

SRA Download Guide

National Center for Biotechnology Information – National Library of Medicine

Version 1.0 Draft A 09 Sep 2009

1 Contents

| | |
|---|---|
| SRA Download Guide | 1 |
| Overview..... | 1 |
| Important Notes on Download Facilities..... | 2 |
| Revision History..... | 2 |
| Related Documents | 2 |
| Static fastq Dump Facility | 2 |
| How to locate static fastq dumps | 2 |
| Static fastq dump format..... | 3 |
| Getting static data from Entrez..... | 4 |
| The Run Browser..... | 5 |
| Filtering and Selection | 5 |
| Downloading Data from the Run Browser | 6 |
| Permanent SRA Objects | 6 |
| Conversion to Popular Formats | 7 |
| Bulk Data Downloads | 7 |
| Bulk Queries using eutils | 7 |
| Using ftp and aspera | 7 |

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).

The screenshot shows the NCBI Short Read Archive search results for the query 'SRA'. The search results are displayed in a table format, showing three experiments. Each experiment entry includes a checkbox, the experiment ID, a description, the submitter, the study, the sample, the instrument, and a download icon. Below each experiment entry is a table of runs with columns for Run ID, # of Spots, and # of Bases.

| Experiment ID | Description | Submitter | Study | Sample | Instrument | Total Runs | Total Spots | Total Bases |
|---------------|---|-----------|---|---|--------------------------|------------|-------------|-------------|
| SRX000600 | Illumina sequencing of Human HapMap individual NA18507 genomic paired-end library | ILLUMINA | Human genome sequencing of an African male individual (HapMap: NA18507) using the Illumina Genome Analyzer (SRP000239) • Summary • Genome Project • All experiments | Human HapMap individual NA18507 (SR0000101) | Illumina Genome Analyzer | 206 | 1.6G | 130.1G |
| SRX000602 | Illumina sequencing of Human HapMap individual NA18507 genomic paired-end library | ILLUMINA | Human genome sequencing of an African male individual (HapMap: NA18507) using the Illumina Genome Analyzer (SRP000239) • Summary • Genome Project • All experiments | Human HapMap individual NA18507 (SR0000101) | Illumina Genome Analyzer | 21 | 10.9G | 147.5M |
| SRX000603 | Illumina sequencing of Human HapMap individual NA18507 genomic paired-end library | ILLUMINA | Human genome sequencing of an African male individual (HapMap: NA18507) using the Illumina Genome Analyzer (SRP000239) • Summary • Genome Project • All experiments | Human HapMap individual NA18507 (SR0000101) | Illumina Genome Analyzer | 7 | 5.3G | 64.9M |

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:

NCBI Site map All databases PubMed Search

Short Read Archive

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST

Study Sample Run Browser Entrez SRA Experiments Entrez PubMed Entrez GEO DataSets Entrez Genome Project Entrez WGS Project Entrez Taxonomy

SRP000239 Human genome sequencing of an African male individual (HapMap: NA18507) using the Illumina Genome Analyzer

Study Type: Whole Genome Sequencing
 Submission: [SRA000271](#) by ILLUMINA on 2008-11-05T22:22:00Z
 Abstract: We have generated paired-end sequence data covering the genome of an African male individual to a sequence depth of more than 40-fold using the Illumina Genome Analyzer. This individual is a member of the population samples described in the PhaseI and PhaseII HapMap Projects and is from the Yoruba in Ibadan, Nigeria (abbreviation: YRI). The DNA identifier for this individual is NA18507
 Description: We obtained the DNA sample NA18507 from The Coriell Institute for Medical Research. We generated two sequencing libraries. The first had a median insert size of 200 bp and was produced following random fragmentation and gel fractionation of the genomic DNA. This library was used for generating most of the data (30x coverage). The second library provided longer range paired-read information. For this library fragments of 2 kb were produced by random fragmentation and gel fractionation. These fragments were circularised and fragmented and the junction fragments were recovered. Paired-end sequencing of the two libraries was carried out using the Illumina Genome Analyzer. We carried out purity-filtering (PF) to remove mixed reads, where two or more different template molecules are close enough on the surface of the flow-cell to form a mixed or overlapping cluster. No other filtering of the data has been carried out prior to submission. The data contain 3.77 billion PF reads from the short-insert library and 296 million PF reads from the long-insert library
 Properties: Project: [29429](#)
 NCBI Link: [NCBI](#) [Entrez \(pubmed\)](#)
 External Link: [NA18507](#)

Download fastq for entire study
 (use Aspera plugin for fast download)

Experiments

Show RUNs for each experiment

| Accession | Spots | Bases |
|---------------------------|--------|--------|
| Total: 6 | 2.0G | 164.5G |
| SRX000600 | 1.6G | 130.1G |
| SRX000601 | 37.6M | 2.7G |
| SRX000602 | 147.5M | 10.9G |
| SRX000603 | 64.9M | 5.3G |
| SRX001539 | 52.0M | 6.1G |
| SRX001540 | 96.1M | 9.3G |

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.

NCBI Site map All databases PubMed Search

Short Read Archive

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST

Study Sample **Run Browser** Entrez SRA Experiments Entrez PubMed Entrez GEO DataSets Entrez Genome Project Entrez WGS Project Entrez Taxonomy

Run Browser

Experiment: [SRX000689](#)
 Illumina sequencing of 1000 Genomes Project Pilot 2 NA19238 paired end RANDOM library

Run: Accession: Search
 Alias: 5730
 Instrument model: Illumina Genome Analyzer II
 Date of run: 2008-07-31T20:34:31Z
 Run center: WUGSC

Statistics:
 Number of spots: 14684999
 Number of reads: 29369998

Other:
 Study: [1000Genomes Project Pilot 2](#)
 Design: Illumina sequencing of 1000 Genomes Project Pilot 2 NA19238 paired end RANDOM library
 Platform: ILLUMINA
 Sample: Human HapMap individual NA19238
 Library Name: 2575169269
 Library Strategy: WGS
 Library Source: GENOMIC
 Library Selection: RANDOM
 Library Layout: PAIRED (ORIENTATION=5'-3'Forward, 5'-3'Reverse, NOMINAL_LENGTH=260, NOMINAL_SDEV=0.0E0)

Find spots: X: Y: Find Download View: reads (customize) signals Intensity graph
 What can the filter be applied to?

1. [SRR003000.1](#)
 name: HWI-EAS324_304RG@1:1061:149
 plate: HWI-EAS324_304RG, lane:8, tile:1, x
 2. [SRR003000.2](#)
 name: HWI-EAS324_304RG@1:1027:142
 plate: HWI-EAS324_304RG, lane:8, tile:1, x
 3. [SRR003000.3](#)
 name: HWI-EAS324_304RG@1:1084:136
 plate: HWI-EAS324_304RG, lane:8, tile:1, x
 4. [SRR003000.4](#)
 name: HWI-EAS324_304RG@1:1040:129
 plate: HWI-EAS324_304RG, lane:8, tile:1, x
 5. [SRR003000.5](#)
 name: HWI-EAS324_304RG@1:1040:155
 plate: HWI-EAS324_304RG, lane:8, tile:1, x
 6. [SRR003000.6](#)
 name: HWI-EAS324_304RG@1:986:137
 plate: HWI-EAS324_304RG, lane:8, tile:1, x
 7. [SRR003000.7](#)
 name: HWI-EAS324_304RG@1:1067:155
 plate: HWI-EAS324_304RG, lane:8, tile:1, x
 8. [SRR003000.8](#)
 name: HWI-EAS324_304RG@1:918:130
 plate: HWI-EAS324_304RG, lane:8, tile:1, x

Reads (joined)
 >gnl|SRA|SRR003000.1 HWI-EAS324_304RG:8:1:1061:149
 GTTAAATTTAAGGCTAARATTCGTTAGCTTTTATTTTATTTTTTACCACGAAAATTT
 AATAAAGCCCT

Intensity graph

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.

NCBI Site map / All databases / PubMed / Search

Short Read Archive

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST

Sequence/Name BLAST

Get Data for Experiment SRX000689

| Accession | # of bases | # of spots total | # of spots filtered |
|---|------------|------------------|---------------------|
| <input type="checkbox"/> SRR002968 | 273.7M | 3.8M | |
| <input type="checkbox"/> SRR002969 | 178.5M | 2.5M | |
| <input type="checkbox"/> SRR002970 | 450.1M | 6.3M | |
| <input type="checkbox"/> SRR002971 | 1.1G | 15.1M | |
| <input type="checkbox"/> SRR002972 | 1.1G | 14.7M | |
| <input type="checkbox"/> SRR002994 | 687.5M | 9.5M | |
| <input type="checkbox"/> SRR002995 | 633.8M | 8.8M | |
| <input type="checkbox"/> SRR002996 | 805.3M | 11.2M | |
| <input type="checkbox"/> SRR002997 | 858.3M | 11.9M | |
| <input type="checkbox"/> SRR002998 | 779.2M | 10.8M | |
| <input type="checkbox"/> SRR002999 | 837.2M | 11.6M | |
| <input checked="" type="checkbox"/> SRR003000 | 1.1G | 14.7M | |
| <input type="checkbox"/> SRR003030 | 794.2M | 11.0M | |
| <input type="checkbox"/> SRR003031 | 844.7M | 11.7M | |

Filter
Search: X: Y:
[What can the filter be applied to?](#)

Format
 filtered clipped FASTA FASTQ

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:

NCBI Site map All databases PubMed Search

Short Read Archive

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST

SRA Format FASTQ Analyses Collections

Download Runs

Volume 1 Volume 2

The Aspera plugin is not installed. Please [Download and Install Aspera Connect](#)

| | | Collapse tree | | |
|--------|--|---------------|----------|------------------|
| srsl | | 33076.86Gb | 3 dirs | 2009-09-08 23:06 |
| ERR | | <1Mb | 1 dir | 2009-09-08 23:06 |
| 000000 | | <1Mb | | 2009-09-08 23:06 |
| ERR | | 836.66Gb | 3 dirs | 2009-05-26 16:13 |
| 000000 | | 100.87Gb | 13 dirs | 2009-07-30 12:18 |
| 000001 | | 657.74Gb | 83 dirs | 2009-08-19 17:44 |
| 000002 | | 78.05Gb | 22 dirs | 2009-06-15 17:43 |
| SRR | | 32240.23Gb | 17 dirs | 2009-09-08 14:07 |
| 000001 | | 9.95Gb | 3 dirs | 2009-06-16 17:49 |
| 000002 | | 4.55Gb | 4 dirs | 2009-07-29 18:10 |
| 000003 | | 12.17Gb | 10 dirs | 2009-07-23 15:10 |
| 000004 | | 14.90Gb | 12 dirs | 2009-07-29 18:10 |
| 000005 | | 835.93Mb | 1 dir | 2009-07-24 12:25 |
| 000010 | | 9.14Gb | 2 dirs | 2009-07-29 18:10 |
| 000012 | | 478.19Mb | 1 dir | 2009-07-08 16:38 |
| 000013 | | 23.76Gb | 9 dirs | 2009-09-07 23:52 |
| 000014 | | 509.62Gb | 48 dirs | 2009-09-08 00:04 |
| 000015 | | 3822.35Gb | 390 dirs | 2009-09-03 11:57 |
| 000016 | | 10197.17Gb | 755 dirs | 2009-09-07 23:45 |
| 000017 | | 4474.95Gb | 833 dirs | 2009-09-07 23:43 |
| 000018 | | 360.70Gb | 322 dirs | 2009-09-07 23:59 |
| nnnn19 | | 4432.66Gb | 271 dirs | 2009-09-07 23:47 |

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility
 National Center for Biotechnology Information | U.S. National Library of Medicine

Last updated: Fri, 21 Aug 2009 Rev. 168948

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.