



Download Guide

1 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.

1.1 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.

1.2 Related Documents

NCBI Large Data Download Best Practices

1.3 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.

1.4 Conventions Used in this Document

`../item`The 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.

1.5 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.

2 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.

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

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: 2675169269
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)

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

Find spots: X: Y: View: reads (customize) signals intensity graph

[What can the filter be applied to?](#)

1. [SRR003000.1](#)
name: HWI-EAS324_304RG:8:1:1061:149
plate: HWI-EAS324_304RG, lane:8, tile:1, x

2. [SRR003000.2](#)
name: HWI-EAS324_304RG:8:1:1027:142
plate: HWI-EAS324_304RG, lane:8, tile:1, x

3. [SRR003000.3](#)
name: HWI-EAS324_304RG:8:1:1084:136
plate: HWI-EAS324_304RG, lane:8, tile:1, x

4. [SRR003000.4](#)
name: HWI-EAS324_304RG:8:1:1040:129
plate: HWI-EAS324_304RG, lane:8, tile:1, x

5. [SRR003000.5](#)
name: HWI-EAS324_304RG:8:1:1040:155
plate: HWI-EAS324_304RG, lane:8, tile:1, x

6. [SRR003000.6](#)
name: HWI-EAS324_304RG:8:1:986:137
plate: HWI-EAS324_304RG, lane:8, tile:1, x

7. [SRR003000.7](#)
name: HWI-EAS324_304RG:8:1:1067:155
plate: HWI-EAS324_304RG, lane:8, tile:1, x

8. [SRR003000.8](#)
name: HWI-EAS324_304RG:8:1:918:130
plate: HWI-EAS324_304RG, lane:8, tile:1, x

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

Intensity graph

2.1 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.

2.2 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.

Accession	# of bases	# of spots total	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	

2.3 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 separated using the customize options in the Run Browser.

3 Installing the Toolkit

The SRA Toolkit can be downloaded from the Software section of the SRA website.

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

The utility is available pre-compiled in 32-bit and 64-bit forms for execution on Linux and Mac systems as well as a 32-bit Windows version. Once you download the toolkit in the desired form, you can decompress using one of these commands from a shell command line:

```
tar xvfz <toolkit file>
```

4 Downloading SRA Format Data

4.1 Changes to SRA FASTQ Dumps

Due to storage space constraints, SRA will be forced to no longer maintain the static FASTQ data dumps for all published submissions. As a result, users should download and install the SRA toolkit so that they can produce their own FASTQ files for submissions from the SRA archive format.

4.2 Downloading SRA Format 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

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.

4.3 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`.

4.3.1 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.

4.3.2 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.

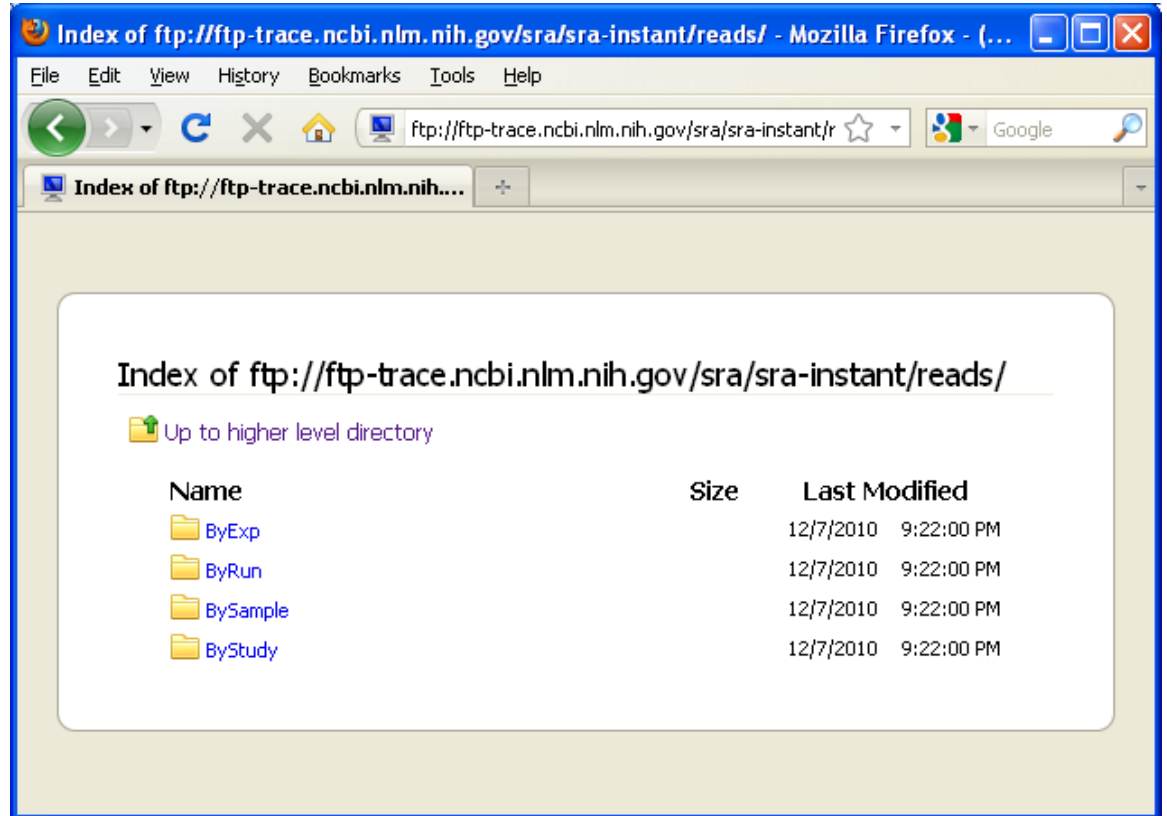
4.4 FUSE Instant Files

SRA has added an on demand file system (FUSE) that will allow the archive to produce .lite.sra files designed for users only wishing to dump FASTQ files. There are issues that prevent SRA from being able to provide FASTQ files on demand in the same format that the static FASTQ dumps have previously been presented in. Users should find that file transfers will be quicker from SRA with the .lite.sra files due to a significant space savings over FASTQ and other text

formats. In addition, there is no need to navigate multiple volumes to locate the storage volume used at NCBI for the run.

The address for SRA FUSE downloads by FTP is:

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



5 Converting SRA format data into FASTQ

5.1 Program Name: fastq-dump

5.2 Synopsis:

```
../bin64/fastq-dump <SRA archive file>
```

5.3 Example:

toolkit installed in: `../sra/bin64/`

downloaded archive in: `../sra/download/SRR000299/`

command to convert the SRA object into fastq data would be:

```
../sra/bin64/fastq-dump ../sra/download/SRR000299/SRR000299.sra
```

Output will be 'SRR000299.fastq' in current directory

This will also work for SRA archives that have been extracted out into their directory structure (contains sub-directories 'col' and files md5, meta, sealed...) using the SRR##### directory as the location.

```
../sra/bin64/fastq-dump ../sra/download/SRR000299/
```

5.4 Options:

The options described in this section are available to modify the behavior of the 'fastq-dump' utility.

Command	Description
Input	
'-A' or '--accession'	Useful when the accession cannot be determined directory from the archive. Also enables modification of the output name used for the fastq files. For example: fastq-dump -A foo SRR000001 Will produce files named 'foo.fastq', 'foo_1.fastq', and 'foo_2.fastq'
'--table <table-name>'	Table name within SRA format, default is SEQUENCE
Processing	
'--split-spot'	Split spots into the individual reads annotated by the submitter or listed within the data file for file types that describe read structure internally.
'-N' or '--minSpotId'	Minimum spot number at which to start the dump process
'-X' or '--maxSpotId'	Maximum spot number at which to stop the dump process For example: fastq-dump -N 5 -X 10 SRR000001 This command will dump six spots starting from spot 'SRR000001.5' and ending in spot 'SRR000001.10'. Filtered spots can result in less than (maxSpotId - minSpotId + 1) total spots output.
'--spot-groups <[list] >'	Filter by SPOT_GROUP (member): name[...]
'-W' or '--clip'	Enables clipping of a spot sequence based on the right clip information. Toggling 'show-clipped' in the 'customize' area for reads in the SRA Run Browser enables observing the effect of this option (e.g. see SRR000001).
'-M' or '--minReadLen'	Allows specification of the desired minimum read length to output. Reads shorter than the provided number will be skipped.
'-R' or '--read-filter <[filter]>'	Split into files by READ_FILTER value or optionally filter by a value: pass reject criteria redacted
'-E' or '--qual_filter'	This option enables quality filtering based on leading/trailing low quality values. As reads have become longer this option has become less important. This filter was used in early 1000 Genomes fastq files and includes no sequences starting or ending with >= 10N
'--skip-technical'	Filter to remove annotated technical reads. This will only remove the technical reads that are provided in the submission metadata or listed in files that describe the read structure. This option does not search for and identify multiplex identifiers or primers in the spots. The --split-spot option must be enabled when using this.
Output	
'-O' or '--outdir'	Indicates the directory where the fastq result should be placed, default is the current directory.
--split-files	Dump each read into a separate file. Files will received a suffix corresponding to the read number
--split-3	Legacy 3-file splitting for mate-pairs used for the 1000 Genomes fastq files. First 2 biological reads satisfying dumping conditions are placed in files *_1.fastq and *_2.fastq. If only 1 biological read is dumpable - it is placed in *.fastq Biological reads 3 and above are ignored.

-G or '--spot-group'	Boolean option that results in fastq files divided into spot groups as defined in the Experiment (or eventually Run) xml. This command: fastq-dump -G SRR051894 Produces these five fragment files: SRR051894.fastq SRR051894_GDSX2KN04_PSORIASISMDA-POOL-738_CB028-01WG.fastq SRR051894_GDSX2KN04_PSORIASISMDA-POOL-738_CB036-01WG.fastq SRR051894_GDSX2KN04_PSORIASISMDA-POOL-738_CD021-01WG.fastq SRR051894_GDSX2KN04_PSORIASISMDA-POOL-738_CD036-01WG.fastq
'-T' or '--group-indirs'	Boolean option directing the utility to produce fastq files in sub-directories rather than producing files within the same directory
'-K' or '--keep-empty-files'	Has no effect - at one time this option would represent all three possible files even if one or two were empty
Formatting	
'-C' or '--dumps' <[cskey]>	Forces color space sequence to be dumped instead of base space. If the optional 'cskey' is provided (i.e. A, C, T, or G), then all fastq files produced will use that key at the start of each color space sequence. (default for SOLiD data)
'-B' or '--dumpbase'	Forces base space sequence to be dumped instead of color space. (default for data other than SOLiD)
'-Q' or '--offset'	Allows using a different offset value to represent a different offset character in the fastq output. For example, using an offset of 64 represents using '@' as the offset character. Default offset is 33.
'--fasta'	Produces fasta files without quality scores
Define	
'-F' or '--origfmt'	Results in fastq containing only the original identifier on the define (i.e. no length or SRR identifier are present)
'-I' or '--readids'	Appends a read index to the run identifier starting with '1' as the first index. Note that this differs from the spot descriptor in the Experiment xml where the read indices start with '0'. In the case of SRR000001, the first spot in each file would have the identifiers 'SRR000001.5.4', 'SRR000001.1.2', and 'SRR000001.1.4'. Note that the first spot sequence in SRR000001.fastq, the fragment file, comes from the second biological/application read which has an index of '4'.
'--define-seq' <fmt>	Allows specification of the sequence define format. For example: -DB "@\$ac.\$si \$sn length=\$rl" This specification produces the same output as the default output. See Appendix D for a more in-depth explanation. Note that submission of a 'fastq-dump' command to a compute farm (e.g. Sun Grid Engine) can require preceding a number of the characters with backslash characters when using this option. The above example might require this version: -DB "@\\\$ac.\\\$si \\\$sn length=\\\$rl"
'--define-qual' <fmt>	Allows specification of the quality define format. For example: -DQ "+\$ac.\$si \$sn length=\$rl" <fmt> is string of characters and/or variables. Variables could be are: \$ac - accession, \$si - spot id, \$sn - spot name, \$sg - spot group (barcode), \$sl - spot length in bases, \$ri - read number, \$rn - read name, \$rl - read length in bases. '[']' can be used for optional output: if all variables in [] yield empty values the whole group is not printed. Empty value is an empty string for text items or 0 for numeric variables. Ex: @\$sn[\$rm]/\$ri - '\$_rn' is omitted if there is no read name
'--helicos'	Helicos style fastq. Omits the accession and length in the base define. Omits the entire quality define.
'-v' or '--version'	Display the version of the program and then quit
'-L' or '--log-level' <level>	Logging level as number or enumeration string. One of (fatal sys int err warn info) or (0-5) is required. Current/default is warn
'-v' or '--verbose'	Increases the verbosity level of the program. Use multiple times for more verbosity.

5.5 Basic Execution of 'fastq-dump' Utility

When executing the 'fastq-dump' utility, the complete path to the utility must be used unless the utility can be located using the current Linux path environmental variable. In its simplest form, the 'fastq-dump' utility executes as follows:

```
fastq-dump <accession archive>
```

The ‘accession archive’ must be located in the current working directory. For example, with SRA format data for SRR000001 in the current working directory:

```
fastq-dump SRR000001.sra
```

To match the format for the legacy 1000 Genome fastq dumps, the following command is used.

```
Fastq-dump --split-3 SRR000001.sra
```

In this case, the result will be the creation of three files in the same directory:

```
SRR000001.fastq
```

```
SRR000001_1.fastq
```

```
SRR000001_2.fastq
```

The file, ‘SRR000001.fastq’, contains fragment read sequences where only a single biological/application read exists (or remained after filtering) for a spot. The first spot in the file will look like this:

```
@SRR000001.6 EM7LVYS01EAHKZ length=250
```

```
ATTCAAACCCTTTTCGGTTCCAACATTTTAGTTTTTCTTTTAACTTTAGAAAACCTGCCTGGATTTCAG
TTATTCATCCATAGCTTGAATATGTCATTTTTCTTACTGCCCTTCTCTCCTCTCAACAACCACTAGTC
ATTCCCTGTAAAAAGTTTGGGAACATTCACAGACTTTTACAGCAGAAAAGTATATAATGTGGTGGCAG
AGAAGCCCTCTTCTCAAAGTCAGTATATTTCTACCACGTGGCGATA
```

```
+SRR000001.6 EM7LVYS01EAHKZ length=250
```

```
<B:9D@,C?+B>)=B:C=D=C==<GC6)<<GC8-"=GC6(>6=FB0=<EA4$<C<:C<<C=<FA/
=<;C==C==<;C=<;==<B;:@8:9<<==<GC8.#8C=;<=<FB0B=;<;C; ;<;@7<@8A:=:<<
<=<:B:FA/==;HD90&:FB0FB09/
<<D=9@9;:=<=HD8, ;<<<<6:FB1:=7<8<C;:=<C=;C<736==C=<FB2<<C=6<<FB1 ; ;8:<
*'&<? ;#)<5B:3*0;A;3:<=4
```

Looking at spot SRR000001.6 in the SRA Run Browser shows that the first application read only has a length of seven, causing the utility to exclude it from the fastq output.

The file, ‘SRR000001_1.fastq’, contains the first read in spots containing paired reads and the file, ‘SRR000001_2.fastq’, contains the second read. The first spot in ‘SRR000001_1.fastq’ will look like this:

```
@SRR000001.1 EM7LVYS01C1LWG length=88
```

```
GGGGGAGCTTAAATTTGAAACTAGAAAAATTTTGAACAAAATAATCATAATTGTTAGCTGATGAAAAA
CTAGAAAAGATTTTCTGAGT
```

```
+SRR000001.1 EM7LVYS01C1LWG length=88
```

```
C91*#==<C=EA.EA/<B=(<<:=HC90'FB5&;B:<GC6
(=D=<<==C=C==B<=<<<=<;<<GC8.#<<9=FB4%<8EA4%87:<<8
```

The first spot in ‘SRR000001_2.fastq’ will look like this:

```
@SRR000001.1 EM7LVYS01C1LWG length=99

GGTATCCCGTAGTGTGCATTCATCCCTGCTCTGGATACAGTCAGCTCCCAAATTCATAAACAACTCC
TTTGTAAGTAACCTCCTTTTGACAGGGGTA

+SRR000001.1 EM7LVYS01C1LWG length=99

B:<|=C?
+<<|<===<=|C<===<FB0=<===<<D=9=|;|=<=<=<|=FB2FB2C<C<|=FB0<C==|C<D@-
<=B:<=C=C|<C=GD7*|=;|=HD90'==
```

If only fragment reads exist for a run, then only one file is generated (e.g. SRR000700). If only paired reads exist for a run, then only the two pair files result (e.g. SRR029338).

To generate fastq for a run archive that is not in the current working directory, a path to the data must be specified:

```
fastq-dump <path to run archive>
```

In this situation, 'path to run archive' can be a complete path or a path relative based on the current working directory.

Default behavior associated with the fastq-dump utility includes:

- 1 Dumping color space sequence for ABI submissions and base space otherwise
- 2 Using a quality offset of 33 (or the '!' character)
- 3 Read names using the SRA run accession.
- 4 Including technical reads as well as all bases in the archive regardless of quality score.

6 Converting SRA format data into SFF

6.1 Program Name: sff-dump

6.2 Synopsis:

```
../bin64/sff-dump -A <SRR_accession> <Path_to_SRR_Directory>
```

6.3 Example:

toolkit installed in: ../sra/bin64/

downloaded SRR in:

```
../sra/download/SRR000299/
```

command to convert the SRA object into a SFF file would be:

```
../sra/bin64/sff-dump -A SRR000299 ../sra/download/SRR000299/
```

This will also work for single-file SRA archives (example SRR000299.sra located in ../sra/download/) using the single file as the directory.

```
../sra/bin64/sff-dump -A SRR000299 ../sra/download/SRR000299.sra
```

6.4 Options:

Command	Description
-A, --accession	Accession to be used in the file output
-O, --outdir	Allows user to specify an output directory. If not used, output will default to the current directory.
-N, --minSpotId	Minimum spot ID to output. The first spot in the output will be the number given for this option.
-X, --maxSpotId	Maximum spot ID to output. The last spot in the output will be the number given. Min and Max spot options can be combined to output subsections of an SRR.
-G, --spot-group	spotgroup-file Split into files by SPOT_GROUP
-GL, --spot-group-list	Filter by SPOT_GROUP (member).
-R, --read-filter	Split into files by READ_FILTER value.
-T, --group-in-dirs	spotgroup-dir Split into subdirectories (of -O) by SPOT_GROUP
-L, --log-level	Log level: 0-13 or fatal sys int err warn info. (default: err)
-H, --help	Prints this help message and version information.

6.5 Description:

The sff-dump utility can be used to convert SRA format data into the SFF (Standard Flowgram Format) file type used by Newbler and other analysis software packages. The file output from the sff-dump utility will not have a manifest or index which may cause issues with the file's usage. To correct this, the file can be rebuilt using the sfffile utility available from 454. An example command for using the sfffile utility for the previously created sff file:

```
sfffile -o SRR000299.sff SRR000299.sff
```

The .lite.sra type files will not work with the sff dump utility. An example of trying to use the .lite.sra archive type with the sff-dump utility is the following:

```
2010-10-22 16:22:33 ../bin64/sff-dump ERROR: column not found while opening
table
within virtual database module - Failed
2010-10-22 16:22:33 ../bin64/sff-dump FATAL: SIGNAL - Segmentation fault
Aborted
```

7 Converting SRA format data into Illumina Native Files

7.1 Program Name: illumina-dump

7.2 Synopsis:

```
../bin64/illumina-dump [options] <Path_to_SRR_Directory>
```

7.3 Example:

toolkit installed in: ../sra/bin64/

downloaded SRR in:

```
../sra/download/SRR000299/
```

command to convert the SRA object into Illumina native files would be:

```
../sra/bin64/illumina-dump -A SRR000299 ../sra/download/SRR000299/
```

This will also work for single-file SRA archives (example SRR000299.sra located in ../sra/download/) using the single file as the directory.

```
../sra/bin64/illumina-dump -A SRR000299 ../sra/download/SRR000299.sra
```

7.4 Options:

Command	Description
-A, --accession	Accession to be used in the file output.
-O, --outdir	Output directory. Default: '.'
-N, --minSpotId	Minimum spot id to output.
-X, --maxSpotId	Maximum spot id to output.
-G, --spot-group	Split into files by SPOT_GROUP (member).
-GL, --spot-group-list	Filter by SPOT_GROUP (member): name[,...].
-R, --read-filter	Split into files by READ_FILTER value, optional filter by a value: pass reject criteria redacted.
-T, --group-in-dirs	Split into subdirectories instead of files.
-K, --keep-empty-files	Do not delete empty files.
-L, --log-level	Logging level: fatal sys int err warn info. Default: err
-H, --help	Prints this message

Format options:

Command	Description
-r, --read	Output READ: "seq". Default: on
-q, --qual1	Output QUALITY, into single (1) or multiple (2) files: "qcal". Default: 1
-p, --qual4	Output full QUALITY: "prb". Default: off
-i, --intensity	Output INTENSITY, if present: "int". Default: off
-n, --noise	Output NOISE, if present: "nse". Default: off
-s, --signal	Output SIGNAL, if present: "sig2". Default: off
-qseq	Output QSEQ format: whole spot (1) or split by reads (2 - default): "qseq". Default: off

7.5 Description:

SRF files can be generated by a two step process of dumping the native Illumina files and then converting the native files into SRF using the illumina2srf utility in io_lib.

An example for the command to use for Illumina2srf v1.12:

```
../illumina2srf -R -P -N read_name:%l:%t: -n %x:%y -o output.srf s_*seq.txt
```

'read_name' and 'output.srf' can be changed to differentiate reads and files. 's_*seq.txt' assumes the original source file names and may need to be changed for Illumina files dumped from the SRA format.

8 Converting SRA format data into ABI-SOLiD Native Files

8.1 Program Name: abi-dump

8.2 Synopsis:

```
abi-dump [options] [ -A ] <accession>
```

8.3 Example:

toolkit installed in: ../sra/bin64/

downloaded SRR in:

```
../sra/download/SRR000299/
```

command to convert the SRA object into SOLiD native files would be:

```
../sra/bin64/abi-dump -A SRR000299 -M 0 ../sra/download/SRR000299/
```

This will also work for single-file SRA archives (example SRR000299.sra located in ../sra/download/) using the single file as the directory.

```
../sra/bin64/abi-dump -A SRR000299 -M 0 ../sra/download/SRR000299.sra
```

8.4 Options:

Command	Description
-A, --accession	Accession to be used in the file output.
-O, --outdir	Output directory. Default: '!'
-N, --minSpotId	Minimum spot id to output.
-X, --maxSpotId	Maximum spot id to output.
-G, --spot-group	Split into files by SPOT_GROUP (member).
-GL, --spot-group-list	Filter by SPOT_GROUP (member): name[,...].
-R, --read-filter	Split into files by READ_FILTER value, optional filter by a value: pass reject criteria redacted.
-T, --group-in-dirs	Split into subdirectories instead of files.
-K, --keep-empty-files	Do not delete empty files.
-L, --log-level	Logging level: fatal sys int err warn info. Default: err
-H, --help	Prints this message

Format options:

Command	Description
-M, --minReadLen	Minimum read length to output. Default: 25
-W, --noclip	Do not clip quality left and right for spot.
-F, --origfmt	Excludes SRR accession on defline.
-B, --noDotReads	Do not output reads consisting mostly of dots.

8.5 Description:

This program will output the .csfasta and _QV.qual file types that SRA refers to as SOLiD native. These files are in color space with a key base (typically T) provided for each read. By default the reads will be named based on the run accession at SRA. The option `-F` can be used to generate read names that do not contain the run accession.

9 Compression by Reference

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 program 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](#).

Appendix A – Toolkit Compiled Binary Components

This list contains the binary files that are useful for operating on archives that have been downloaded from the SRA archive. There are additional compiled binaries included that are used for converting source data into the SRA archive format.

Program Name	Program Function
abi-dump	convert sra format archive data into SOLiD native data

fastq-dump	convert sra format archive data into fastq file types
illumina-dump	convert sra format archive data into illumina native file types
sff-dump	convert sra format archive data into sff files
kar	interconverts monolithic file .sra archives and directory based sra archive formats
sra-dump	view sra format archive data as text
vdb-dump	view sra format archive data as text
sra-stat	view statistics for an archive