

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.

Experiment (1 results found)
SRX220898 Illumina HiSeq 2500 paired end sequencing; UCSF_NA12878_AgilentV4
View all 1 results
Study (1 results found)
SRP012400 Next Generation Sequencing Standard Reference Materials Project
View all 1 results
Study (Sequence) (1 results found)
PRJNA162355 Next Generation Sequencing Standard Reference Materials Project
View all 1 results

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\[accn\]](https://www.ncbi.nlm.nih.gov/sra/SRX1608032[accn])
 - 2) SRX1608029: [https://www.ncbi.nlm.nih.gov/sra/SRX1608029\[accn\]](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)

NCBI Resources How To Sign in to NCBI

BioProject BioProject NA12878 Search

Create alert Advanced Help

Project Types
Umbrella (1)
Primary submission (66)

Data Types
Exome (3)
Genome sequencing (4)
Other (20)
Phenotype/genotype (1)
Random survey (1)
Targeted locus (1)
Variation (19)

Project Data
Nucleotide (11)
Assembly (10)
SRA (30)
dbVar (1)
GEO DataSets (19)

Scope
Monoisolate (40)
Multi-isolate (23)
Multi-species (2)
Synthetic (1)

Organism Groups
Human (65)
Mammals (65)

Clear all

Display Settings: Summary, 20 per page, Sorted by Default order

Search results
Items: 1 to 20 of 67

1. [NA12878 AmpliSeq Exome Sequencing](#)
Project data type: Raw sequence reads
Scope: Multispecies
University of Brescia
Accession: PRJNA390249 ID: 390249

2. [Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation \(CNV\) analysis in humans \[Aqilent022060\]](#)
Organism: Homo sapiens
Taxonomy: ID: 9606
Project data type: Variation
Scope: Multisolate
Life Sciences, The University of the West Indies
Accession: PRJNA380069 ID: 380069

3. [Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation \(CNV\) analysis in humans \[Aqilent023642\]](#)
Organism: Homo sapiens
Taxonomy: ID: 9606
Project data type: Variation
Scope: Multisolate
Life Sciences, The University of the West Indies
Accession: PRJNA380069 ID: 380069

Filters: Manage Filters

Find related data
Database: Select

Find items

Search details
NA12878[All Fields]

Search See more...

Recent activity
Turn Off Clear

NA12878 (67)
SRP012400 (72)
SAMN00801888 (541)

We are going to explore the “[Next Generation Sequencing Standard Reference Materials Project](#)” project, PRJNA162355

NCBI Resources How To

BioProject BioProject Advanced

Display Settings: Send to:

Next Generation Sequencing Standard Reference Materials Project (human) Accession: PRJNA162355 ID: 162355

Development of characterized gDNA reference sequence generated by various Next Generation sequencing technologies. [More...](#)

See Genome Information for Homo sapiens

Navigate Across
32392 additional projects are related by organism.

Accession	PRJNA162355
Scope	Multisolate
Organism	Homo sapiens [Taxonomy ID: 9606] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo; Homo sapiens
Submission	Registration date: 27-Apr-2012 GeT-RM NGS Reference Material Project

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	72
Capillary Traces (Trace Archive)	1
OTHER DATASETS	
BioSample	2

Click on 72 SRA experiments.

NCBI Resources How To

SRA SRA Advanced

Access Public (72)

Source DNA (72)

Type exome (12) genome (16)

Other aligned data (37)

Clear all

Show additional filters

Summary 20 per page Send to:

View results as an expanded interactive table using the RunSelector. **Send results to Run selector**

Links from BioProject

Items: 1 to 20 of 72

<< First < Prev Page 1 of 4 Next > Last >>

1. [Get-UM: NA12878- NIST GIAB Illumina Molecule](#)
1 ILLUMINA (Illumina HiSeq 2500) run: 22.7M spots, 91.5G bases, 1.9Gb downloads
Accession: SRX1608032
2. [Get-UM: NA12878- NIST GIAB Nextera Garvan Institute Exome](#)
4 ILLUMINA (Illumina HiSeq 2500) runs: 82.8M spots, 16G bases, 6.9Gb downloads
Accession: SRX1608029
3. [Get-UM: NA12878- NIST GIAB Mt Sinai PacBio](#)
1 PACBIO_SMRT (PacBio RS II) run: 40.8M spots, 176.9G bases, 75.4Gb downloads
Accession: SRX1607993
4. [ACH-NA12878- Illumina Trusight One](#)
1 ILLUMINA (Illumina MiSeq) run: 4.7M spots, 1.4G bases, 380.6Mb downloads
Accession: SRX710374
5. [NIST-NA12878- RM8398- Complete Genomics Library 2](#)
2 COMPLETE_GENOMICS (Complete Genomics) runs: 680.4M spots, 40.6G bases, 25.9Gb downloads
Accession: SRX1608032

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

Facets

- Run
- ☒ BioSample
- Sample name
- Center
- Library name
- Platform
- MBases
- MBytes
- ☒ Assay Type
- AssemblyName

Center

- ncbi [9]
- Assay Type
- ☒ wxs [9]
- BioSample
- samn00001696 [6]
- ☒ samn00801888 [9]

Hide common fields

BioProject: [PRJNA162355](#)

Consent: public

LibrarySource: GENOMIC

Organism: Homo sapiens

SRA Study: [SRP012400](#)

	Runs	Bytes	Bases	Download
Total:	83	1.08 Tb	2.16 T	RunInfo Table Accession List
Selected:				RunInfo Table Accession List

9 Runs found

Run	BioSample	Sample name	Center	Library name	Platform	MBases	MBytes	Assay Type	AssemblyName
<input type="checkbox"/> SRR504510	SAMN00801888	NA12878	NCBI	ARUP_NA12878_exome	ILLUMINA	6,919	3,344	WXS	GRCh37
<input type="checkbox"/> SRR521650	SAMN00801888	NA12878	NCBI	ARUP_NA12878_exome	ILLUMINA	436	201	WXS	GRCh37
<input type="checkbox"/> SRR867061	SAMN00801888	NA12878	NCBI		ILLUMINA	12,189	4,664	WXS	GRCh37
<input type="checkbox"/> SRR1586015	SAMN00801888	NA12878	NCBI		ILLUMINA	1,316	380	WXS	GRCh37
<input type="checkbox"/> SRR3197783	SAMN00801888	NA12878	NCBI		ILLUMINA	3,726	1,734	WXS	GRCh37
<input type="checkbox"/> SRR3197784	SAMN00801888	NA12878	NCBI		ILLUMINA	3,974	1,841	WXS	GRCh37
<input type="checkbox"/> SRR3197785	SAMN00801888	NA12878	NCBI		ILLUMINA	3,657	1,687	WXS	GRCh37
<input type="checkbox"/> SRR3197786	SAMN00801888	NA12878	NCBI		ILLUMINA	3,911	1,796	WXS	GRCh37
<input type="checkbox"/> SRR1238548	SAMN00801888	NA12878	NCBI		ABI_SOLID	17,679	9,305	WXS	GRCh37

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

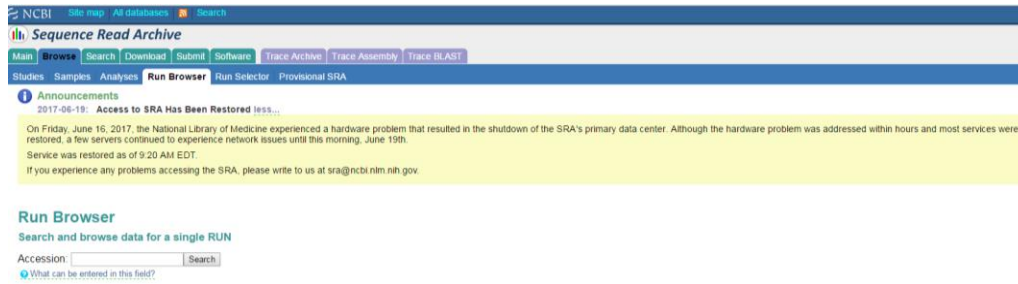
Exercise 2:

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

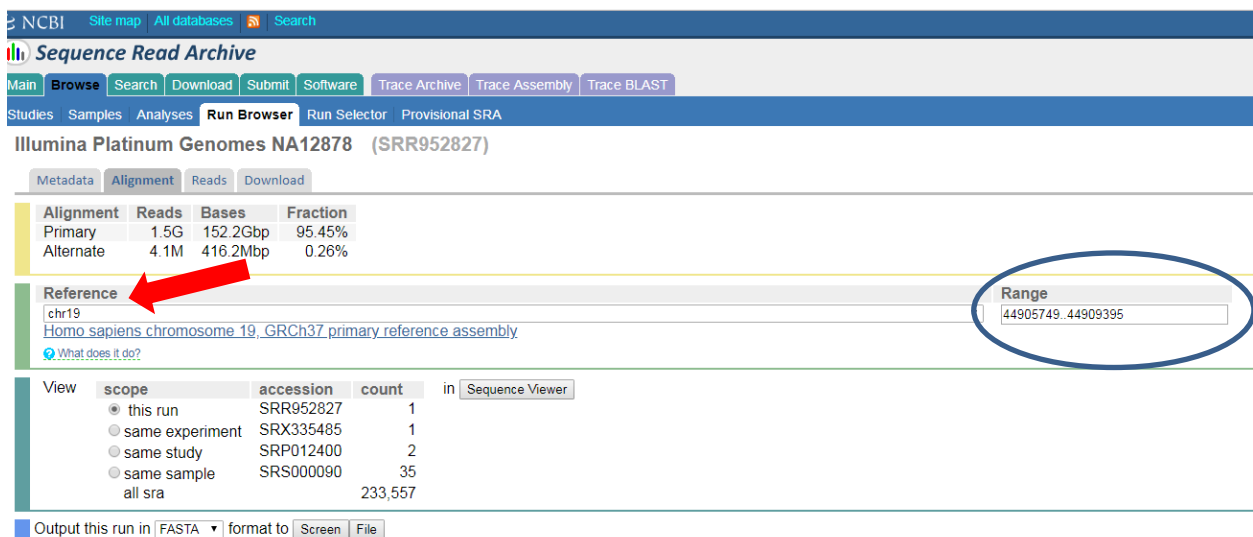
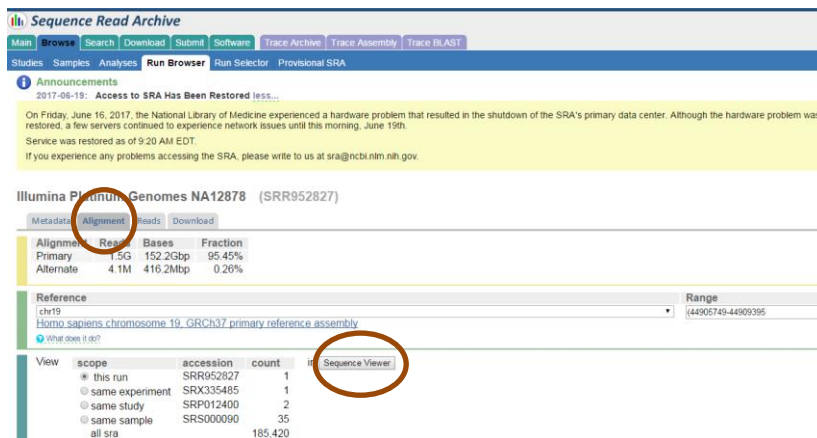
<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?>

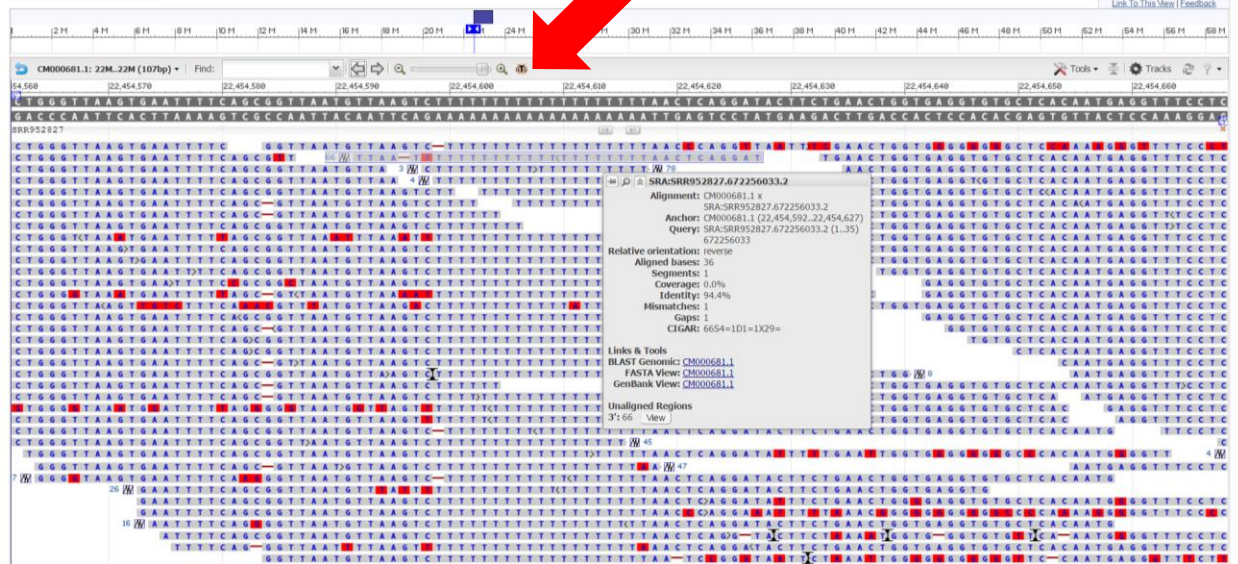
Select “Browse” → “Run Browser” to get to the following option (shown below)



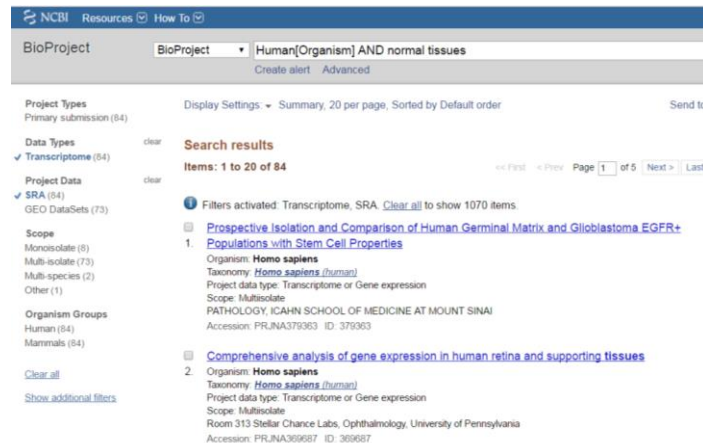
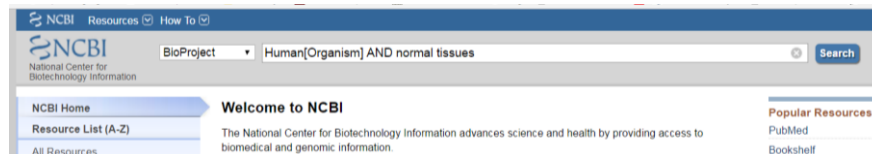
Type in the SRR952827 id on the box.

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





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.



BioProject BioProject Advanced

Display Settings: Send to: v

HPA RNA-seq normal tissues Accession: PRJEB4337 ID: 231263

RNA-seq was performed of tissue samples from 95 human individuals representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes.

Accession	PRJEB4337
Data Type	Transcriptome or Gene expression
Scope	Monoisolate
Publications	Fagerberg L <i>et al.</i> , "Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics", <i>Mol Cell Proteomics</i> , 2013 Dec 5;13(2):397-406
Submission	Registration date: 12-Dec-2013 Science for Life Laboratory, Stockholm, Sweden

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	171
PUBLICATIONS	
PubMed	1
PMC	1
OTHER DATASETS	
BioSample	95

▼ SRA Data Details

Parameter	Value
Data volume, Gbases	503
Data volume, Tbytes	0.27

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

NCBI Resources How To Sign in to NCBI

BioSample BioSample Advanced Search Help

Organism: Customize ...

Attribute name: tissue: Customize ...

Access: Public (95)

Other: Used by SRA (95)

Clear all Show additional filters

Summary: 20 per page Sort by Default order Send to: Filters: Manage Filters

Links from BioProject

Items: 1 to 20 of 95

1. [Sample from Homo sapiens](#)
Identifiers: BioSample: SAMEA104252679; BioSample: SAMEA2153031; SRA: ERS327025
Organism: Homo sapiens
Accession: SAMEA2153031 ID: 2438369
[BioProject](#) [SRA](#)

2. [Sample from Homo sapiens](#)
Identifiers: BioSample: SAMEA104252729; BioSample: SAMEA2154965; SRA: ERS327024
Organism: Homo sapiens
Accession: SAMEA2154965 ID: 2438368
[BioProject](#) [SRA](#)

3. [Sample from Homo sapiens](#)
Identifiers: BioSample: SAMEA104252711; BioSample: SAMEA1968968; SRA: ERS327023
Organism: Homo sapiens
Accession: SAMEA1968968 ID: 2438367
[BioProject](#) [SRA](#)

4. [Sample from Homo sapiens](#)
Identifiers: BioSample: SAMEA104252712; BioSample: SAMEA2162328; SRA: ERS327022
Organism: Homo sapiens
Accession: SAMEA2162328 ID: 2438366
[BioProject](#) [SRA](#)

Find related data

Database: SRA

Links to SRA experiments

Find items

Recent activity Turn Off Clear

BioSample for BioProject (Select 231263) (95) BioSample

SRA Links for BioProject (Select 231263) (171) SRA

HPA RNA-seq normal tissues BioProject

Human[Organism] AND normal tissues AND ("Transcriptome gene expre... (84) BioProject

Human[Organism] AND normal tissues AND ("Transcriptome gene expre... (799) BioProject

See more...

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.

NCBI Resources How To Sign In

SRA SRA Advanced Search

Access: Public (171)
Source: RNA (171)
[Clear all](#)
[Show additional filters](#)

Summary 20 per page Send to Filters: [Manage Filters](#)

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

Find related data
Database: [Select](#)

Find items

Recent activity
[Turn Off](#)
BioSample for BioProject (Select 231 (95))
SRA Links for BioProject (Select 231 (171))
HPA RNA-seq normal tissues
Human[Organism] AND normal tissue (transcriptome gene expression) (84)
Human[Organism] AND normal tissue (transcriptome gene expression) (799)

Links from BioSample
Items: 1 to 20 of 171

1. [HPA RNA-seq normal tissues](#)
1 ILLUMINA (Illumina HiSeq 2000) run: 9.1M spots, 1.8G bases, 930.1Mb downloads
Accession: ERX289565

2. [HPA RNA-seq normal tissues](#)
1 ILLUMINA (Illumina HiSeq 2000) run: 5.6M spots, 1.1G bases, 529Mb downloads
Accession: ERX289567

3. [HPA RNA-seq normal tissues](#)
1 ILLUMINA (Illumina HiSeq 2000) run: 8.8M spots, 1.8G bases, 888.8Mb downloads
Accession: ERX289491

4. [HPA RNA-seq normal tissues](#)
1 ILLUMINA (Illumina HiSeq 2000) run: 24.4M spots, 4.9G bases, 3.6Gb downloads
Accession: ERX289490

5. [HPA RNA-seq normal tissues](#)
1 ILLUMINA (Illumina HiSeq 2000) run: 8.8M spots, 1.8G bases, 880.9Mb downloads
Accession: ERX289489

6. [HPA RNA-seq normal tissues](#)
1 ILLUMINA (Illumina HiSeq 2000) run: 15.8M spots, 3.2G bases, 1.4Gb downloads
Accession: ERX289488

7. [HPA RNA-seq normal tissues](#)

NCBI Resources How To Sign In

SRA Home Help

SRA Advanced Search Builder

(#24) AND liver

[Edit](#) [Clear](#)

Builder

Recent Query: #24

AND All Fields liver [Show index list](#)

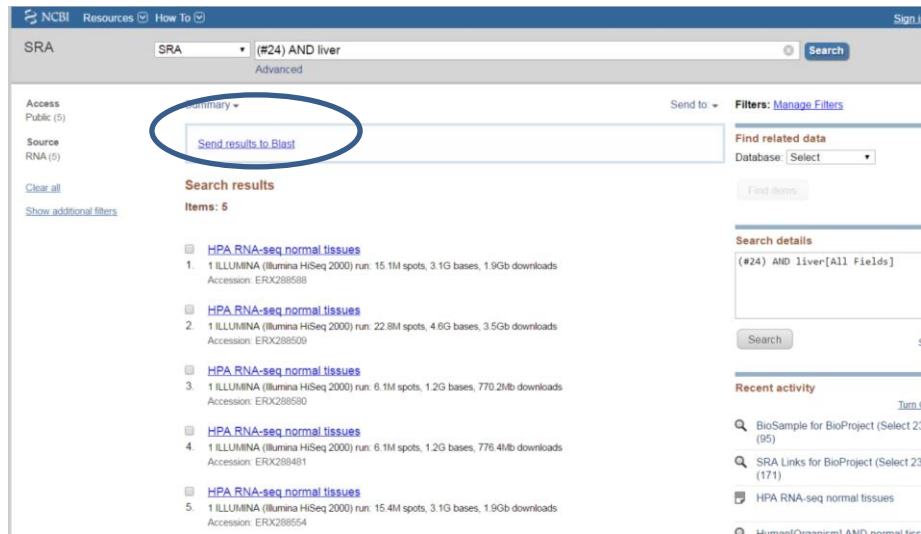
AND All Fields [Show index list](#)

[Search](#) or [Add to history](#)

History [Download history](#) [Clear history](#)

Search	Add to builder	Query	Items found	Time
#24	Add	SRA Links for BioSample (BioSample for BioProject (Select 231263))	171	11:56:00

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.

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information

BLAST » blastn suite

Sequence Read Archive Nucleotide BLAST

blastn

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

NG_012859

Clear Query subrange

From 4892 To 9686

Or, upload file

Choose File | No file chosen

Job Title

NG_012859 Homo sapiens apolipoprotein C1 (APOC1),...

Enter a descriptive title for your BLAST search

Choose Search Set

Sequences: 131,030,434

SRA Experiment set (SRX)

ERX288588

ERX288509

ERX288580

ERX288481

ERX288554

Enter an SRA accession (experiment, study, or submission), title, the scientific name or tax id. Only 20 top suggestions will be shown.

Program Selection

Optimize for

☒ Highly similar sequences (megablast)

☐ More dissimilar sequences (discontiguous megablast)

☐ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Job title: NG_012859:Homo sapiens apolipoprotein C1 (APOC1),...

RID MN638H2014 (Expires on 06-23 00:01 am)

Query ID NG_012859.1

Description Homo sapiens apolipoprotein C1 (APOC1), RefSeqGene on chromosome 19

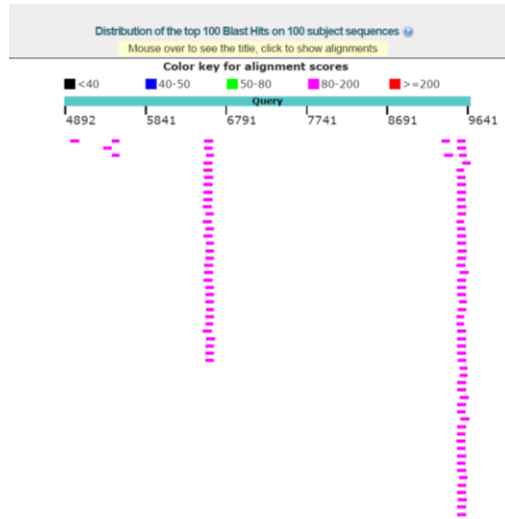
Molecule type nucleic acid

Query Length 4794

Database Name SRA

Description BLASTN 2.6.1+ » See details

Program BLASTN 2.6.1+ » Citation



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.