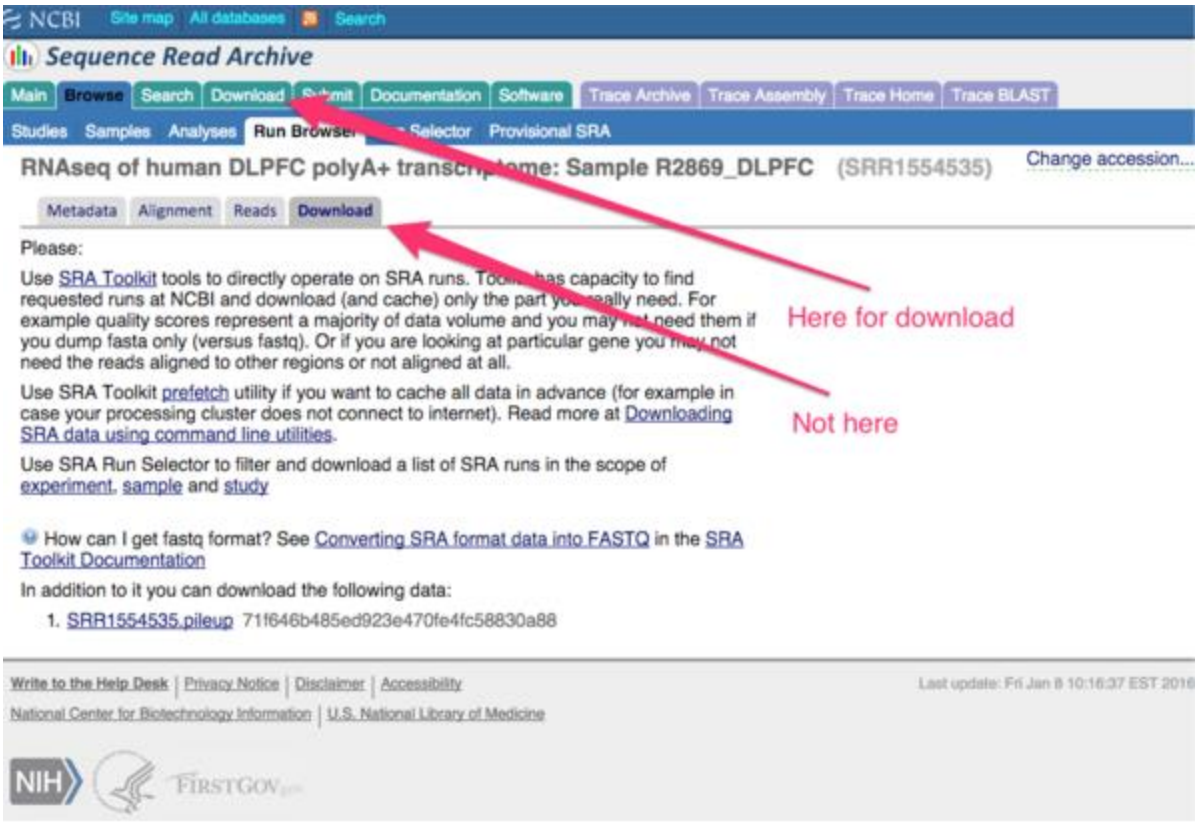
Getting the data from SRA

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.



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.