

# Downloading Public Data

Author: Mary Allen

## From NIH GEO

### Downloading one fastq

```
#!/bin/bash
#SBATCH --mail-type=ALL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --nodes=1 # Run on a single node
#SBATCH --ntasks=1 # Number of CPU (processor cores i.e. tasks) In this example I use 1. I only need one, since none of the commands I run are parallelized.
#SBATCH --mem=1gb # Memory limit
#SBATCH --time=24:00:00 # Time limit hrs:min:sec
#SBATCH --job-name=featurecounts # Job name
#SBATCH --mail-user=email@colorado.edu # Where to send mail
#SBATCH --partition short # Job queue
#SBATCH --output=/scratch/Users/username/eofiles/%x_%j.out
#SBATCH --error=/scratch/Users/username/eofiles/%x_%j.err

filename=SRR15283267
outdir=

module load sra/2.9.2

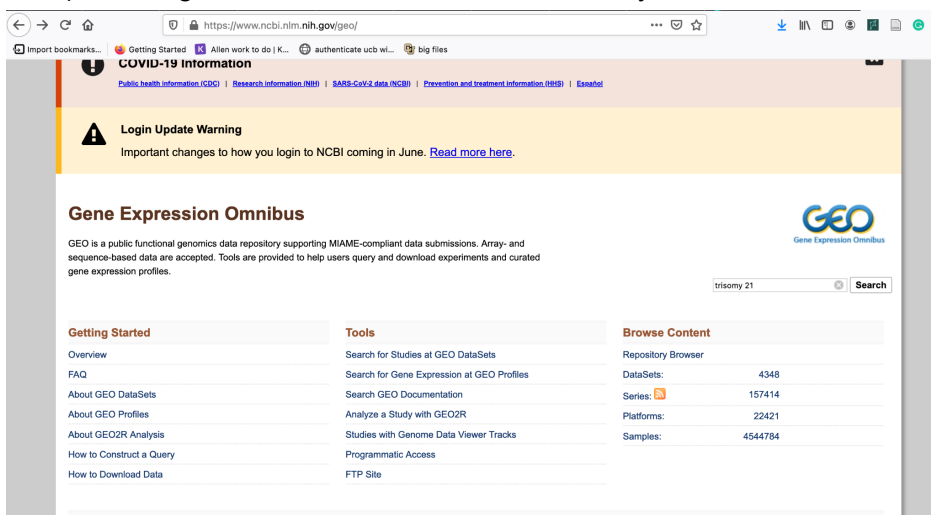
fastq-dump -O $outdir -split-3 $filename
```

The `-split-3` flag is essential and should be default. That flag will do nothing if you have single end data. If you have paired end data you NEED it. Paired end data will come down as one file instead of two files (the R1, which means read1, and R2 files)

## Getting a SRR# for the fastq you want

### 1. Get the SRP# number

1) Go to geo and find the SRP# for the data you care about



The screenshot shows the NCBI GEO website. At the top, there is a navigation bar with links for COVID-19 information, Public health information (CDC), Research information (NBI), SARS-CoV-2 data (NCBI), and Prevention and treatment information (HHS). Below this is a yellow banner with a login update warning. The main content area is titled "Gene Expression Omnibus" and includes a search bar with the text "trisomy 21" and a "Search" button. Below the search bar, there are three columns of links: "Getting Started" (Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, How to Download Data), "Tools" (Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, Search GEO Documentation, Analyze a Study with GEO2R, Studies with Genome Data Viewer Tracks, Programmatic Access, FTP Site), and "Browse Content" (Repository Browser, DataSets: 4348, Series: 157414, Platforms: 22421, Samples: 4544784).

https://www.ncbi.nlm.nih.gov/gds/?term=trisomy 21

NCBI Resources How To maallen3 My NCBI

GEO DataSets GEO DataSet: trisomy 21 Search

Create alert Advanced

**COVID-19 Information**  
[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

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

DataSets (11)  
 Series (309)  
 Samples (1,811)  
 Platforms (7)

Organization  
 Customize ...

Study type  
 Expression profiling by array  
 Methylation profiling by array  
 Customize ...

Author  
 Customize ...

Attribute name  
 tissue (873)  
 strain (156)  
 Customize ...

Publication dates  
 30 days  
 1 year

**Search results**  
 Items: 1 to 20 of 2138

1. [Blood and immune cell development in human fetal bone marrow and in Down syndrome](#)  
 (Submitter supplied) Throughout postnatal life, haematopoiesis in the bone marrow (BM) maintains blood and immune cell production. Haematopoiesis first emerges in human BM at 11-12 post conception weeks while fetal liver (FL) haematopoiesis is still expanding. Yet, almost nothing is known about how fetal BM evolves to meet the highly specialised needs of the fetus and newborn infant. Here, we detail the development of fetal BM including stroma using single cell RNA-sequencing. more...

Organization: Homo sapiens  
 Type: Expression profiling by high throughput sequencing; Other  
 Platform: GPL24676 16 Samples  
 Download data: CSV  
 Series Accession: GSE166895 ID: 200166895  
[Analyze with GEO2R](#) [SRA Run Selector](#)

2. [Mapping the Cellular Origin and Early Evolution of Leukemia in Down Syndrome](#)  
 (Submitter supplied) This SuperSeries is composed of the SubSeries listed below.

**Find related data**  
 Database: Select  
 Find items

**Search details**  
 "down syndrome" [MeSH Terms] 0  
 trisomy 21[All Fields]

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE166895

QC) FBM MNCs.  
 preQC\_ADTraw\_FL-FBM-CB.csv.gz: RAW Protein - Post-CITE-seq-Count (pre-QC) Combination of FL-FBM-CB.  
 preQC\_mRNARaw\_FBM-MNCs.csv.gz: RAW mRNA - Post-cell ranger (pre-QC) FBM MNCs.  
 preQC\_mRNARaw\_FL-FBM-CB.csv.gz: RAW mRNA - Post-cell ranger (pre-QC) Combination of FL-FBM-CB.  
 Raw\_ADT\_CD34SInu\_FLFBM.csv.gz: RAW Protein - Expression data.  
 Raw\_mRNA\_CD34SInu\_FLFBM.csv.gz: RAW mRNA - Sinusoidal endo RNA.

Contributor(s) [Hannah R, Wilson NK, Quiroga Londoño M](#)  
 Citation missing *Has this study been published? Please notify GEO.*  
 Submission date Feb 16, 2021  
 Last update date Jul 09, 2021  
 Contact name Rebecca Hannah  
 E-mail(s) [rh60@cam.ac.uk](mailto:rh60@cam.ac.uk)  
 Organization name University of Cambridge  
 Street address Cambridge Institute for Medical Research, University of Cambridge, Cambridge Biomedical Campus, Wellcome Trust / MRC Building, Hills Road  
 City Cambridge  
 ZIP/Postal code CB2 0XY  
 Country United Kingdom

Platforms (1) [GPL24676](#) Illumina NovaSeq 6000 (Homo sapiens)  
 Samples (16) [GSM5087731](#) RBG33493\_F90\_F103\_F107\_mRNA  
[GSM5087732](#) RBG33493\_F90\_F103\_F107\_protein  
[GSM5087733](#) RBG33494\_CIMA-60\_CIMA-58\_F103\_mRNA

**Relations**  
 BioProject [PRJNA702168](#)  
 SRA [SRP306736](#)

**Download family** **Format**

## Getting a list of all SRRs for this identifier.

1. Paste the SRP into this website and search- <https://trace.ncbi.nlm.nih.gov/Traces/study/>
2. Download both accession and metadata
3. metadata will download a tab delimited text file of all SRR in that SRP
  - a. This is a start for your metadata table you will need for Deseq2
4. Accessions are just the SRRs and this is the file you will need to download stuff

## Arrange a out directory

1. Make a new outdirectory on the supercomputer
2. Copy the scripts I made for downloading a fastq from GEO

```

[maallen3@ip-172-31-38-192 day10]$ mkdir -p /scratch/Users/maallen3/day10/
[maallen3@ip-172-31-38-192 day10]$ cd /scratch/Users/maallen3/day10/
[maallen3@ip-172-31-38-192 day10]$ scp /scratch/Shares/public/sread2021/scripts/day10/downloadfromgeo/* .
[maallen3@ip-172-31-38-192 day10]$ ls
downloadafastq.sh  downloadall.sh  test_SRR_Acc_List.txt
[maallen3@ip-172-31-38-192 day10]$ █

```

## Upload your SRR\_ACC\_List.txt to the super computer

```

cu-biot-6-10:~ maryallen$ rsync Downloads/SRR_Acc_List.txt maallen3@18.219.252.252::/scratch/Shares/public/sread2021/scripts/day10/downloadfromgeo/

```

Run the sbatch script

1. Make a directory for the fastq files to go into
2. Edit the [downloadall.sh](#) script.
  - a. Change your email!!!!
3. Run the two scripts by typing

```
bash downloadall.sh <pathtoyour_SRR_AccList.txt> <outdir>
```

```

[~bash-4.2$ bash downloadall.sh SRR_Acc_List.txt fastqfiles/
SRR11856162

```

### 4. Under the hood

- a. The main script, called `downloadfastq.sbatch`, uses a program called `fastq-dump` to download public data from GEO.
  - i. **IMPORTANT!!! Use the `-split-3` flag every time!** If you are doing single end data that flag does nothing (it won't hurt you), but if you are using paired end data then that flag outputs the fastq files as two files. If you don't use this flag you will get one file with both the forward and reverse reads in it!!!!

```

maryallen — ssh allenma@fiji.colorado.edu — 155x46
#!/bin/bash
#SBATCH --mail-type=ALL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --nodes=1 # Run on a single node
#SBATCH --ntasks=1 # Number of CPU (processor cores i.e. tasks) In this example I use 1. I only need one, since none of the commands I run
#SBATCH --mem=1gb # Memory limit
#SBATCH --time=24:00:00 # Time limit hrs:min:sec

module load sra/2.9.2

fastq-dump -O $outdir -split-3 $filename
~
~
~
~
~
~
~
~
~
~
~

```

- b. This script, `downloadall.sh`, just runs the other script via `sbatch`.
  - i. It reads the `SRR_ACC_List.txt` file and submits each SRR to a different CPU to download.

```
#!/bin/bash
#went to https://trace.ncbi.nlm.nih.gov/Traces/study/?go=home and searched for SRP002796 to download both the SRR_ACC_List.txt and the Sra-run-table
#to run type bash downloadall.sh <path_to_SRR_ACC_List.txt> <outdir> <email>

mkdir -p $2

IFS=''
while read var
do
echo $var
if [ -n "$var" ];
then
sbatch -J $var --mail-user=$3 --output=${2}/${var}_%j.out --output=${2}/${var}_%j.err --export=filename=$var,outdir=$2 downloadafastq.sh
fi
done < $1
```

c. If it works you will have fastq files in your outdirectory:

```
-bash-4.2$ cd fastqfiles/
-bash-4.2$ ls
eando SRR11856162_2.fastq SRR11856163_2.fastq SRR11856164_2.fastq SRR11856165_2.fastq SRR11856167_2.fastq
SRR11856162_1.fastq SRR11856163_1.fastq SRR11856164_1.fastq SRR11856165_1.fastq SRR11856167_1.fastq
-bash-4.2$ cd eando/
-bash-4.2$ ls
SRR11856162.7773752.err SRR11856163.7773753.out SRR11856165.7773755.err SRR11856166.7773757.out SRR11856168.7773759.err SRR11856169.7773760.out
SRR11856162.7773752.out SRR11856164.7773754.err SRR11856165.7773755.out SRR11856167.7773758.err SRR11856168.7773759.out
SRR11856163.7773753.err SRR11856164.7773754.out SRR11856166.7773757.err SRR11856167.7773758.out SRR11856169.7773760.err
-bash-4.2$ cd ..
-bash-4.2$ wc -l *.fastq
 3279023 SRR11856162_1.fastq
 3496411 SRR11856162_2.fastq
 3234267 SRR11856163_1.fastq
 3304597 SRR11856163_2.fastq
 3982336 SRR11856164_1.fastq
 4165733 SRR11856164_2.fastq
 4456303 SRR11856165_1.fastq
 4791089 SRR11856165_2.fastq
 5132192 SRR11856166_1.fastq
 5523829 SRR11856166_2.fastq
 5650163 SRR11856167_1.fastq
 5909149 SRR11856167_2.fastq
 7930501 SRR11856168_1.fastq
 8625340 SRR11856168_2.fastq
 9010660 SRR11856169_1.fastq
 9629699 SRR11856169_2.fastq
 88121292 total
-bash-4.2$
```

d. You should do a wc -l on the files to check them.

- i. SRR#\_1.fastq and SRR#\_2.fastq represents read 1 and read 2 of a pair. Therefore SRR#\_1.fastq and SRR#\_2.fastq should have the same line numbers.

## From CistromeDB (processed ChIP-seq)

1. Go to <http://cistrome.org/db/#/>
2. Pick an organism, cell line and TF
3. YOU can do a lot on this site
  - a. download the bed file
  - b. look at the quality of each chip
  - c. See what motif was most enriched in this chip
  - d. Find genes that may be regulated by this TF
  - e. They also have a site you can search a gene to see what TFs bind it
    - i. <http://dbtoolkit.cistrome.org/>

