SRA Download Guide

National Center for Biotechnology Information – National Library of Medicine

Version 1.0 Draft A 09 Sep 2009

1 Contents

2 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 final usable form.

2.1 Important Notes on Download Facilities

A number of users have the question: Why can't I get SRA data in my favorite format ?

- The SRA is an archive of data and does not have the resources to develop format conversions for all possible formats that users may wish. In any case, these formats (and some formats have multiple flavors) 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 some 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 is a superior technology: it provides faster bandwidth, higher level flow control, user level encryption, and ability to download trees of components.

2.2 Revision History

2.3 Related Documents

3 Static fastq Dump Facility

3.1 How to locate static fastq dumps

Currently the SRA provides a static dump of "fastq" form data. These datasets are structured by various SRA views:

Submission - ftp://ftp.ncbi.nlm.nih.gov/sra/Submissions/SRA000/SRA000001/
Study - ftp://ftp.ncbi.nlm.nih.gov/sra/Studies/SRP000/SRP000001/
Sample - ftp://ftp.ncbi.nlm.nih.gov/sra/SeqSamples/SRS000/SRS000002/
Experiment - ftp://ftp.ncbi.nlm.nih.gov/sra/static/SRX000/SRX000001/
Run -
http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=viewer&m=data&s=viewer&run=SRR000001

Given an accession you can construct a canonical path to retrieve the files in each view. Divide the accession by 1000 to get the group branch. If you are looking up a run, access the Run

Browser to retrieve the Experiment, and construct a path using the Experiment subdir. Once you have a path to the component you wish to download, you can give that address to your ftp or aspera client for download.

3.2 Static fastq dump format

The static fastq dump is provided to present to users the sequence of maximum utility. It has been selected, filtered, and substringed in order to provide the usable biological sequence. There are other facilities in SRA that can return the raw, untreated sequence.

Up to three files are produced for each run, containing only biological sequence (linkers and adapters have been removed):

SRR000001.fastq – Fragment library data, or unpaired mates from a paired library.
SRR000001_1.fastq – First mate sequence.
SRR000001_2.fastq – Second mate sequence in the submitted orientation.

Reads are sorted by SRA accession (so all three files can be read concurrently).

The format for each file is the same:

```
@$acc $readname length=$length
$bases
+$acc $readname length=$length
$qualities
```

where

$acc – The SRA accession of the read, for example SRR000001.42
$readname – The original readname reconstructed from the spot address (platform specific)
$bases – The base calls ACTG (or color calls for SOLiD data: 0123 with leading T/G). Note that no call (. Or N) is represented as having quality score 0, rather than printing no call.
$qualities – The quality score in text format. The quality scoring system is log probability of error (commonly referred to as phred), with the value range 0..63 (0 signifies no call), and presented in ascii characters 33..97 with ascii 33 as the basis character (equivalent to 0).

Here is an example record:

```
@SRR000065.1 EN21GUZ01A02PT length=39
GGGGAAAGTGGAGAAGAATCCAGAAGATAGGAGTATCCA
+SRR000065.1 EN21GUZ01A02PT length=39
A91(C@/<<A<<<A<<A;<A;<<A<<;<<A;<<<<<A;<
```

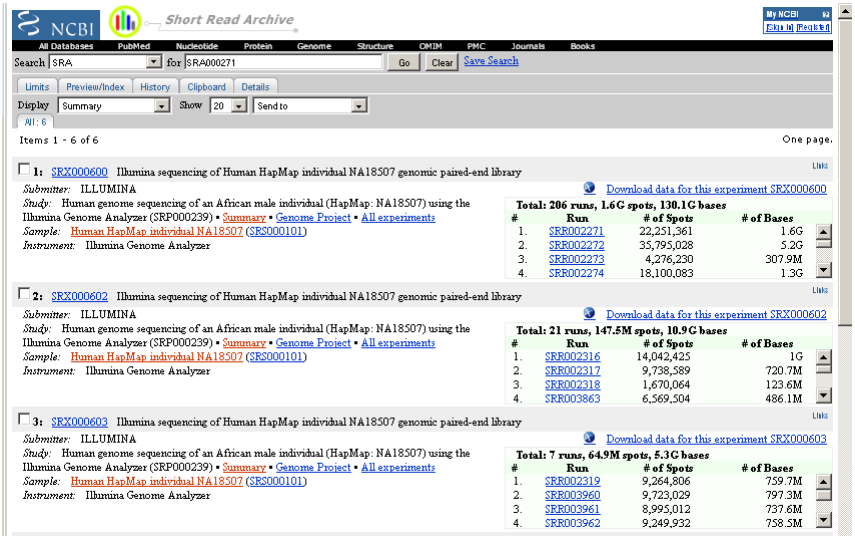
Static fastq data have been subjected to the following treatment:

- Adapters, linkers have been removed when entirely recognized by the SRA
- Mate pairs have been extracted as specified by the run's SRA Experiment spot descriptor and presented in distinct files.
- Fragments and failed mates are presented in a third file (or sometimes the only file).
- Reads less than 25 bp are not presented with this facility
- Reads containing null base calls (signified by quality 0) in their first 25 bp are not presented.
- Left and right clip points are applied yielding a subsequence of biological data that is worth analyzing.

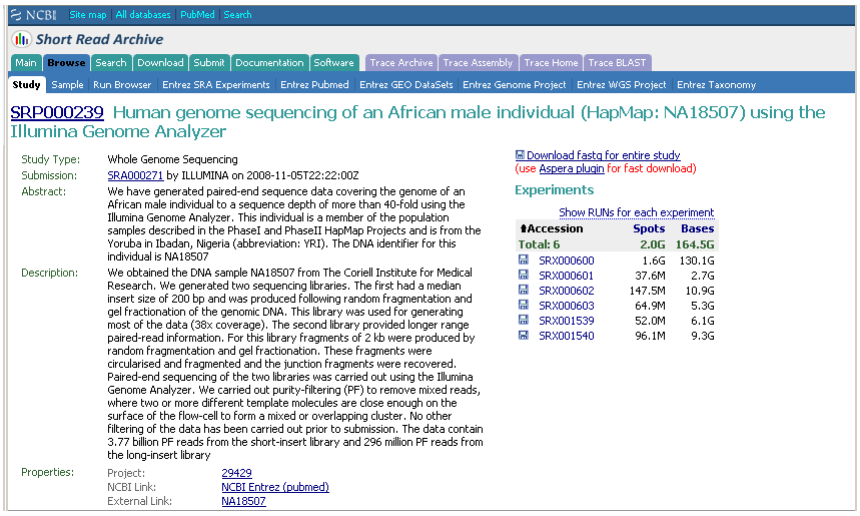
The requirements for this treatment were determined by the 1000 Genomes Project. The separation of fragments/failed mates, and each good mate into separate files is done in order to support certain aligners that consider fragments and mate pairs differently.

3.3 Getting static data from Entrez

When you access an SRA record through the Entrez system, you can land on a particular component and request download of all the sequencing data for that component. The data are provisioned through the static fastq facility. Below is a display of SRA Experiment components returned for a query. Above the run list for each experiment is a download icon that allows you to download via ftp (or aspera, if it's been configured for your browser).



A similar download icon is available from the Study Report, and would allow you to download all the runs in a SRA Study (typically a much bigger dataset). In the Entrez SRA Experiment report, follow the "Summary" link to get the Study Report:



4 The Run Browser

The SRA Run Browser can display sequencing and instrumentation data on a given run. You need to know the run's accession. Typically you invoke the Run Browser as a click through from Entrez SRA Experiment report.

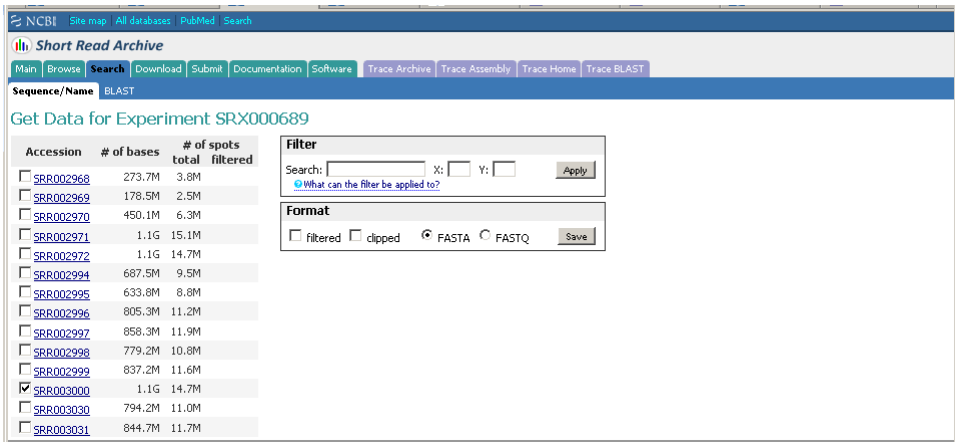
4.1 Filtering and Selection

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.

4.2 Downloading Data from the Run Browser

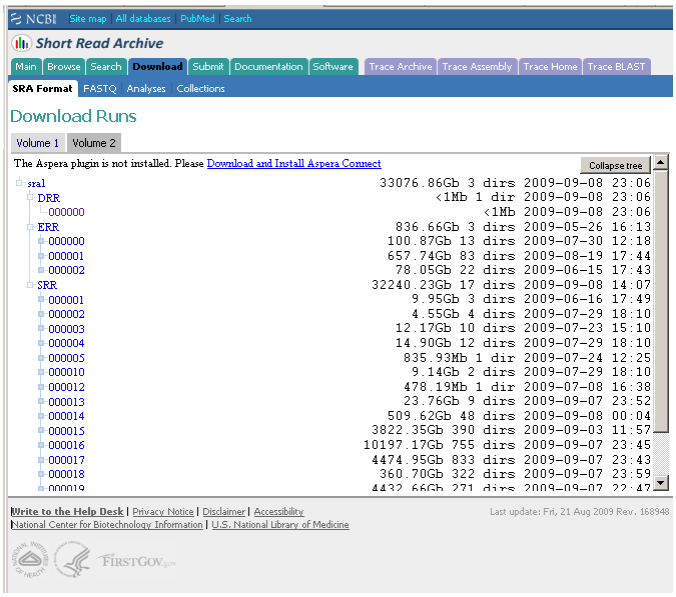
You can download data from one or more runs in an SRA Experiment in fasta form and a simple fastq form that has none of the treatments of the static fastq dump. The download dataset will however reflect the filtering and selection you may have performed.



5 Permanent SRA Objects

Permanent SRA objects amalgamate all the sequencing data including base calls, qualities, intensities, and are stored in the following location, arrayed in batches of 1024 accessions. These objects are the core storage components that make up the SRA, and public access is provided to them so that users may download all or portions of certain runs.

The address of a desired run is computable in a canonical way:



5.1 Conversion to Popular Formats

Content under development.

6 Bulk Data Downloads

Content under development.

7 Bulk Queries using eutils

Content under development.

8 Using ftp and aspera

Content under development.