fasta/qual, and Illumina native. Please note that not all SRA data can be converted into all of the above data formats (this is determined by the submitted format: data submitted as fastq does not contain the information necessary to build an sff file, for example). Please see this page for more information. If you are only interested in fasta- or fastq- formatted data, please see the following question.

We store data in the SRA format due to storage space constraints — storing the data in different formats (.fastq, .sff, etc.) requires far more storage space than saving the data in a single archive format and allowing users to generate multiple different file types from the same source.

I am only interested in fasta- or fastq- format data. How can I download this directly?

From this page, you may enter an Experiment accession (SRX accession), or a comma-separated list of Experiment accessions, that will take you to a page like this on which you may specify (1) from which Run(s) you would like to download data, (2) the formatting (fasta or fastq), and (3) options relating to sequence filtering, trimming, and clipping. The latter page may also be reached through the Run Browser by selecting the "Reads" tab and then clicking "Filtered download" (leave the filter empty to request all reads).

fastq-dump outputs color space fastq containing numbers in the sequence reads instead of "normal" base space fastq.

Fastq files that have numbers instead of letters are in color space format. This format is a feature unique to the methodology used by the ABI SOLiD sequencer. To learn more about color space and its relationship to base space (ATCG) see the ABI webinar on the on the fundamentals of 2 Base Encoding and Color Space, and the ABI documentation "Color Space Analysis in the SOLiD™ System: the Theory, Advantages and Solutions".

In order to dump a "normal" base space fastq file, please run the fastq-dump command with the "-B" option. Conversely, to dump base space data as color space fastq, please use the "-C" option.

Using the SRA Toolkit to convert .sra files into other formats

What is the purpose of the SRA toolkit?

The SRA Toolkit, and the source-code SRA System Development Kit (SDK), will allow you to programmatically access data housed within SRA and convert it from the SRA format to the following formats:

- ABI SOLiD native (colorspace fasta / qual)
- fasta
- fastq
- sff
- sam (human-readable bam, aligned or unaligned)
- Illumina native

You can also use the toolkit to convert from the formats listed below into the SRA format (not required for submission, but will allow you to use the SRA Toolkit to archive or analyze your data):

- fastq or fasta/qual pairs
- AB SOLiD-SRF
- AB SOLiD-native
- Illumina SRF
- Illumina native
- sff
- Aligned bam

The SRA toolkit is available in versions compatible with Linux, Windows and Mac operating systems.

How do I download and install the SRA Toolkit?

The SRA Toolkit can be obtained from SRA Software page. Please note that as of version 2.3.2, only 64-bit versions of the Toolkit are being produced. The reasons for this decision are manifold, but are primarily due

to the limited memory and processing capacities of 32-bit operating systems, which are insufficient for handling large SRA data files. Legacy versions of the Toolkit, including previous 32-bit versions, are available here, but please note that we are serving these files "as is" – we are happy to assist with usage (email sra@ncbi.nlm.nih.gov), but bugs (known and unknown) will not be addressed. It is strongly recommended that you configure the Toolkit prior to using it to extract data.

How do I use the SRA Toolkit to convert data into a particular format?

The SRA Toolkit contains a series of independent data-"dump" utilities that will allow you to convert SRA data into different file formats. As of version 2.3.2, the list of "dumpers" that are included with the toolkit include:

- fastq-dump: Converts data to fastq and fasta format.
- sam-dump: Converts data to sam (human-readable bam). Data submitted as aligned bam are output as aligned sam, while other formats are output as unaligned sam.
- sff-dump: Converts data to sff format. Note that only data submitted as sff can be converted back to this format.
- abi-dump: Converts data to csfasta/csqual format. Note that data submitted in base-space can be represented in color-space, but please be aware of the advantages / disadvantages of converting between different encodings.
- illumina-dump: Converts data to Illumina native and qseq formats.
- vdb-dump: Exports the vdb-formatted data of the .sra file.

Each of the above links will open the current documentation / help page for the respective utility, which include frequently used options and their definitions, usage examples, and common errors messages / solutions. Please send all Toolkit questions to: sra@ncbi.nlm.nih.gov

I'm having problems using the toolkit, and the documentation doesn't cover the problem I'm having. Who do I contact for help?

Send any toolkit questions you have to: sra@ncbi.nlm.nih.gov

Be sure to provide as much detail as possible so that we may more quickly diagnose your problem: Your operating system, Toolkit version, the command that you are attempting to execute, error messages and/or the "ncbi_error_report.xml" (if one was generated).

Searching for Data

Sequence Read Archive Submissions Staff

Created: August 22, 2011; Updated: March 18, 2014.

This section of the Sequence Read Archive (SRA) Knowledge Base contains:

- Instructions for searching for specific data types
- Suggestions for search terms
- Instructions for using SRA search tools
- Answers to common problems that happen while searching
- Information on how to interpret your search results.

To begin searching this section of the Sequence Read Archive Knowledge Base, you can either:

- **Enter** your **search word(s)** text in the text box **at the top of the page** and click on the "Go" button,

  OR
- **Clickon** any of the "SRA Knowledge Base" **sub-categories listed** in the "Contents" section **below** to go to the sub-category of your choice.

Searching for Species-Specific Data

How do I find SRA data for African strains of fruit fly using Entrez SRA?

Enter the search phrase **fruit fly AND Africa** or **Drosophila AND Africa** into the Entrez SRA search box, and click the "search" button.

- You can make your search more specific by using Entrez search limits. Click the "Limit" link to specify a value for any or all of the following:
  - Publication date
  - Search Field Tags
  - Access level
  - Date Modified
  - Molecule
  - Availability of loadable data

Once you have specified your limits, click the "Search" button on the Limit page to apply your limits to the search terms you entered in the search box at the top of the page.

- You can also make your search more specific by using the Entrez "Advanced" search. Click the "Advanced" link to go to the SRA Advanced Search page, which contains:
  - "Search Builder"
    This tool helps you select search fields and field values to create a search statement that reflects your specific search requirements.
  - "Search History"
    This tool keeps a record of up to 100 searches and their results so that you do not duplicate searches and can combine previous searches for more specificity.
- For help using Entrez "Limits" or "Advanced" search options, see the Entrez Help documentation.

Note: Since SRA is a raw data archive (not curated) and can index only those terms that a submitter provides in the metadata of their submission, your search results may not include all possible studies associated with your search term.

How do I find SRA data for African strains of fruit fly using the "Object Search" tab on the SRA home page?

Enter the search phrase **fruit fly AND Africa** or **Drosophila AND Africa** into the "Search for SRA related objects" search box, and click the "search" button. You can use any Entrez search statement in SRA's "Search for related objects", but this search page does not have the "Limit" and "Advanced" search features that the Entrez SRA search has.

Note: Since SRA is a raw data archive (not curated) and can index only those terms that a submitter provides in the metadata of their submission, your search results may not include all possible studies associated with your search term.

How do I find SRA entries for sequenced complete genomes of the Euryarchaeota group of archaea?

1. Go to the Entrez Taxonomy page.
2. Enter **Euryarchaeota[organism] AND taxonomy_sra[filter]** in the search box at the top of the page.
3. Click the "Search" button.
4. The response page provides you with a list of the organisms in this phylum that have published SRA sequencing.
5. Click the organism of interest to go to the Taxonomy Browser page for that organism.
6. Click on the link for "SRA Experiments" under "Entrez Records" for this organism to go to the SRA database.
7. You will see a list of experiments (SRX/ERX/DRX accessions).
8. Click on the SRX record of interest. The record will include:
    a. Links to each run in the experiment
    b. Experiment design information
    c. Submission accession and submitter name
    d. Study summary and abstract (if available)
    e. Experiment sample (SRS) information
    f. Links to a list of experiments (SRX)
    g. Library information (if available)
    h. Platform information
    i. Processing information (base calls and quality scores)
    j. Spot descriptor information
    k. Links to related information
9. At this point, you can do one of the following:
    a. **Click on the "All experiments" link** for a complete list of the experiments included in the study.
    b. **Download the data for a particular experiment** by clicking on the SRA or the SRA-lite download links. These links will take you to the SRA FTP site where the files are available for download. If there are no download links in the experiment record, please contact sra@ncbi.nlm.nih.gov.
    c. **Click on one of therun links(**SRR/ERR/DRR**accessions)** to go to the SRA Run Browser, where you can look at individual reads, search the spots for a specific sequence, or download a set of spots from the run.

Note: Since SRA is a raw data archive (not curated) and can index only those terms that a submitter provides in the metadata of their submission, your search results may not include all possible studies associated with your search term.

Note: If you choose to download files, install the "Aspera Connect" plug-in to transfer files at a significantly faster rate than ftp.

Searching for Specific Study Types

How do I find all instances of a specific study type (e.g. human genomic re-sequencing experiments) in the SRA?

1. Go to the Entrez SRA page.
2. Enter **human[organism] AND study type resequencing[Properties] AND biomol genomic[Properties]** in the search box at the top of the page.
3. Click the "Search" button.
4. Your search results will be presented as a list of human genomic resequencing experiments (SRX/ERX/DRX accessions) available in the SRA.
5. Click on the SRX record of interest. The record will include:
    a. Links to each run in the experiment
    b. Experiment design information
    c. Submission accession and submitter name
    d. Study summary and abstract (if available)
    e. Experiment sample (SRS) information
    f. Links to a list of experiments (SRX)
    g. Library information (if available)
    h. Platform information
    i. Processing information (base calls and quality scores)
    j. Spot descriptor information
    k. Links to related information
6. At this point, you can do one of the following:
    a. **Click on the "All experiments" link** for a complete list of the experiments included in the study.
    b. **Download the data for a particular experiment** by clicking on the SRA or the SRA-lite download links. These links will take you to the SRA FTP site where the files are available for download. If there are no download links for the experiment record, please contact sra@ncbi.nlm.nih.gov.
    c. **Click on one of therun links(**SRR/ERR/DRR**accessions)** to go to the SRA Run Browser, where you can look at individual reads, search the spots for a specific sequence, or download a set of spots from the run.

Note: Since SRA is a raw data archive (not curated) and can index only those terms that a submitter provides in the metadata of their submission, your search results may not include all possible studies associated with your search term.

Note: If you choose to download files, install the "Aspera Connect" plug-in to transfer files at a significantly faster rate than ftp.

Searching for Specific Molecule Types

How do I find all of the human transcriptome (RNA) data in SRA?

1. Go to the Entrez SRA page.
2. Enter **human AND RNA** in the search box at the top of the page.
3. Click the "Search" button.
4. Your search results will be presented as a list of human transcriptome experiments (SRX/ERX/DRX accessions) available in the SRA.
5. Click on the SRX record of interest. The record will include

   a. Links to each run in the experiment
   b. Experiment design information (if available)
   c. Submission accession and submitter name
   d. Study summary and abstract (if available)
   e. Experiment sample (SRS) information
   f. Links to a list of experiments (SRX)
   g. Library information (if available)
   h. Platform information
   i. Processing information (base calls and quality scores)
   j. Spot descriptor information
   k. Links to related information
6. At this point, you can do one of the following:
   a. **Click on the "All experiments" link** to see a complete list of the experiments included in the study.
   b. **Download the data for a particular experiment** by clicking on the SRA or the SRA-lite download links. These links will take you to the SRA FTP site where the files are available for download. If there are no download links visible for the experiment record, please contact sra@ncbi.nlm.nih.gov.
   c. **Click on one of therun links(**SRR/ERR/DRR**accessions)** to go to the SRA Run Browser, where you can look at individual reads, search the spots for a specific sequence, or download a set of spots from the run.

Note: Search results may not include all studies that used the specified platform or instrument since SRA is a raw data archive (not curated) and can index only those terms that a submitter provides in the metadata of their submission.

Note: If you choose to download files, install the "Aspera Connect" plug-in to transfer files at a significantly faster rate than ftp.

Searching for Data Generated by a Specific Platform or Instrument

Can I search for SRA data by platform/instrument type?

Yes, you can search SRA for data by platform/instrument type using Entrez SRA.

- **Search for the platform/instrument typeby placingquotation marks around the name of the platform/instrument** (e.g. **"ab solid system 3 0"**) **or by using a qualifier at the end of a compound word** (e.g. **ab solid system 3 0[word]***or***ab solid system 3 0[all]**).
- **If you search for an unquoted or unqualified phrase** (e.g. **ab solid system 3 0**), Entrez will not treat the words as a name – they will be treated as list of words joined by the Boolean operator AND. This means Entrez will search for each of the words in all fields irrespective of order, and **your search results may contain records that would not normally be included in the results of a search for the name of the platform/instrument type**.
- To see how Entrez interpreted your query, look in the "Search Details" text box on your result page.

Your search results will be presented as a list of experiments (SRX/ERX/DRX accessions) whose metadata contains the phrase you entered in the query box.

1. Click on the SRX record of interest. The record will include:
   a. Links to each run in the experiment
   b. Experiment design information (if available)

c. Submission accession and submitter name
d. Study summary and abstract (if available)
e. Experiment sample (SRS) information
f. Links to a list of experiments (SRX)
g. Library information (if available)
h. Platform information
i. Processing information (base calls and quality scores)
j. Spot descriptor information
k. Links to related information

2. At this point, you can do one of the following:
   a. **Click on the "All experiments" link** to see a complete list of the experiments included in the study.
   b. **Download the data for a particular experiment** by clicking on the SRA or the SRA-lite download links. These links will take you to the SRA FTP site where the files are available for download. If there are no download links for the experiment record, please contact sra@ncbi.nlm.nih.gov.
   c. **Click on one of therun links(**SRR/ERR/DRR**accessions)** to go to the SRA Run Browser, where you can look at individual reads, search the spots for a specific sequence, or download a set of spots from the run.

Note: Search results may not include all studies that used the specified platform or instrument since SRA is a raw data archive (not curated) and can index only those terms that a submitter provides in the metadata of their submission.

Note: If you choose to download files, install the "Aspera Connect" plug-in to transfer files at a significantly faster rate than ftp.

Searching for Recently Submitted Data Accession Numbers

I can't find the SRA accession number of data I just submitted using an Entrez search. The query browser states that the link is not public, even though we have not placed any holds on our submitted data.

Once you have submitted, your data will **not** be available right after you submit it — it will take 24 hours for Entrez to re-index and include your newly submitted data as a search result.

If 24 hours have passed since you submitted, and you cannot get your accession number through an Entrez search, your accession may be in a default "HUP" (Hold Until Publication) state since all SRA submissions default to a HUP status even if you did not place a hold (i.e. set a release date) on the original submission.

To determine the status of your submission, do the following:

First, check to see if your submission or individual accessions in it are in a default hold:

1. Log in to SRA submissions using your NCBI PDA account.
2. Look at the "State" column of your login page. This column will contain the text "HUP" for submissions that are in the default HUP state.
3. Click the accession of the submission you are trying to find.
4. Check the status of the "Released" column for each of the individual accessions within your submission. If the "Released" column is blank (does not contain a future release date for that accession) the accession is in the default HUP state.

To release your accession from the default hold, click the "Release submission now" button, or set a release date. Once the hold on a submission is released, you can find your submission using an Entrez search in 24 hours.

I completed a submission to SRA this morning, but cannot find my accession in Entrez when I searched for it this afternoon.

Once you have submitted, your data will **not** be available right after you submit it — it will take 24 hours for you to get your new accession number as a search result.

If 24 hours have passed since you submitted, and you cannot get your accession number through an Entrez search, your accession may be in a default "HUP" (Hold Until Publication) state, since all SRA submissions default to a HUP status even if you did not place a hold (i.e. set a release date) on the original submission.

To determine the status of your submission, do the following:

First, check to see if your submission or individual accessions in it are in a default hold:

1. Log in to SRA submissions using your NCBI PDA account.
2. Look at the "State" column of your login page. This column will contain the text "HUP" for submissions that are in the default HUP state.
3. Click the accession of the submission you are trying to find.
4. Check the status of the "Released" column for each of the individual accessions within your submission. If the "Released" column is blank (does not contain a future release date for that accession) the accession is in the default HUP state.

To release your accession from the default hold, click the "Release submission now" button, or set a release date. Once the hold on a submission is released, you can find your submission in 24 hours using an Entrez search.

Searching for New Studies added to SRA

Does the SRA have a notification system in place to let people know when SRA data from new GEO studies are added to SRA?

SRA does not have a notification system that will alert users when new GEO studies have been added. **But you can set up your own notification system using an Entrez search in conjunction with My NCBI.** To do this, you will construct an Entrez search of GEO DataSets for studies that contain SRA data. Then, using My NCBI, you will set up an automated query as an update of your first query (thereby reporting only new entries), that will run periodically (daily, weekly, monthly, etc.) and generate an email notification.

1. Go to the GEO DataSets (gds) page.
2. Click on the "Advanced" link at the top of the page.
3. Go to the "Search Builder" section of the page.
4. Activate the drop-down menu that contains the words "All Fields" and select "Filter".
5. Click on the "Show Index" link to generate a pop-up menu that lists all available filters for GEO DataSets.
6. Scroll down the list (which is arranged alphabetically), select **gds sra[Filter]**, and then click on the "Add to Search Box" button. This will place your selected filter in the search box at the top of the page.
7. Click the "Search" button at the top of the page.
8. Once you have the response page listing all the GEO DataSets containing SRA data, click on the "Save Search" link located to the right of the search box at the top of the page to go to "My NCBI".
9. Sign in to My NCBI or register for a My NCBI account by clicking on the "Register for an account" link. Help for My NCBI is available online.
10. When the My NCBI "Saved Searches" page appears, click the "Save" button to save your search and generate a response page asking for your search settings.

11. Fill out the response page, which will allow you to select how often (monthly, weekly, daily) you want an update showing new results for the search you just conducted sent to an email address you provide.

You may also perform this search programmatically using Eutils.

Note: You will not be able to use the above instructions to search BioProject, dbGaP or BioSample since these resources currently do not support email updates. Only GEO DataSets and Genome currently provide an SRA filter and support email updates.

As BioProject, dbGaP and BioSample do not support email updates, you will have to perform a manual search for new studies/projects that contain SRA data in these resources and scan the results of your search for new studies yourself.

Which NCBI resources support an Entrez filtered search for studies/projects that contain SRA data and what are the search terms I use for them?

Resources supporting Entrez filtered search for projects containing SRA data:

Resource	Search Term
BioSample	"biosample sra"[Filter]
dbGaP	"gap sra"[Filter]
Genome	"genome sra"[Filter]
BioProject	"bioproject sra"[Filter]
GEO DataSets	"gds sra"[Filter]

Searching for Data Generated by a Specific Next-Generation Sequencing Application

Can I search SRA for data generated by a specific next-generation sequencing application like RNA-seq, ChIP-seq or CLIP-seq?

Yes you can by going to the Entrez SRA page, and entering **"RNA-seq"**, **"CHIP-seq"** or **"CLIP-seq"** in the search box at the top of the page and clicking the "Search" button.

If a submitter of RNA-seq, CHIP-seq, or CLIP-seq data didn't happen to use the term RNA-Seq, CHIP-seq or CLIP-seq anywhere in their submission, the above search will not find their records. If this is the case, you could try the following:

- Formulate your search using terms that would describe the data generated from that specific technology.

  For example, If the application name ChIP-seq wasn't indexed in Entrez (it is indexed; the name is only being used for the purpose of this example), you could use the following search terms to find data generated by the ChIP-seq method:
  - "protein"
  - "chromatin"
  - "protein" AND "chromatin"

  Notes:

    - The above search terms may not turn up data from the same ChIP-seq studies since SRA is a raw data archive (not curated) and can index only those terms that a submitter provides in the metadata of their submission.
    - Depending on which terms the submitter used in their submission, the data generated by these searches may or may not contain the application term (e.g. ChIP-seq), so you may have to click on the "Study Summary" link within the experiment entry of interest to see if the submitter mentions the application name.
    - The results for a search of this type will include more than just the data generated by the next-generation sequencing application, so you will have to review the search results yourself and look for data generated by the application of interest.
- If the above search strategy yields nothing, search PubMed using the application name and limit the search to the title (**"ChIP-seq"[Title]**) to find studies whose title contains the text: ChIP-seq. Open the PubMed record of interest in the resulting page. If SRA data was submitted for that publication, there will be a link to it in the PubMed record.

I searched SRA for data generated by a new next-generation sequencing application. I used the name of that application as my search term, but found no records. What did I do wrong?

If the sequencing technology is relatively new, it may not be indexed in SRA yet, so you won't get any results for your search. If this is the case, create your search using terms that would describe the data generated from that specific technology.

For example, If the application name ChIP-seq wasn't indexed in Entrez (it is indexed; it is only being used for the purpose of this example), you could use the following search terms to find data generated by the ChIP-seq method:

- "protein"
- "chromatin"
- "protein" AND "chromatin"

Notes:

1. The search terms you use may not give you all the SRA data associated with that search term since SRA is a raw data archive (not curated) can only index terms that are provided in the metadata of the submission.
2. Depending on which terms the submitter used in the text of their submission, the data generated by these searches may or may not contain the application term (See note number 1 above), so you may have to click on the "Summary" link within an experiment entry of interest to see if the submitter mentions the application name.
3. The results for a search of this type will include more than just the data generated by the next-generation sequencing application, so you will have to review the search results yourself for data generated by the application of interest (e.g. ChIP-seq).

If the above search strategy yields nothing, search PubMed using the application name and limit the search to the title ("ChIP-seq"[Title]) to find studies whose title contains the text: "ChIP-seq". Open the PubMed records of interest in the resulting page. If SRA data was submitted for that publication, there will be a link to it in the PubMed record.

If a next-generation sequencing application is not indexed by name in SRA, please contact sra@ncbi.nlm.nih.gov.

Searching for Data by Submitter or Author

Can I search SRA by a submitter or author's name?

Unfortunately, SRA is not indexed by the author/submitter name, so you cannot use this as a search term. Instead, search PubMed using the author/submitter's name. If there is a publication by the suthor/submitter that is associated with SRA data, when you find the publication in PubMed, you will also find a link to the SRA data within the PubMed publication record.

Unfortunately, SRA and Pubmed are not reciprocally indexed, so you cannot search SRA with the author/submitter's name to get a link for a PubMed article.

Searching for PubMed Articles Associated with SRA Data

How do I find PubMed articles that have mouse SRA data associated with them?

To find PubMed articles that have mouse SRA data associated with them, do the following:

1. Go to the Entrez SRA page.
2. Enter **mus musculus[organism] AND "sra pubmed"[filter]** in the search box at the top of the page.
3. Click the "Search" button.
4. Your search results will be presented as a list of mouse experiments (SRX/ERX/DRX accessions) in the SRA that have been reported in a PubMed article.

Searching for Data Referenced in a Publication

The paper I'm reading says all the data described in it was deposited in SRA, but it doesn't list any accession numbers. How do I find the data?

There are many reasons that accessions may be unavailable (e.g. the submission may be incomplete or the submission may be released but not indexed). If the paper you are reading mentions that the data was submitted to SRA, but does not provide accession numbers for the data, please contact sra@ncbi.nlm.nih.gov.

There are SRA accession numbers in the PubMed article I'm reading, but there are no links to SRA from the accessions.

We must create a link *from SRA to the accession* in a PubMed article. If you find a PubMed article containing SRA accessions that do not have links to SRA, please contact sra@ncbi.nlm.nih.gov, and we'll insert the links.

Unable to Find Data in SRA Using ID given in Publication

When I search for a submission accession (SRA) provided in a publication, I get a list of accessions for experiments (SRX) rather than the actual submission. How do I get data for a submission accession (SRA)?

An SRA submission accession number does not represent a single object the way a GenBank accession number represents a single object (a sequence) submitted to GenBank. The SRA submission accession is an artificial packaging construct that groups together a number of objects, and therefore has no specific response page.

For example, the GenBank accession NM_001048036.2 represents a single object: a GenBank sequence submission (in this case, a RefSeq mRNA from the domestic dog). An SRA (submission) accession number (e.g. *SRA010122*), doesn't represent a single object – it is a reference number that represents a group of objects that includes all of the following:

Study (a metadata object that describes the entirety of a sequencing study — these are represented by SRP/ERP/DRP accession numbers) .The study accession housed by submission SRA01012 is SRP001451.

Experiments (metadata objects that describes the experiments within the study — these are represented by SRX/ERX/DRX accession numbers). One of the experiment accessions contained in SRA010122 is SRX013300

Runs (objects containing sequencing run data files and their associated metadata— these are represented by SRR/ERR/DRR accession numbers) One of the run accessions contained in SRA010122 is SRR030732

Samples (metadata objects that describe material sequenced in the study — these are represented by SRS/ERS/DRS accession numbers). One of the run accessions contained in SRA010122 is SRS007212

Analysis (a packaging construct for data objects [and their associated metadata] generated by different types of sequence analysis: e.g. assemblies, alignments, etc.)

Because a submission accession (SRA) is not an object itself but a container for objects (SRP, SRX, SRR, etc.), when you search for a particular SRA accession, you will get is a list of the experiments (SRX accessions) from the submitted study. The experiment entries contains links that will take you to the all the other objects contained in a particular submission.

In time, we hope to generate a response page specific to submission (SRA) accession searches. This response page will contain links to all the objects contained by the submission accession.

Searching using BLAST

Can I conduct a sequence similarity search (BLAST) against data in the SRA?

Yes you can. NCBI offers a nucleotide BLAST service for sequence similarity searches of SRA transcript and genomic libraries. You can access this service using one of the following routes:

- Go to the BLAST home page, and select the "Search SRA transcript and genomic libraries" link located in the "Specialized BLAST" section
- Go to the SRA home page, click the "Search" tab, select the "BLAST" link, and when you see the response page, select the "Search transcripts" link to get to the BLAST SRA page.

Note: Currently, BLAST only works with data sets that have long sequence reads. This means that an SRA BLAST search will not work with any data sets that may have been submitted for a species that have short sequence reads.

If you require a sequence similarity search against species data that is in the SRA but is not listed as available for BLAST, SRA does offer a limited sequence similarity search for a subset of spots with in a run. Go to the run browser page for a run of interest and enter your sequence in the "Filter" text box and click the "Find" button. You can also search for spots within a run using:

- The name of the spot

  OR
- The sample pool member

I would like to BLAST my sequence against SRA, but I don't see the species I would like to BLAST against listed in the search set menu.

Because a successful BLAST search requires reads longer than some SRA data sets contain, not all SRA submissions can be used for a BLAST search.

If the species that you are interested in BLASTing against isn't in the SRA BLAST list, then it means that data with long enough reads to be useful for BLAST has not been submitted for it yet.

If you require a sequence similarity search against a species data that is in the SRA but is not listed as available for BLAST, SRA does offer a limited sequence similarity search for a subset of spots with in a run. Go to the run browser page for a run of interest and enter your sequence in the "Filter" text box and click the "Find" button. You can also search for spots within a run using:

- The name of the spot

  OR
- The sample pool member

Problems Searching SRA

Why don't I get results when I search for Black Mustard (Brassica nigra) data? How do I find black mustard data in SRA?

Below are two negative search scenarios for black mustard that may describe why you are not getting results:

Scenario # 1: Although the publication you read does not provide SRA accession numbers, it does provide sequencing details that lead you to believe the author may have deposited the data in the SRA, yet when you look for the data in SRA you can't find it.

Frequently, when SRA accessions are not included in a paper, it is because the data was not actually submitted to the SRA — it was submitted to another NCBI resource (e.g. GenBank or others) instead. So if you get a negative result for a particular search in SRA, try searching other NCBI resources for the sequence data.

Scenario # 2: Often, when you search for SRA data for a particular species one day, you won't find anything, but the next day you perform the same search, and you might get 150 data records. The reason this happens is that there will be a long period where no one submits data for a particular species, and then a research group completes a project and submits all of their data to SRA at one time.

You can use Entrez and "My NCBI" to set up an automated periodic search of SRA for the term "black mustard" (or "Brassica nigra") that will notify you when updates to this search are available:

a. Go to the Entrez SRA page.
b. Enter your search term in the text box at the top of the page. (e.g. black mustard or Brassica nigra)
c. Click on the "Advanced" link.
d. Go to the "Search Builder" section of the response page.
e. Activate the drop-down menu that contains the words "All Fields" and select "Filter".
f. Click on the "Show Index" link to generate a pop-up menu that lists all available filters for SRA.

   Note: Only GEO DataSets and Genome currently provide an SRA filter and support email updates. BioProject, dbGaP and BioSample currently do not support email updates. For those resources that do not support email updates, you will have to perform a manual search for new studies/projects that contain SRA data in these resources and scan the results of your search for new studies yourself.
g. Scroll down the list (which is arranged alphabetically), select either **sra gds** or **sra genome**, and click on the "Add to Search Box" button. This will place your selected filter in the search box at the top of the page.
h. Click the "Search" button.
i. Once you have the response page listing all the GEO DataSets containing SRA data, click on the "Save Search" link to go to "My NCBI".
j. Sign in to My NCBI or register for a My NCBI account by clicking on the "Register for an account" link. Help for My NCBI is available online.

k. When the My NCBI "Saved Searches" page appears, click the "Save" button to save your search and generate a response page asking for your search settings.
l. Fill out the response page, which will allow you to select how often (monthly, weekly, daily) you want an update showing new results for the search you just conducted sent to an email address you provide.

You may also perform this search programmatically using Eutils.

I searched SRA using a specific term but don't get any results. Are there other ways of looking for SRA data associated with that search term?

Try using your search term in either BioSample or BioProject and check your results for links to SRA experiments. If this doesn't work, try using Eutils.

Understanding SRA Search Results

SRA Response Page Links

The Entrez SRA search response page provides a list of links to SRX/ERX/DRX accessions. What are these accessions and where do these links go?

These accessions are for experiments (SRX) from various studies associated with the search terms you used. The SRX/ERX/DRX links go to an experiment report page that provides:

- Links to each run in the experiment
- Experiment design information
- Submission accession and submitter name
- Study summary and abstract (if available)
- Experiment sample (SRS) information
- Links to a list of experiments (SRX)
- Library information (if available)
- Platform information
- Processing information (base calls and quality scores)
- Spot descriptor information
- Links to related information

Note: If you choose to download files, install the "Aspera Connect" plug-in to transfer files at a significantly faster rate than ftp.

Each SRX entry in the Entrez SRA search response page contains a list of SRR/ERR/DRR accessions that are also links. What are these accessions and where do these links go?

These accessions are for runs for that experiment (SRR). The links go to the Run Browser page for each run.

The Run Browser page provides you with specific information about the run, including read and intensity data, and **will allow you to download data from a particular run experiment or study.**

You can also **use the Run Browser to search the spots of a particular read for a specific sequence**:

1. Enter sequence data into the "Filter" text box.
2. Click the "Find" button.
3. If your sequence is found, a list of the spots that contain the sequence entered will appear just below the grey box containing the "Find Spots" query, as will links to runs that contain the sequence data you are looking for.m
4. You can also use the "Filter" to search for the name of a spot or the name of a sample pool member.

If you want to download the data you find using the run browser's filter, click the "Filtered Download" button. The response page allows you to download the data in a selected format (filtered, clipped, FASTA and FASTQ).

Confused About which Accession Number to Use

When I use Entrez SRA to search for data in SRA, which link on the search response page do I use to see the data ?

When you get the response page from an SRA Entrez search, you will see a list of experiment accessions (SRX) from studies associated with the search terms you used. The SRX/ERX/DRX links go to an experiment report page that provides:

- Links to each run (SRR) in the experiment
- Experiment design information
- Submission accession and submitter name
- Study summary and abstract (if available)
- Experiment sample (SRS) information
- Links to a list of experiments (SRX)
- Library information (if available)
- Platform information
- Processing information (base calls and quality scores)
- Spot descriptor information
- Links to related information

Note: If you choose to download files, install the "Aspera Connect" plug-in to transfer files at a significantly faster rate than ftp.

When I use SRA Search browser to search for an SRA submission, which link on the search response page do I use to see the data?

When you get the response page from the SRA "Search for Related objects", the most helpful link is the SRA Studies link. The link will provide you with a list of links to all the studies associated with the search terms you used. Once you are on the response page, click on the SRP accession number for the study of interest to go to the study summary page. The study summary page includes:

- Submission accession and submitter name
- Study abstract (if available)
- A text description of the study (if available).
- Links to associated NCBI resources (BioProjects, GEO, WGS, dbGaP, PubMed, etc.).
- Links to associated external resources.
- Links to all experiments in study
- Link to Entrez summary documentation for all experiments.
- Links to download study reads

Note: If you choose to download files, install the "Aspera Connect" plug-in to transfer files at a significantly faster rate than ftp.

Why does SRA have so many different accession types? A single GenBank ID number accesses all the information you need for a particular molecule.

An SRA submission accession number does not represent a single object the way a GenBank accession number represents a single object (a sequence) submitted to GenBank. The SRA submission accession is a

modular data construct that groups together a number of objects, and therefore has no specific response page.

For example, the GenBank accession NM_001048036.2 represents a single object: a GenBank sequence submission (in this case, a RefSeq mRNA from the domestic dog). An SRA (submission) accession number (e.g. *SRA010122*), doesn't represent a single object – it is a reference number that represents 5 distinct data types that are independently reusable.

Examples of repeated and independent use of SRA Accessions:

1. A submitter can reference (link to) a sample (SAMN accession) multiple times from any experiment they submit. A good case in point for this example is HapMap samples, some of which have been referenced (linked to) in experiments from 5 or 6 different submitting centers (see SAMN00004417).
2. The set of variables captured in an Experiment (SRX) accession is similar to what a biologist might think of as a "library", and includes the nucleic acid type extracted from the sample, as well as the selection type applied to enrich the intended sequencing target. This information can be referenced (linked to) by any run (SRR) produced using the set of variables described by that SRX.
3. The SRP accession, which provides all the metadata describing a particular study, references (links to) all the experiments (SRX) contained within the study, while each of the SRX accessions for that study all reference the same SRP accession. Since the SRP accession ultimately references (links to) all 5 data types in a study, it can be used as a starting point to access any of the data in that study.

Benefits of SRA modular data construction to the submitter:

- It is a **time-saver** since the submitter can directly reference (link to) existing data or metadata for a study, analysis, sample, experiment, or run from any point within a submission, instead of having to physically enter the data.
- If a particular instance of any one of the 5 data types in a particular study is not yet available, once the submitter creates a particular instance of a data type , that data can be referred to again and again in the current submission or in a future submission.

Benefits of SRA modular data construction to the user:

- It is **a time-saver** since the user can look a single instance of any one of the 5 data types from a particular study without having to download the data for the entire study.
- Once the metadata for a particular experiment (SRX) or sample (SRS) is generated, there is **uniformity of description** within the Sequence Read Archive for that particular experiment or sample, since the metadata description can be used over and over again in the same submission or in other submissions.
- The user can **make logical connections between studies** when different studies use the same samples. Because the samples are linked to the experiments, which in turn are interlinked to studies, runs and analyses, it is easy for the user to compare data between studies.

What do the different SRA accessions represent?

There are 6 different SRA accession types:

Accession Prefix	Accession Name	Definition	Example
SRA	SRA submission accession	The submission accession represents a virtual container that holds the objects represented by the other five accessions and is used to track the submission in the archive.	Since the SRA accession number is an artificial packaging construct, there is no example available since the SRA accession number has no specific response page
SRP	SRA study accession	A Study is an object that contains the project metadata describing a sequencing study or project. Imported from BioProject.	HTML
SRX	SRA experiment accession	An Experiment is an object that contains the metadata describing the library, platform selection, and processing parameters involved in a particular sequencing experiment.	HTML
SRR	SRA run accession	A Run is an object that contains actual sequencing data for a particular sequencing experiment. Experiments may contain many Runs depending on the number of sequencing instrument runs that were needed.	HTML
SRS	SRA sample accession	A Sample is an object that contains the metadata describing the physical sample upon which a sequencing experiment was performed. Imported from BioSample.	HTML
SRZ	SRA analysis accession	An analysis is an object that contains a sequence data analysis BAM file and the metadata describing the sequence analysis.	

The first letter of the accession prefix shows which INSDC archive the data originated from (S =NCBI-SRA, E = EMBL-SRA, D = DDBJ-SRA:

Data originating from EMBL-EBI:

Accession Prefix	Accession Name
ERA	ERA submission accession
ERP	ERA study accession
ERX	ERA experiment accession
ERR	ERA run accession
ERS	ERA sample accession
ERZ	ERA analysis accession

Data originating from DDBJ:

Accession Prefix	Definition
DRA	DRA submission accession
DRP	DRA study accession
DRX	DRA experiment accession

Table continued from previous page.

Accession Prefix	Definition
DRR	DRA run accession
DRS	DRA sample accession
DRZ	DRA analysis accession

Matching Search Results to SRA Accessions Provided in Publication

How do I know which accessions mentioned in a paper correspond to specific data mentioned in the paper (e.g. which accession represents replicate 1? replicate 2? Illumina? SOLiD?).

The only way you can determine what the SRA accessions provided in a publication actually represent in terms of the data discussed in the publication is to contact the author of the publication and ask.

Currently, there is no requirement for authors to use a specific SRA accession in their publications. If there was, we would recommend that authors use the SRP (study) accession, which would provide the reader/user with a complete overview of the study and a set of links to all the data from that study.

If you are a submitter/author whose paper describes many experiments with identical information, we suggest that you describe how your experiments differ from one another or how the samples used in the experiments differ from each other. If you describe your experiments/samples in this way, a reader/user will be able to more easily differentiate the samples and libraries used in the experiments.

Search Results All Look the Same

Why do all the experiments for a single study all look the same? For example, SRP001156 has 9 different SRX accessions, but when I click on each, the description is the same. Is there a difference between the SRX accessions?

Many of the experiments do have similarities, but you can see some differences if you look closely at the samples used in the experiments as well as the number of runs, spots and base calls. If you still have trouble differentiating between experiments, contact the submitting author. Since SRA is a raw data archive (not curated), submitting authors have the opportunity to provide descriptions of the experiments when they submit, and it is up to them to provide a description that will differentiate one experiment from another in the same study.