

Getting the data from SRA

The previous section helps us to locate the samples. We now need to get the data out of SRA in the form of a set of FASTQ files.

On the webpage <http://www.ncbi.nlm.nih.gov/biosample/2999518>, in the top right corner there is a header called “Related Information”, with a link to SRA. Clicking on that link takes us to an SRA page

http://www.ncbi.nlm.nih.gov/sra?LinkName=biosample_sra&from_uid=2999518. Here we see important information. First, internally in SRA this BioSample is called SRX683793; two Runs with ids SRR1554535 and SRR2071346 are associated with this sample. The ids beginning with SRX are called experiment ids and the ones beginning with SRR are called run ids. In this case it means that this particular sample was sequenced twice. Each sequencing run will give us a FASTQ file, and we will therefore have two FASTQ files associated with this sample.

Side note: internally in SRA all data is stored in a special SRA format, which is - frankly - irritating to deal with. But it allows us to potentially retrieve the data in FASTA, FASTQ and SRA formats. We want to get the FASTQ files.

Clicking on either of the Run ids at the bottom of the page takes us to <http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1554535>, a screen shot is below.

NCBI Site map All databases Search

Sequence Read Archive

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

Studies Samples Analyses Run Browser Selector Provisional SRA

RNAseq of human DLPFC polyA+ transcriptome: Sample R2869_DLPFC (SRR1554535) Change accession...

Metadata Alignment Reads Download

Please:

Use [SRA Toolkit](#) tools to directly operate on SRA runs. Toolkit has capacity to find requested runs at NCBI and download (and cache) only the part you really need. For example quality scores represent a majority of data volume and you may not need them if you dump fasta only (versus fastq). Or if you are looking at particular gene you may not need the reads aligned to other regions or not aligned at all.

Use SRA Toolkit [prefetch](#) utility if you want to cache all data in advance (for example in case your processing cluster does not connect to internet). Read more at [Downloading SRA data using command line utilities](#).

Use SRA Run Selector to filter and download a list of SRA runs in the scope of [experiment](#), [sample](#) and [study](#)

How can I get fastq format? See [Converting SRA format data into FASTQ](#) in the [SRA Toolkit Documentation](#)

In addition to it you can download the following data:

1. [SRR1554535.pileup](#) 71f646b485ed923e470fe4fc58830a88

Write to the Help Desk | Privacy Notice | Disclaimer | Accessibility

Last update: Fri Jan 8 10:16:37 EST 2016

National Center for Biotechnology Information | U.S. National Library of Medicine

NIH FIRSTGOV

Confusingly, the page you land at is labeled “download” but you can only download a pileup file. Instead click at the top, at the other (!) download tab. Here you get taken to a page where you can only enter experiment ids, beginning with SRX. Doing this, lands you on a page like this

<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?exp=SRX683792&cmd=search&m=downloads&s=seq>, which allows you to download FASTQs.

An alternative to the web interface is to use the sra toolkit, a command line utility you can find at <http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>. There are several command line utilities for accessing SRA. A classic is the fastq-dump command; an example of using it is

```
fastq-dump -v --gzip SRR1554534
```

This will output a gzipped FASTQ file. The utility only appears to support run ids, not experiment ids.