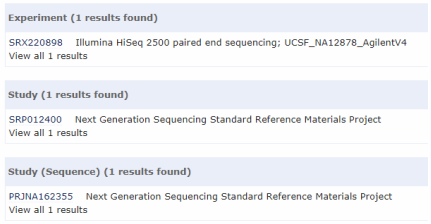
Short summary of layout of NGS data
S. Ravichandran, PhD. (modified on 08/14/2018)

We are focusing on the NGS data in NCBI. Sequence Read Archive is the database that is associated with NGS data. The data layout is very complicated (at least for me ☺). Let us look at a project and try to understand how the data is stored in NCBI.

Here I had gone to the ENA database (relevant European SRA database; https://www.ebi.ac.uk/ena/data/search?query=PRJNA162355 to get the left side result figure.



Project: PRJNA162355, https://www.ncbi.nlm.nih.gov/bioproject/162355

PRJNA162355 has the following two BioSamples:

https://www.ncbi.nlm.nih.gov/biosample/801888
https://www.ncbi.nlm.nih.gov/biosample/1696

Study: SRP012400, https://www.ncbi.nlm.nih.gov/sra/?term=SRP012400

This project has sequence data in two different databases:

72 Experiments (SRA Data)

1) SRX1608032: https://www.ncbi.nlm.nih.gov/sra/SRX1608032[accn]
2) SRX1608029: https://www.ncbi.nlm.nih.gov/sra/SRX1608029[accn]
   and so, on

one more dataset in Trace Archive

https://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=retrieve&val=ncbi_project_id=162355

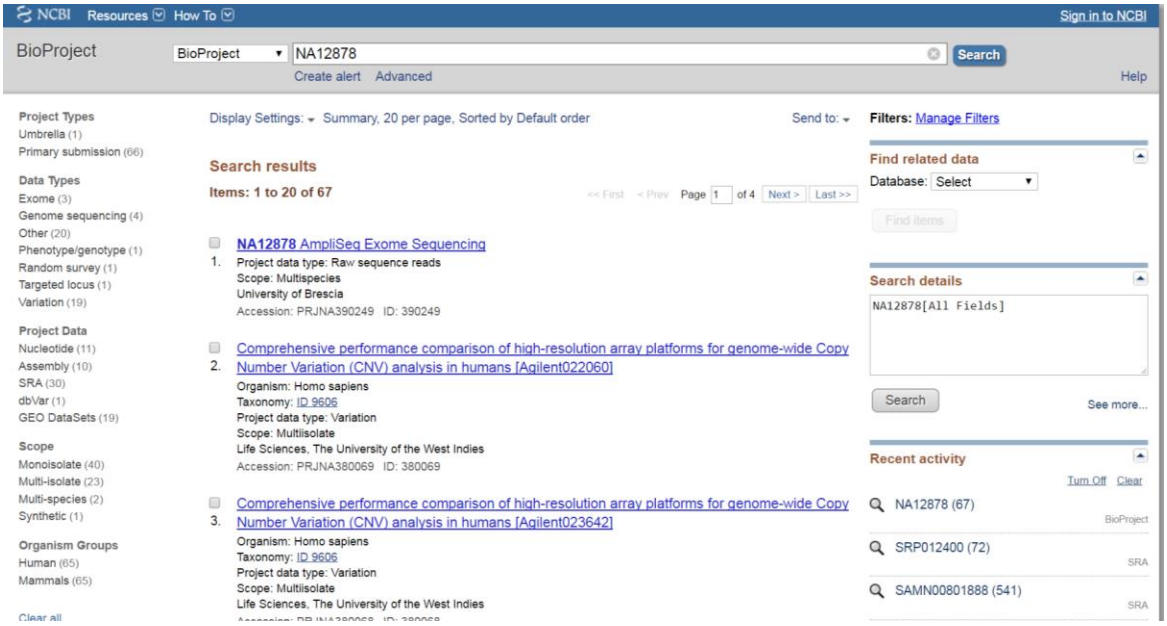
Experiment, SRX1608029, contains one run, SRR3197790

Experiment, SRX1608029, contains 4 runs, SRR3197783 .. SRR3197786.

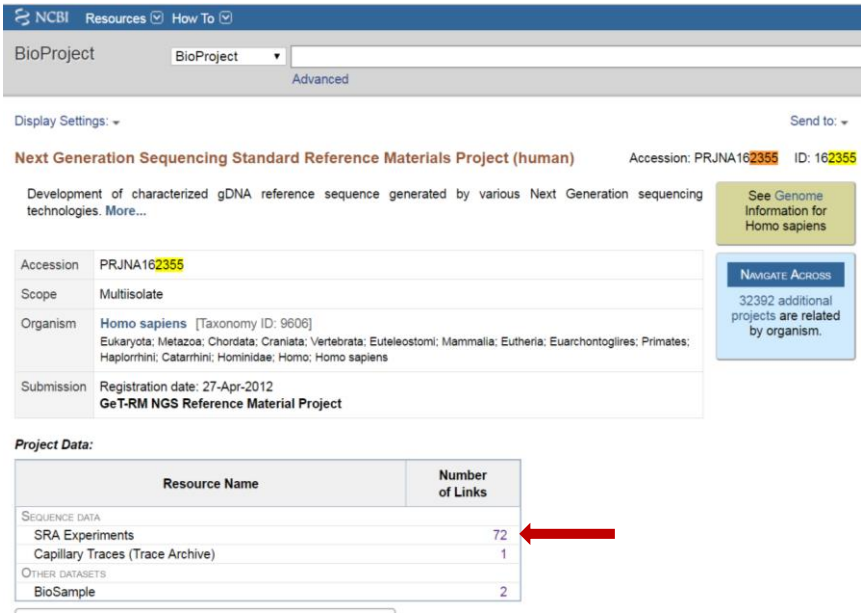
How to extract FASTQ ids for download using command line interfaces?

Let us look at the individual NA12878.

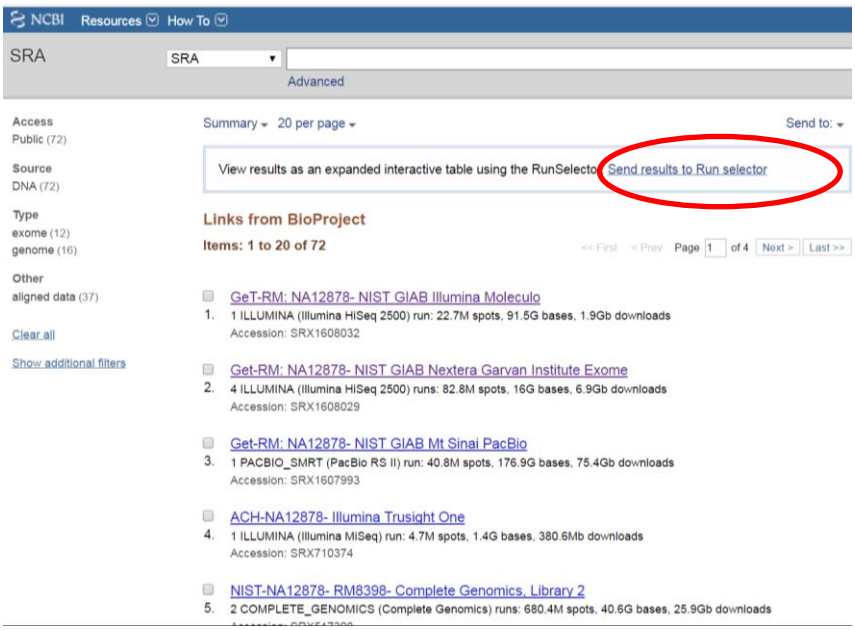
Go to BioProject, and type NA12878 (Hit "Clear All" before the search)



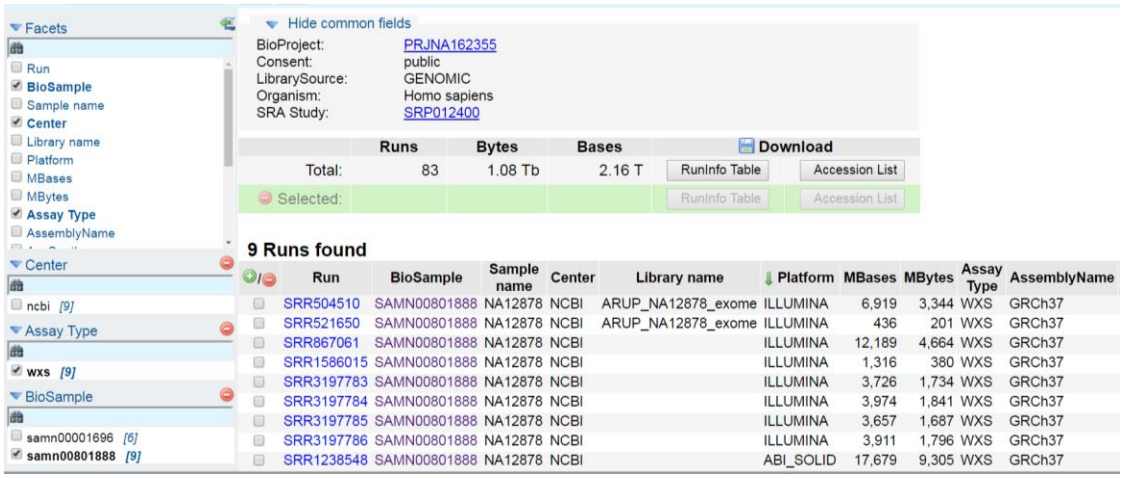
We are going to explore the "Next Generation Sequencing Standard Reference Materials Project" project, PRJNA162355



Click on 72 SRA experiments.



Click to send the results to a Run Selector. From the SRA Run Selector, use the left-hand side options to select, “BioSample”, “Assay Type” and then choose “wxs” and finally the sample samn00801888



Use “Runinfo Table” or “Accession List” for use with other command-line interfaces.,

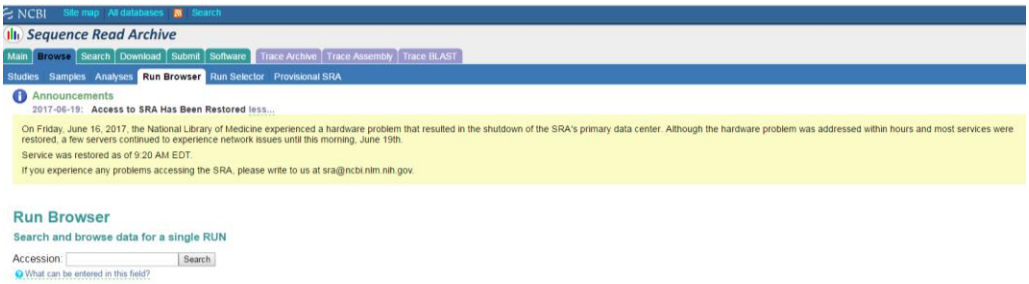
Exercise 2:

Are there any reads that have been mapped and readily available in NCBI?
The answer is yes. It is available for some of the SRA runs.

Go to SRA run Browser

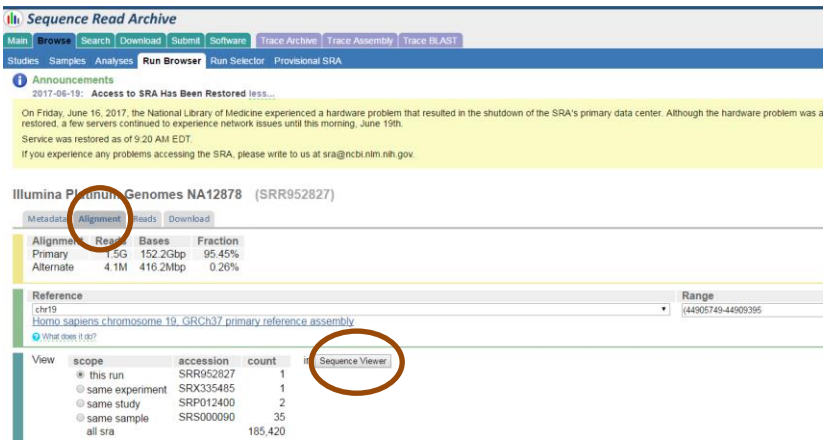
https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?

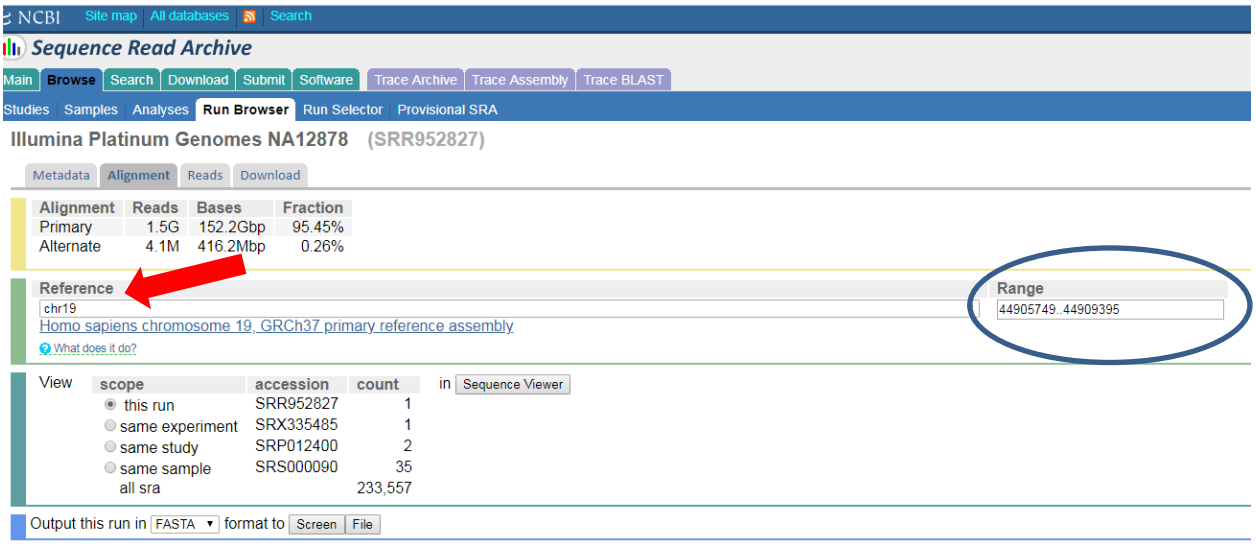
Select “Browse” → “Run Browser” to get to the following option (shown below)



Type in the SRR952827 id on the box.

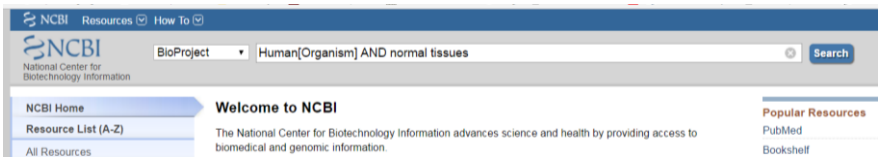
We are going to use APOE gene region, Chr19: 44905749..44909395 to see what reads from SRR952827 had been aligned.



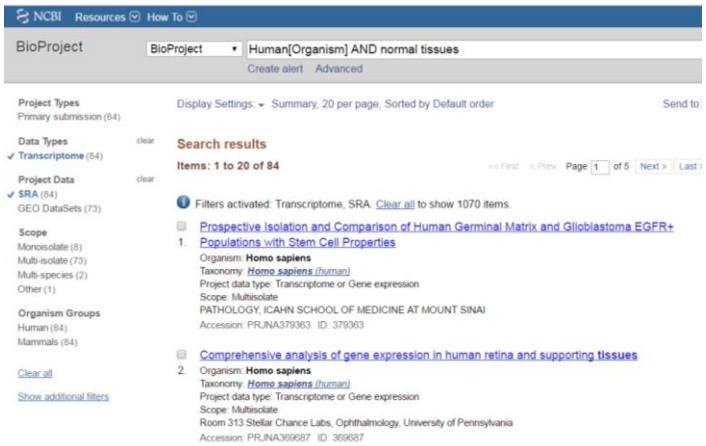


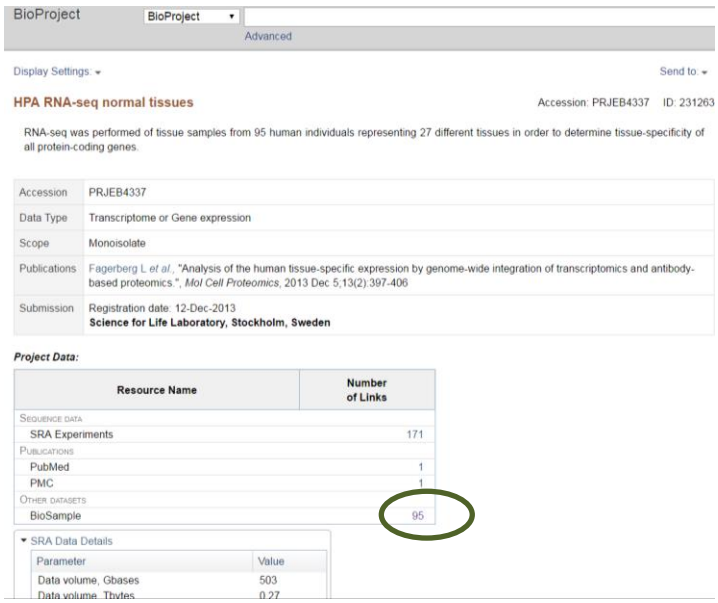
Exercise 3:

SRA BLAST is a useful procedure for reads that have not been aligned to the reference genome. Here we are going to use reads from a study and align it to human genome using BLASTN to accomplish this task.

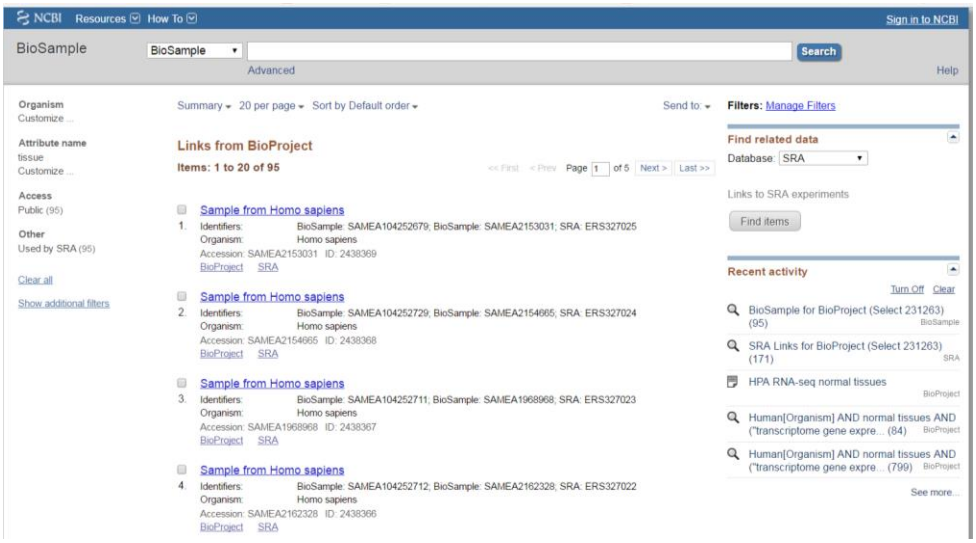


Select “Transcriptome” and “SRA” samples (use left-hand-side filters)

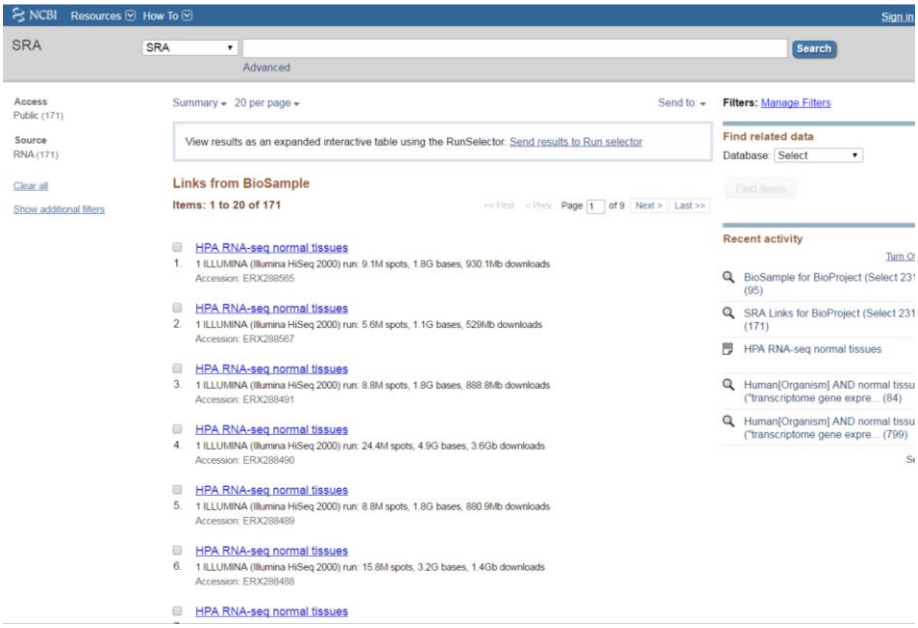


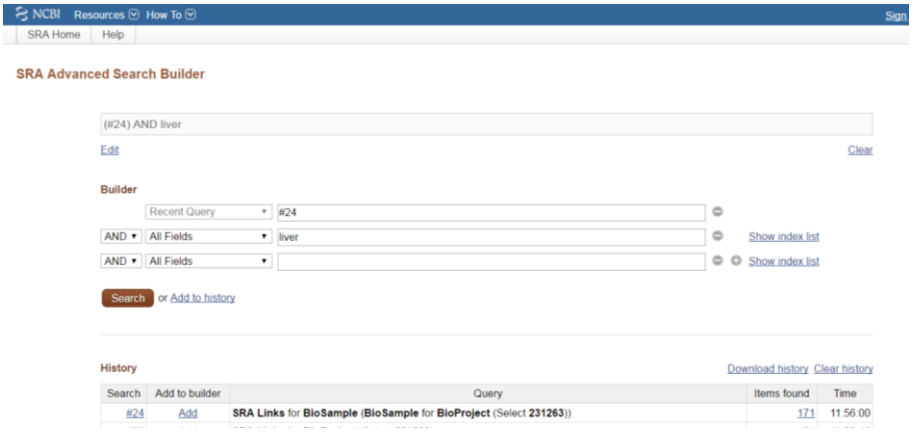


We are going to go into the 95 biosamples link and find related information on SRA Database.

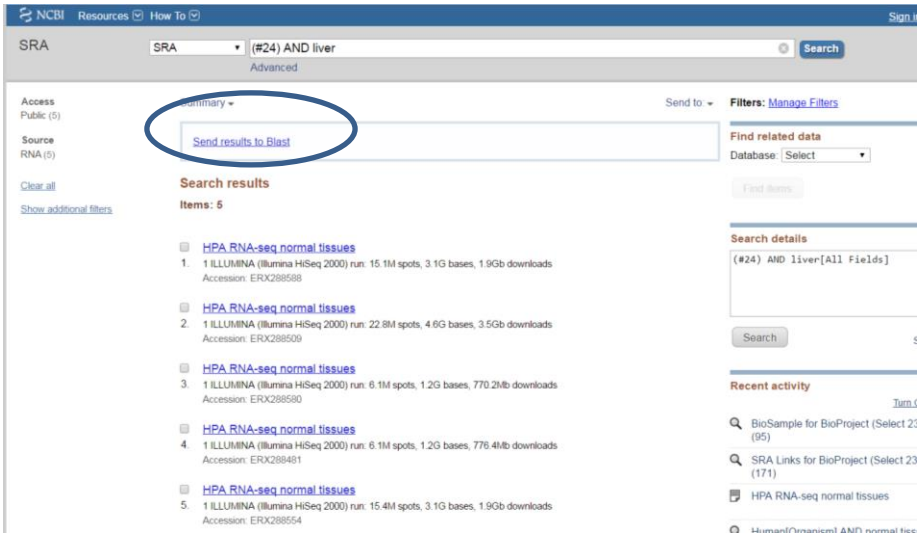


Select one of the projects from the hits, HPA RNA-seq normal tissues. Click to go into the project and using the related information on the right-hand-side to go into SRA. use advanced options to restrict sample only from liver.



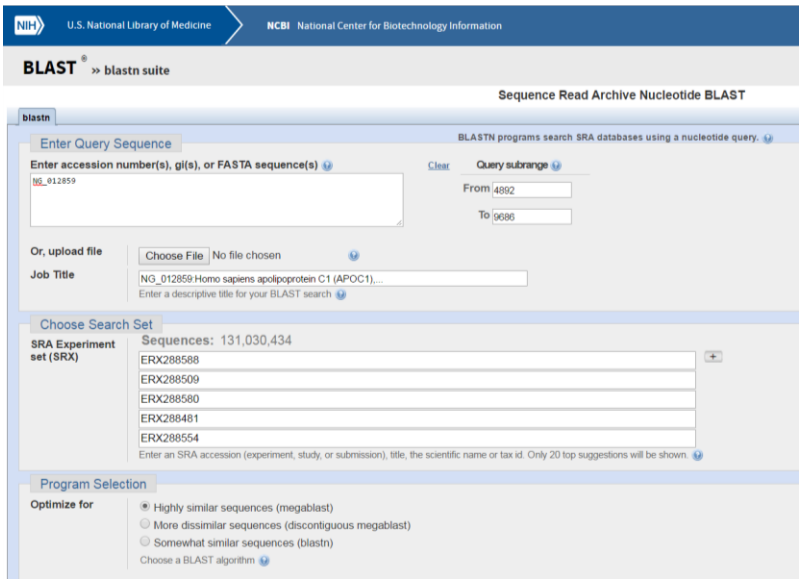


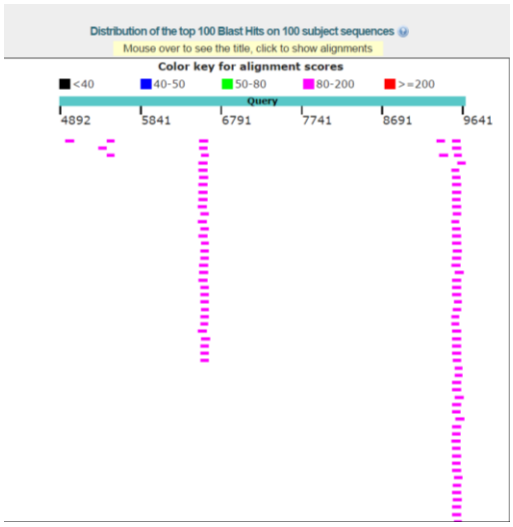
Use Advanced options and in “ALL Fields”, type in “liver” (without quotes). You should see 6 samples.



We are going to restrict ourselves only to APOC1 gene region (NG_012859: 4892-9686)

The SRA experiments used are ERX288588, ERX288509, ERX288580, ERX288481 and ERX288554 and we are using MegaBlast for this exercise. Here your query is a section of human genome and the database is created using the above mentioned ERX experiments.





Exercise 4 (optional):

If you have SRA Toolkit (NCBI application), you can use it to do accomplish many NGS related tasks such as extracting and analyzing SAM files.
For example, the following command will download aligned reads from SRR925743 into BRCA1.sam file:

sam-dump -aligned-region 17:41243452-41277500 SRR925743 > BRCA1.sam

(Note that for the above line to work in Linux/Mac, sam-dump, sra-toolkit executable should be in your path. Also for windows, the above line must be modified from sam-dump to sam-dump.exe, and maybe you must use the full path of sam-dump.exe)

Once you have successfully downloaded the SAM file, you can then view the contents of the SAM file using NCBI Genome Workbench.