# Downloading Public Data

Author: Mary Allen

# From NIH GEO

## Downloading one fastq

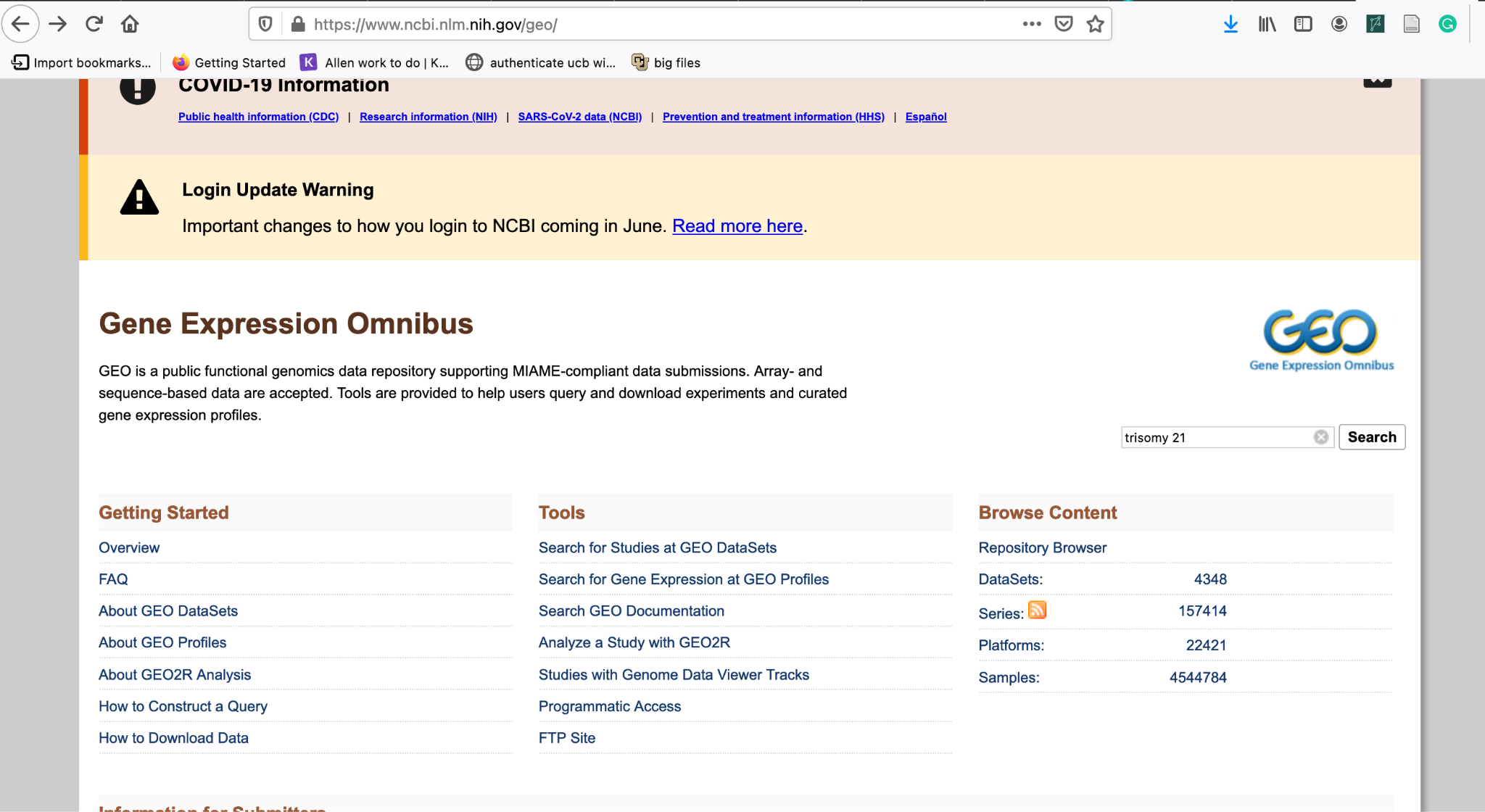
## 

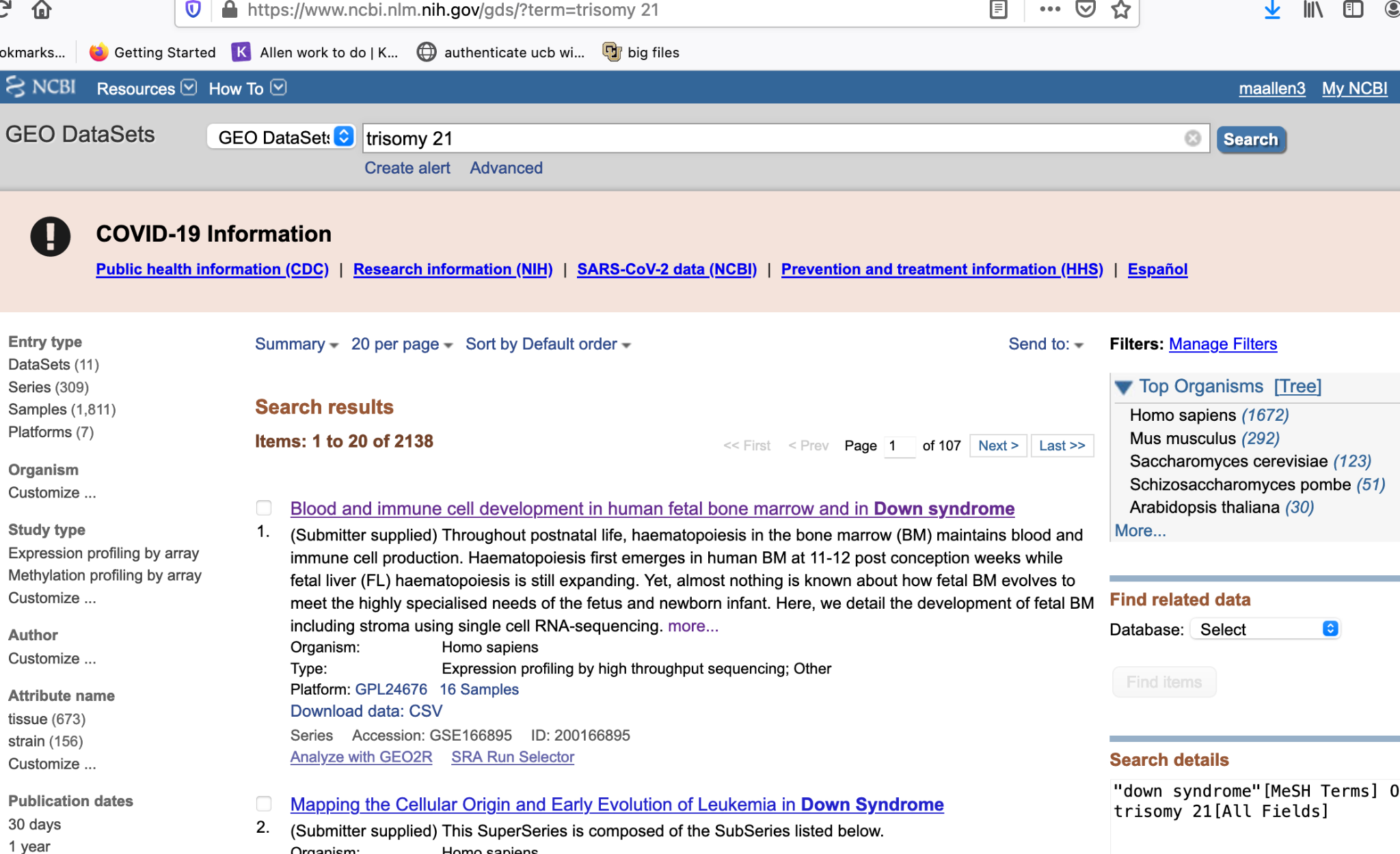
The -split-3 flag is essential and should be default. That flag will do nothing if you have single end data. If you have paired end data you NEED it. Paired end data will come down as one file instead of two files (the R1, which means read1, and R2 files)

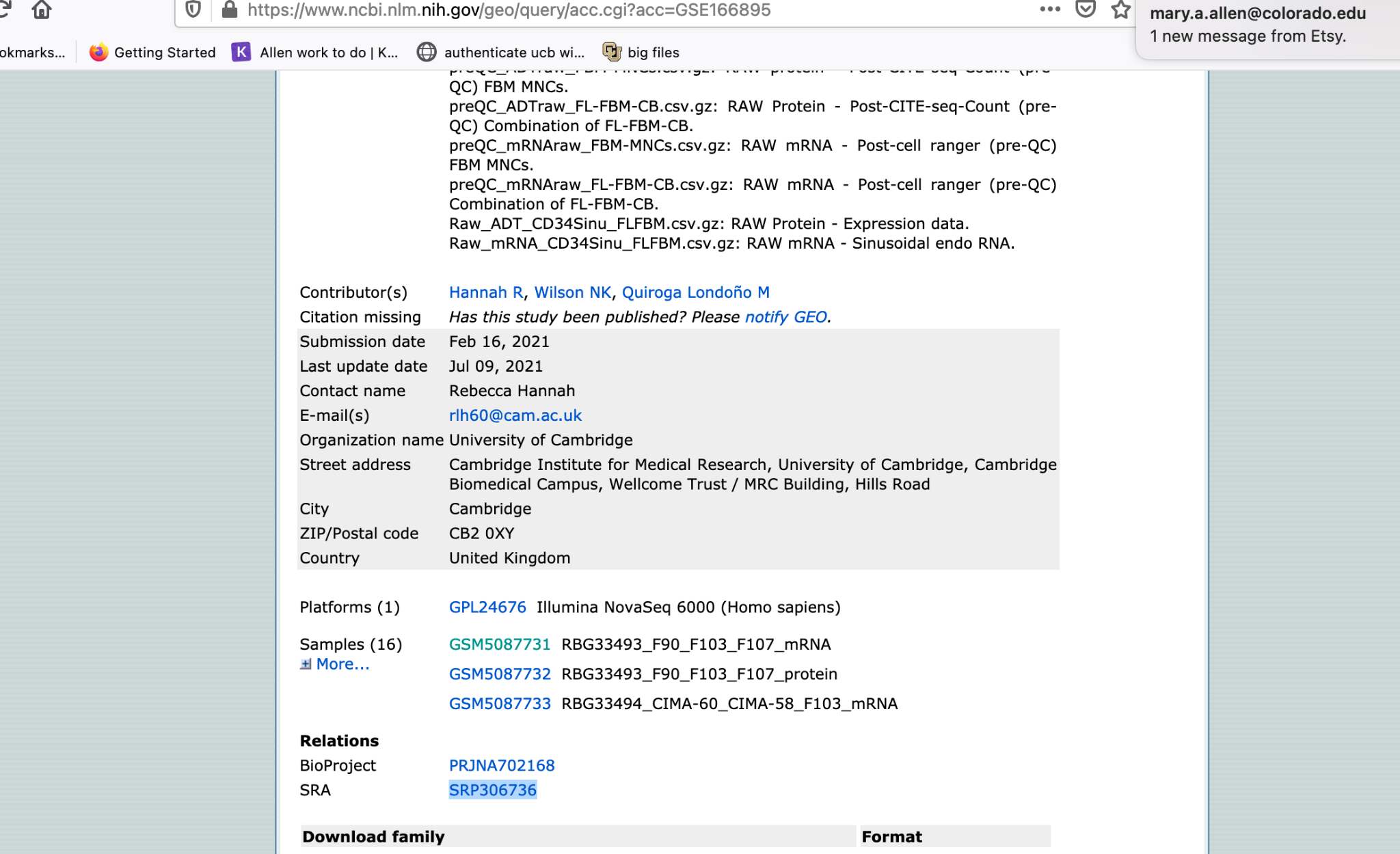
## Getting a SRR# for the fastq you want

## Get the SRP# number

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





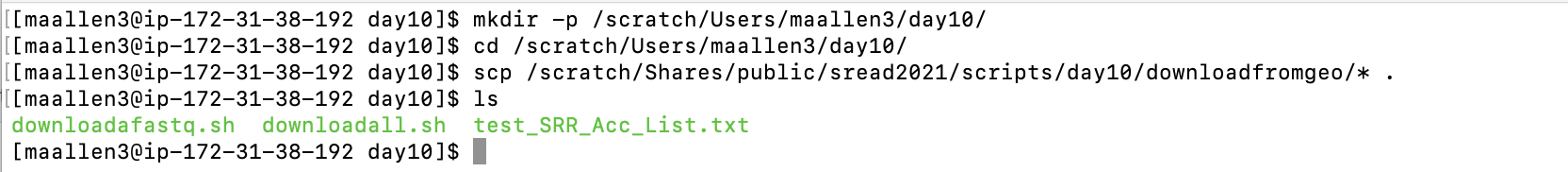


## Getting a list of all SRRs for this identifier.

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
   1. 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



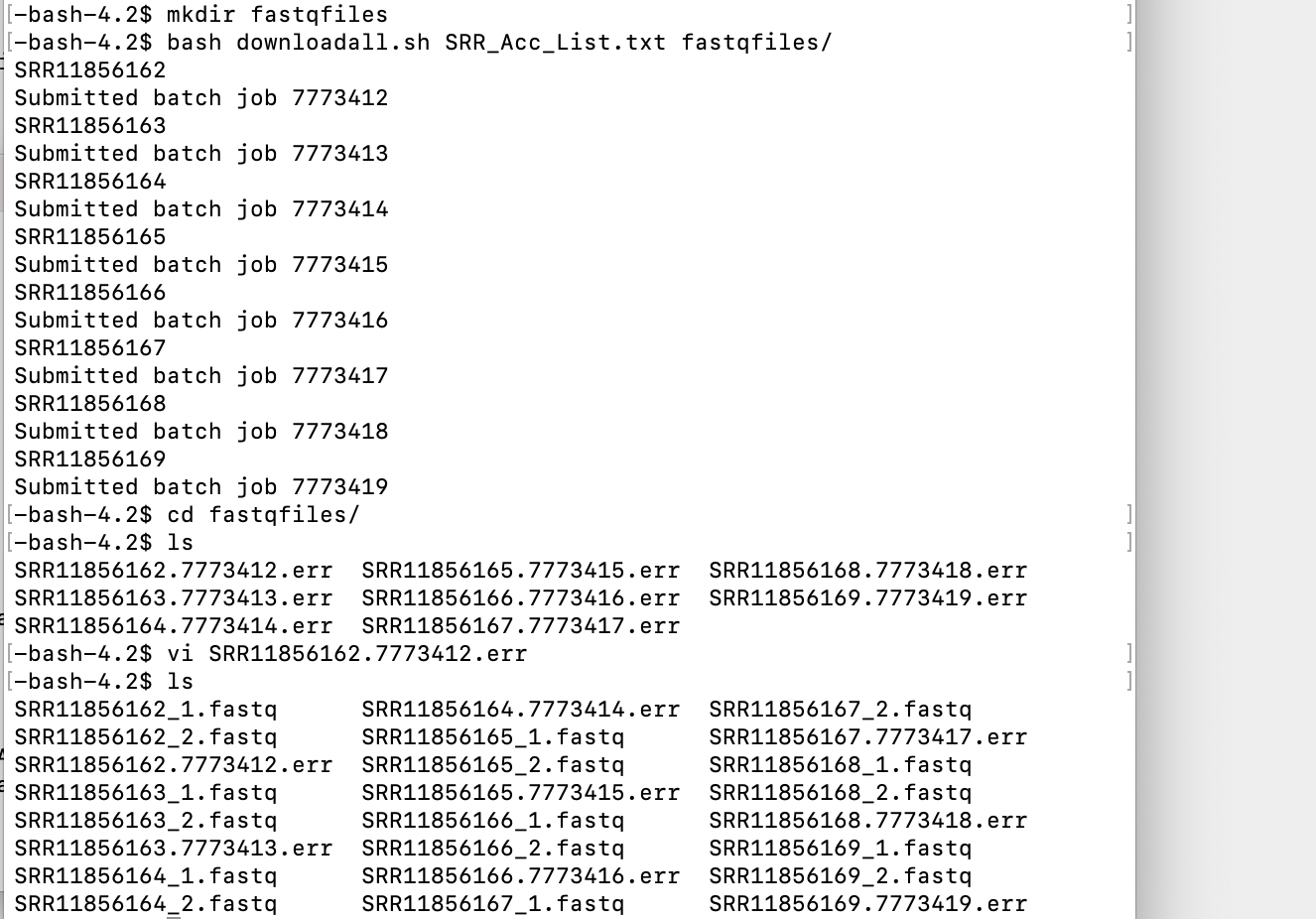
## Upload your SRR\_ACC\_List.txt to the super computer



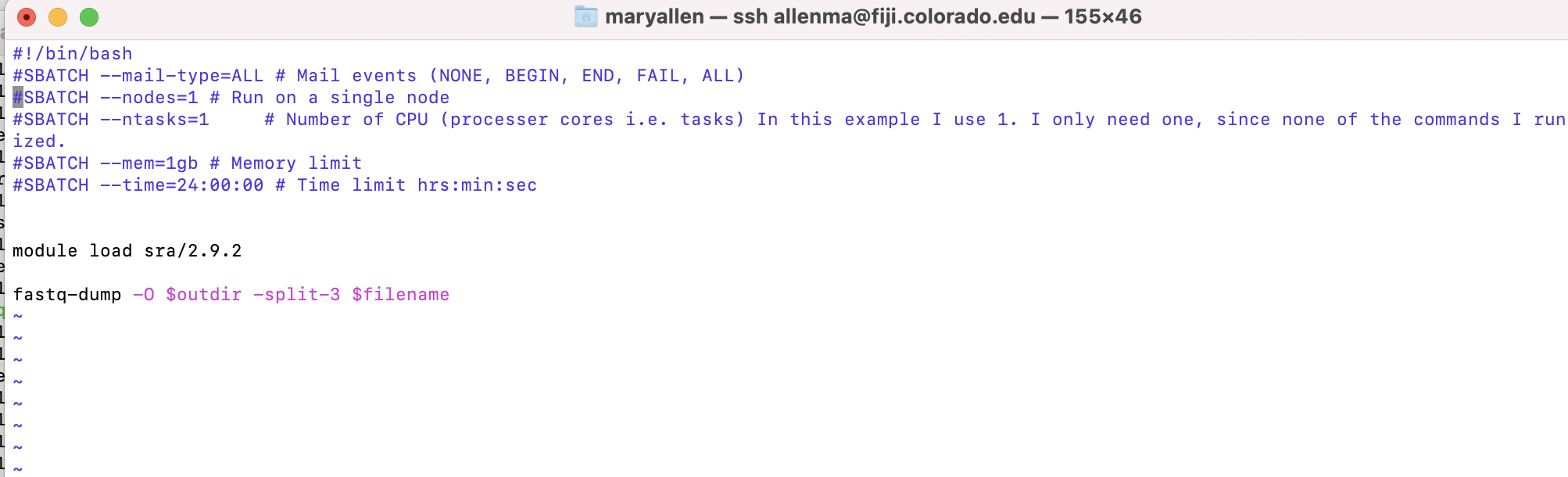
Run the sbatch script

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

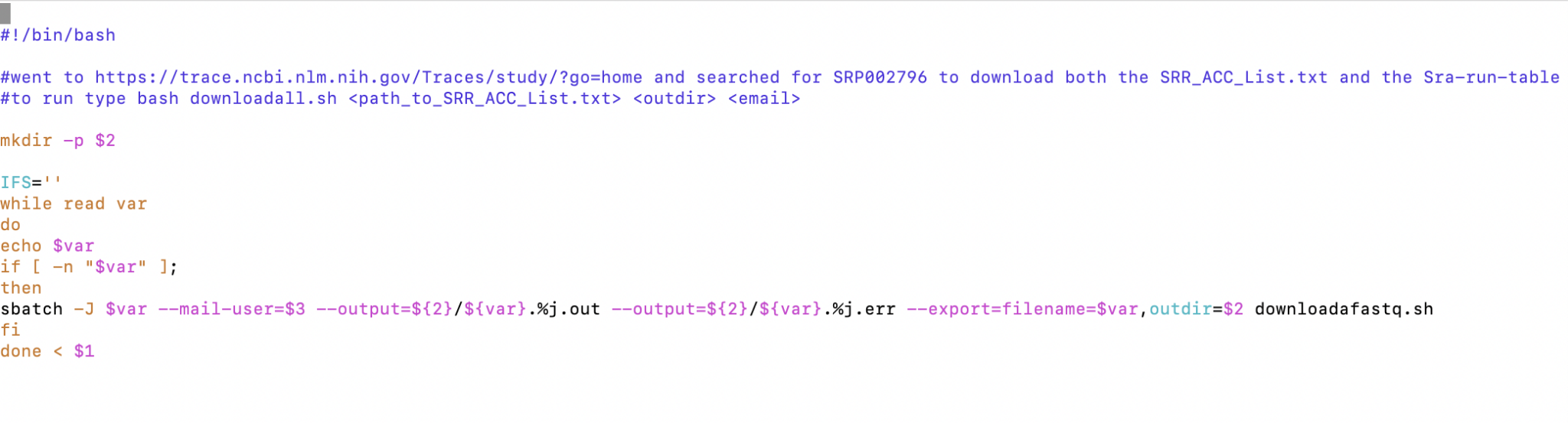
bash downloadall.sh <pathtoyour\_SRR\_AccList.txt> <outdir>



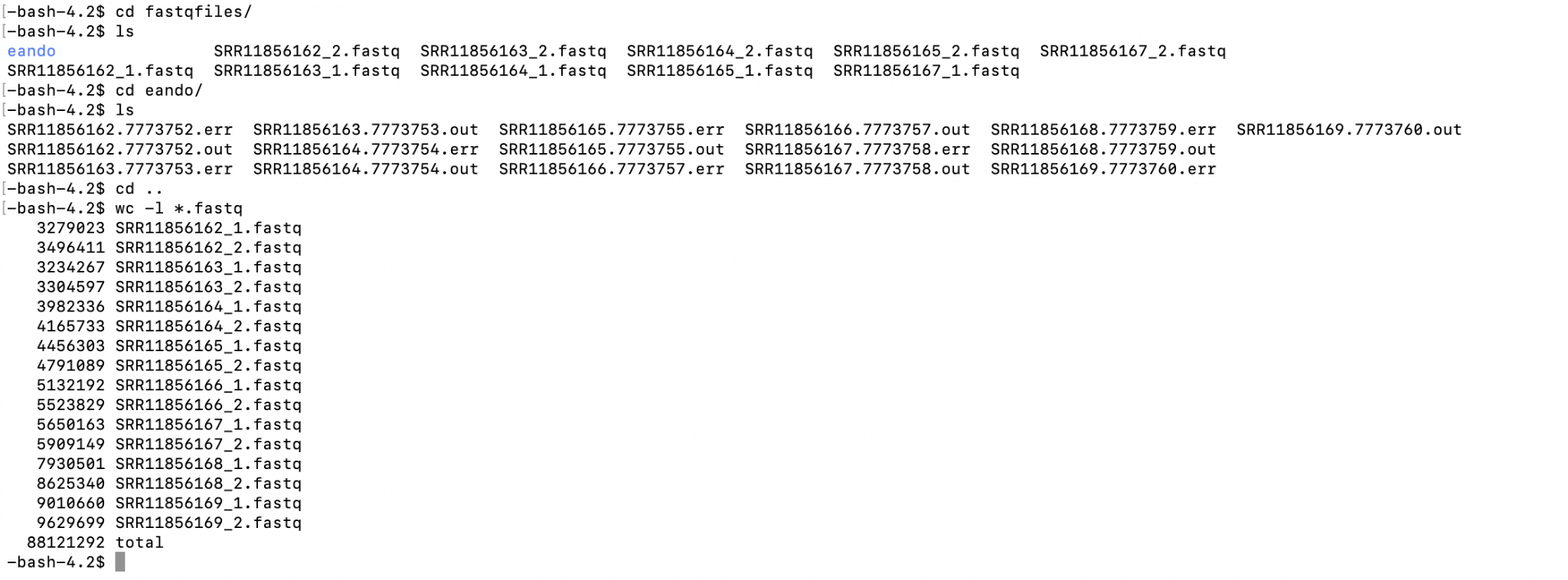
1. Under the hood
   1. The main script, called downloadfastq.sbatch, uses a program called fastq-dump to download public data from GEO.
      1. IMPORTANT!!! Use the -split-3 flag every time! If you are doing single end data that flag does nothing (it won’t hurt you), but if you are using paired end 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!!!!



* 1. This script, downloadall.sh, just runs the other script via sbatch.
     1. It reads the SRR\_ACC\_List.txt file and submits each SRR to a different CPU to download.



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



* 1. You should do a wc -l on the files to check them.
     1. 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) hg19 or hg38?

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