Download GEO Data (SRA Toolkit)

The Sequence Read Archive (SRA) is a publicly accessible archive for high throughput sequencing data. The <u>SRA Toolkit</u> from NCBI is a collection of tools for using data in the INSDC SRA. It takes the following steps to download data from SRA:

Install and Config SRA Toolkit

```
# Download and extract the latest version
$ wget --output-document sratoolkit.tar.gz https://ftp-
trace.ncbi.nlm.nih.gov/sra/sdk/2.10.9/sratoolkit.2.10.9-ubuntu64.tar.gz
$ tar -vxzf sratoolkit.2.10.9-ubuntu64.tar.gz

# Append the path to your PATH environment variable:
$ export PATH=$PATH:/path/to/sratoolkit.2.10.9-ubuntu64/bin

# Verify the installation
$ which fastq-dump
```

Download Data from SRA

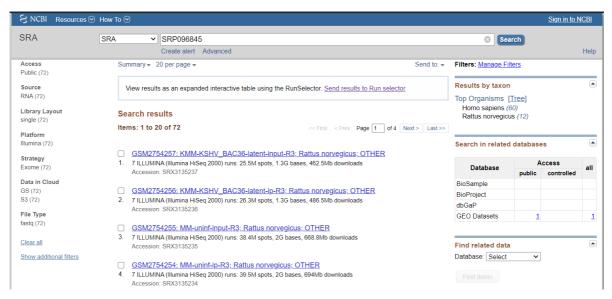
1. Access the GEO summary page by searching "GSE93676" on GEO website.



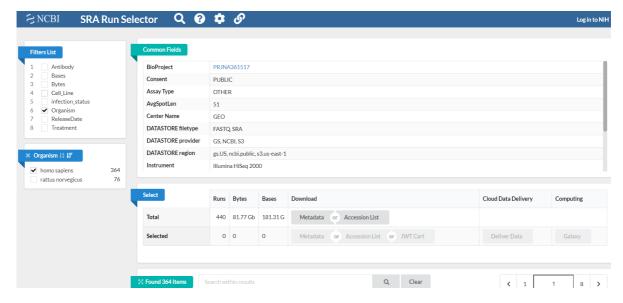
2. Find a link for "SRA" under the heading "Relations".

```
Submission date Jan 16, 2017
Last update date May 15, 2019
Contact name Hui Liu
                lhcumt@hotmail.com
E-mail(s)
Organization name China University of Mining and Technology
Street address #1 Daxue Road
City
                Xuzhou
State/province
                Jiangsu
                221116
ZIP/Postal code
Country
                 China
Platforms (2)
                 GPL11154 Illumina HiSeq 2000 (Homo sapiens)
                 GPL14844 Illumina HiSeq 2000 (Rattus norvegicus)
Samples (72)
                 GSM2460344 iSLK-uninf-input
■ More...
                 GSM2460345 iSLK-uninf-ip
                 GSM2460346 iSLK-KSHV_BAC16-latent-input
Relations
                 PRJNA361517
BioProject
SRA
                 SRP096845
```

3. Click on the link (SRP096845) which sends you to a page of all the biological samples with specific runs and files in this study.



4. To find files of interest in one comprehensive list, navigate to the bottom of the page then click: "send to" > "Run Selector" > "go". Use "Filter List" to narrow down the choices.



5. Extract FastQ files from SRA-accession using SRA-Toolkit

```
#!/bin/bash
cd /path/to/raw_data/homo/
fetch_dump(){
prefetch $1
fastq-dump $1
}
export -f fetch_dump
for s in SRR5978827 SRR5978828 SRR5978829 SRR5978834 SRR5978835 SRR5978836
SRR5978869 SRR5978870 SRR5978871 SRR5179446 SRR5179447 SRR5179448
do
fetch_dump ${s}
done
```

```
#!/bin/bash
cd /path/to/raw_data/mm10/
fetch_dump(){
prefetch $1
fastq-dump $1
}
export -f fetch_dump
for s in SRR866997 SRR866998 SRR866999 SRR867000 SRR867001 SRR867002 SRR866991
SRR866992 SRR866993 SRR866994 SRR866995 SRR866996
do
fetch_dump ${s}
done
```

For paired-end data, you need to add --split-files option in the fastq-dump command.

```
fastq-dump --split-files SRR866997
```