Download Guide

Overview

The purpose of this document is to review to users types of data that are available for download from the SRA, how to download datasets of interest, and how to transform the download components into various formats.

Important Notes on Download Facilities

A number of users have asked why SRA does not provide data in their format of interest.

- The SRA does not have the resources to develop format conversions for all possible formats that users may wish. In any case, these formats often have multiple versions and change quickly as new bioinformatics tools and methods become popular.
- Instead, one basic format (SRA) is provided by the Archive for all publicly available data. A toolkit is also provided that supports conversion to several popular formats. The toolkit is also easily extended to supply data in other formats.
- The SRA is a high throughput resource that relies on streaming output. For this reason certain file types that require indexing or that require evaluation of the data stream in order to know how best to compress it cannot be served efficiently. This is the reason that SRF is not supported.
- Users are advised to switch from ftp to aspera for bulk downloads. Aspera provides: faster bandwidth, higher level flow control, user level encryption, and ability to download trees of components.

Related Documents

NCBI Large Data Download Best Practices

Notices

Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government, and shall not be used for advertising or product endorsement purposes.

Conventions Used in this Document

../itemThe location of the item will be specific to the user's installation.

<item>The items inside the angle brackets are user supplied.

[item]The item inside the square brackets is optional.

{item1|item2}The user must select one of the options inside the curly brackets.

Software Version

This guide is current to SRA Toolkit version 2.1.6 released September 10, 2011. Instructions for previous versions of the SRA Toolkit may be different from those provided in this guide.

We recommend that users stay current with SRA Toolkit updates to benefit from feature additions and bug fixes.

The Run Browser

The SRA Run Browser can display sequencing and instrumentation data on a given run. Typically the Run Browser is reached as a click through from Entrez SRA Experiment report. Users may also navigate by entering a run accession directly in the Run Browser.

Filtering and Selection

In the Run Browser, you can filter and subset reads according to certain regular expression pattern matching:

- Sequence substring: one of the biological reads for a spot should contain the substring
  Examples: ATTGGA, ^ATTGGA, ATTGGA$, ATGDNNAT, ATGGA&GCGC
  See "SRA nucleotide search expressions" for more details.
- Name of a spot you are looking for.
  Example: EXWA4RL02G9Z6H
- Name of a spot plus a window in pixels around it.
  Example: EXWA4RL02G9Z6H X=100 Y=100 - will return all spots located within 200 pixels (in X and Y) from a given spot.

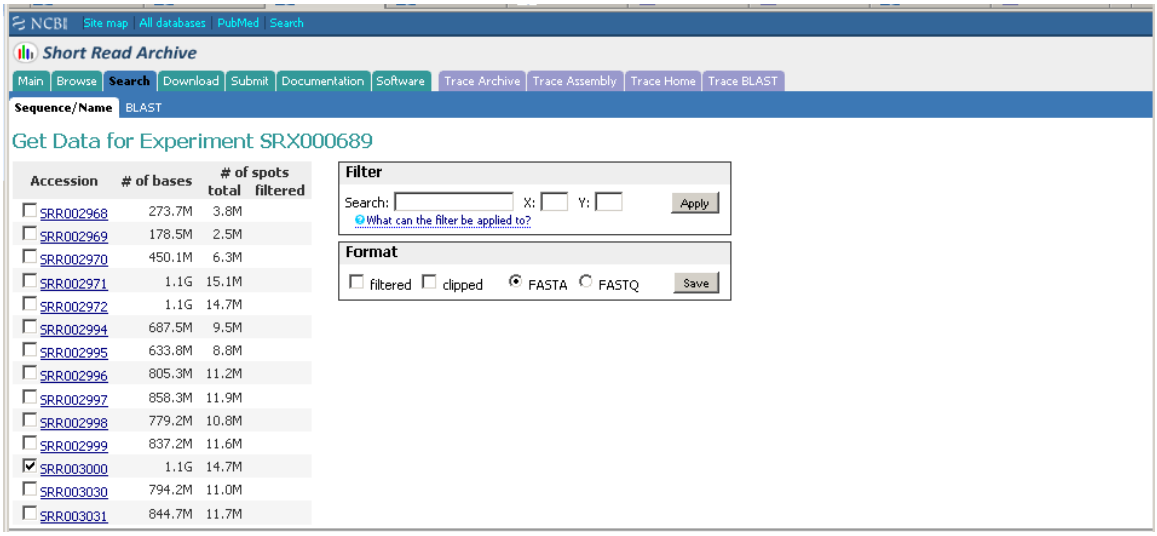
SRA Handbook

SRA Handbook

SRA Handbook

Downloading Data from the Run Browser

You can download data from one or more runs in an SRA Experiment in fasta form and a simple fastq form that has none of the treatments of the static fastq dump. The download dataset will however reflect the filtering and selection you may have performed.



Format of Run Browser Data

The Run Browser supports IUPAC basespace and colorspace base data. The quality scores are in the Phred scale with 0 being at ASCII character 33 or '!'. Reads can be viewed combined or seperated using the customize options in the Run Browser.

Using SRA Data

The SRA Toolkit is provided for accessing the content of SRA archive files.

The SRA Toolkit can be installed from pre-compiled binaries or built from source code using the SRA Software Development Kit

http://ftp-private.ncbi.nlm.nih.gov/sra/sdk

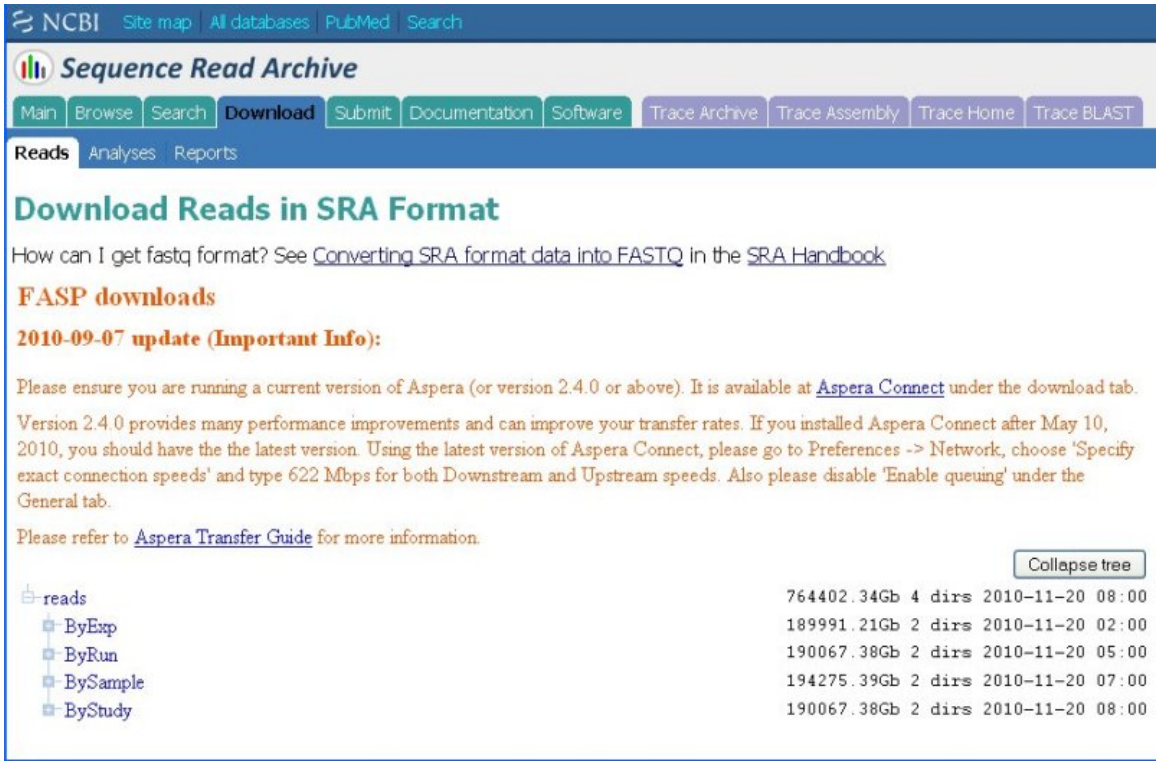
SRA Toolkit Documentation can be located here.

Downloading SRA Data

From a browser like Mozilla Firefox or Internet Explorer users can download SRA Format archives using FTP or Aspera. The URL from which to download SRA Archives is:

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?
cmd=show&f=faspftp_runs_v1&m=downloads&s=download_sra

ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/



Within the SRR (for data submitted to NCBI), ERR (submissions to EBI), and DRR (submissions to DDBJ); users should see a series of numbered sub-directories. To determine which sub-directory the run of interest will be in, take the first six characters of the accession. For example, SRR066661 would be in volume SRR066.

Aspera Connect

Due to the potential for large SRA archive sizes, we recommend that users install and exercise the Aspera Connect client when possible. There is no cost associated with installing the Aspera Connect plug-in. Aspera Connect can be used either as an internet browser plug-in or as the command-line program ascp. Additional information about Aspera can be found in the Aspera Transfer Guide or by visiting Aspera documentation for ascp.

Download by Internet Browser with Aspera Connect Plugin Installed

As an example for using the ascp program, here are the steps to download the data for SRR096072 through an internet browser. ~69KB for the lite.sra

1. Go to http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=download_reads
2. Open the ‘ByRun’ tree by clicking the ‘+’ at the branch point.
3. Open the ‘litesra’ tree by clicking the ‘+’ at the branch point.
4. Open the ‘SRR’ tree by clicking the ‘+’ at the branch point.
5. Open the ‘SRR096’ tree by clicking the ‘+’ at the branch point.
6. Open the ‘SRR096072’ tree by clicking the ‘+’ at the branch point.
7. Click the link for ‘SRR096072.lite.sra’ to download the file.

With the protocol and options removed, the username and location for downloading this run is:

anonftp@ftp-trace.ncbi.nlm.nih.gov:22/sra/sra-instant/reads/ByRun/litesra/SRR/SRR096/SRR096072/SRR096072.lite.sra

This location can be used to download via FTP by adding ftp:// to the beginning.

Download by Command-line ascp

To download SRR096072 using the ascp command program instead, the command in Unix and OSX looks like:

../ascp -i ../asperaweb_id_dsa.putty -L <log directory> -k 1 –QTr -l

200m anonftp@ftp-trace.ncbi.nlm.nih.gov:/sra/sra-instant/reads/ByRun/litesra/SRR/SRR096/SRR096072/SRR096072.lite.sra <save location>

Windows users will keep the / for the remote paths but need to use \ for local paths and quote any paths that contain spaces. For information on the options used in the command, please see the Aspera Transfer Guide. The result of the above command will look like:

SRR096072.lite.sra 100% 69KB 312Kb/s 00:01

Completed: 69K bytes transferred in 1 seconds

(448K bits/sec), in 1 file.

Note that the location used for downloading with ascp is the same as the web browser location but with the port 22 assignment removed. Because the convention for locating the data is conserved, users can build the location themselves if they understand the parts.

To build the location for a Run, the parts are:

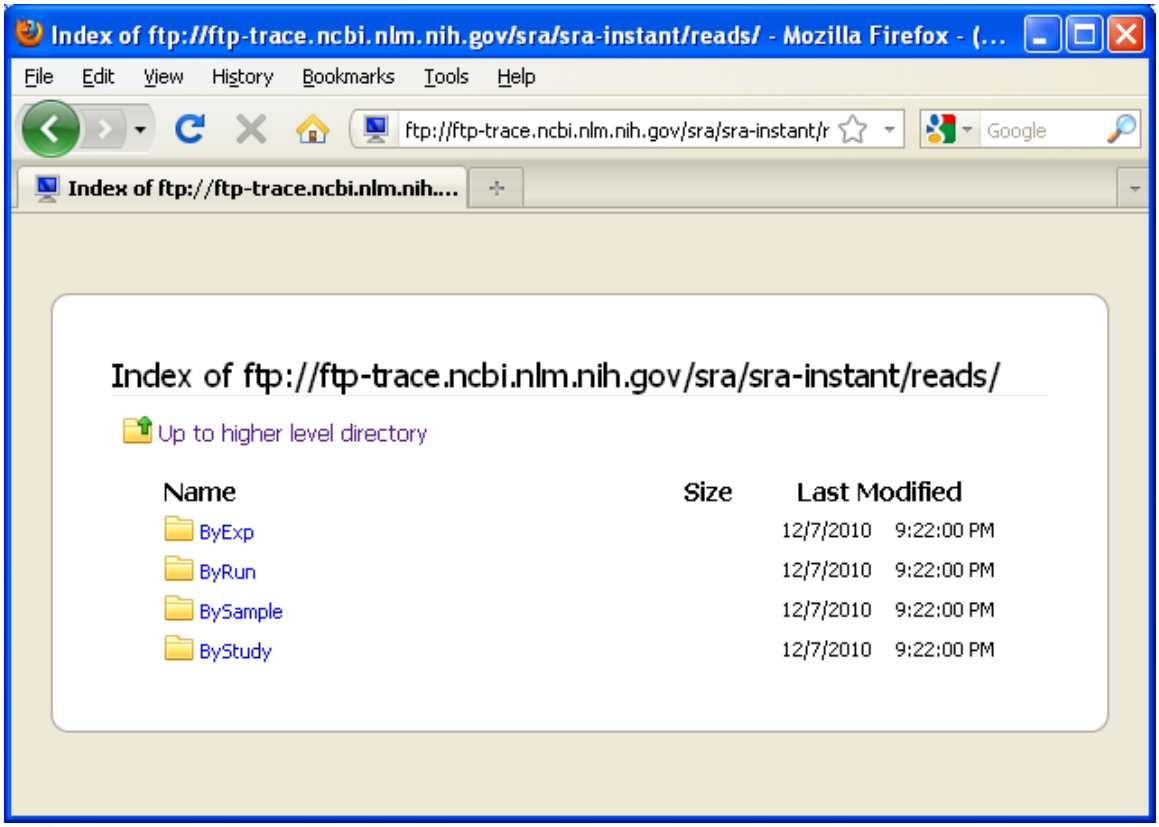
anonftp@ftp-trace.ncbi.nlm.nih.gov:/sra/sra-instant/reads/ByRun/{litesra|sra}/{SRR|ERR|DRR}/<first 6 characters of accession>/<accession>/<accession>.[lite.]sra

For all other object types (Submission, Study, Sample, or Experiment), the format is:

anonftp@ftp-trace.ncbi.nlm.nih.gov:

/sra/sra-instant/reads/By<object type>/{litesra|sra}/<object prefix>/<first 6 characters of accession>/<accession>/<run accession>/<run accession>.[lite.]sra

The object information must agree with each other. For example, experiments will need to use ByExp and accessions with the prefix SRX, ERX, or DRX. If the Run accession is left off the end of the location and the –r switch is used, all sub-directories and files in the directory indicated will be downloaded.



Reference Compression

Compression by Reference is a sequence alignment compression process for storing sequence data. Currently BAM, Complete Genomics, and Illumina export.txt formats contain alignment information. Compression by Reference only stores the difference in base pairs between sequence data and the segments it aligns to. The decompression process to restore original data such as FastQ dump would require fast access to the actual sequences of the references. NCBI recommends that SRA users dedicate local disk space to store local references downloaded from the NCBI SRA site. Linked references should be in a location accessible by the SRA Toolkit software.

Archives that have been compressed by reference bear the .csra extension. Because the reference sequence is not contained in the archive, only the differences, a copy of the reference will be required to decode the compressed archive. The toolkit now contains two programs to assist with this process.

The program config-assistant.pl is used to define the location of the reference sequence files. This progam requires that Perl be installed for the operating system. When config-assistant is run, the default is /home/USERNAME/ncbi/refseq. If you define your own personal directory or a group directory so that multiple users can use the same common references, make sure not to include a slash '/' at the trailing end of the path.

The program reference-assistant.pl is used to download the relevant reference sequences that are using in the compressed archive. The program uses the program wget to download the reference files. An executable binary of wget is included in the install of the SRA toolkit. The reference-assistant will ask for cSRA files to test and will download any reference files used

in the cSRA that are not present in the reference directory that was set by config-assistant.pl previously.

The programs config-assistant.pl and reference-assistant.pl require an installation of Perl to be available to the toolkit.

Once the necessary reference files are present, other toolkit programs like vdb-dump and fastq-dump can be used as normal on the archive.

Additional information about compression by reference is available at the SRA Software Documentation Page or at http://ftp.ncbi.nlm.nih.gov/sra/doc/toolkit_download_and_usage.txt