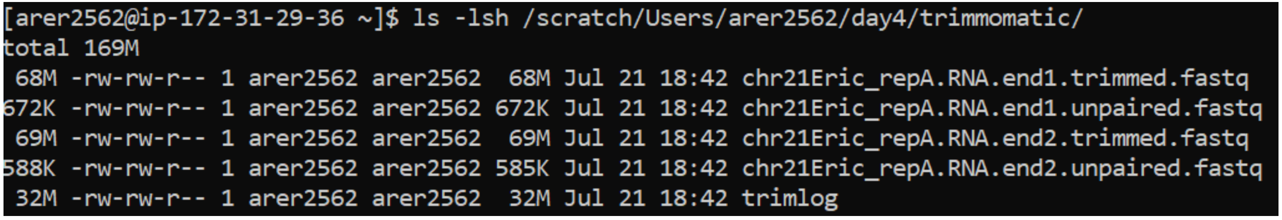
Day 4 Worksheet – Read mapping and visualization

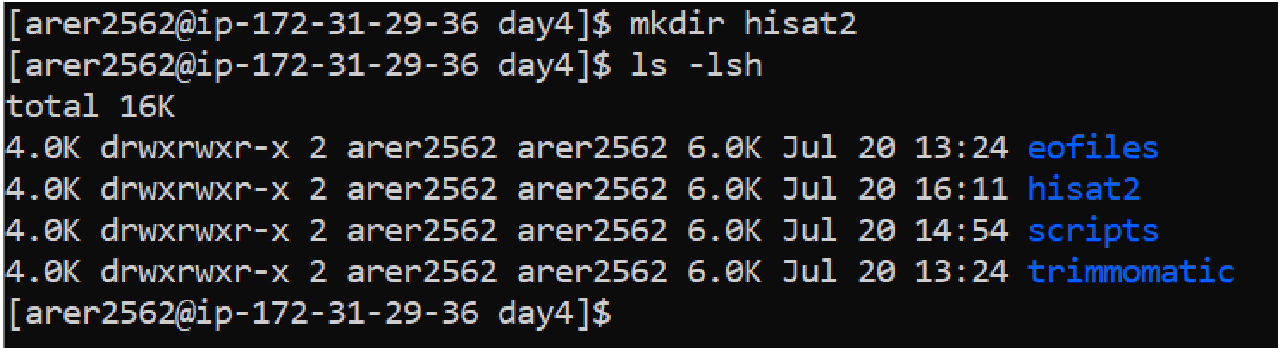
Author: Qing Yang, 2021
Edited: Ariel Eraso, Lynn Sanford
2023

1. Make sure you have the following files in your …**/day4/trimmomatic/** directory:



```
[arer2562@ip-172-31-29-36 ~]$ ls -lsh /scratch/Users/arer2562/day4/trimmomatic/
total 169M
 68M -rw-rw-r-- 1 arer2562 arer2562  68M Jul 21 18:42 chr21Eric_repA.RNA.end1.trimmed.fastq
672K -rw-rw-r-- 1 arer2562 arer2562 672K Jul 21 18:42 chr21Eric_repA.RNA.end1.unpaired.fastq
 69M -rw-rw-r-- 1 arer2562 arer2562  69M Jul 21 18:42 chr21Eric_repA.RNA.end2.trimmed.fastq
588K -rw-rw-r-- 1 arer2562 arer2562 585K Jul 21 18:42 chr21Eric_repA.RNA.end2.unpaired.fastq
 32M -rw-rw-r-- 1 arer2562 arer2562  32M Jul 21 18:42 trimlog
```

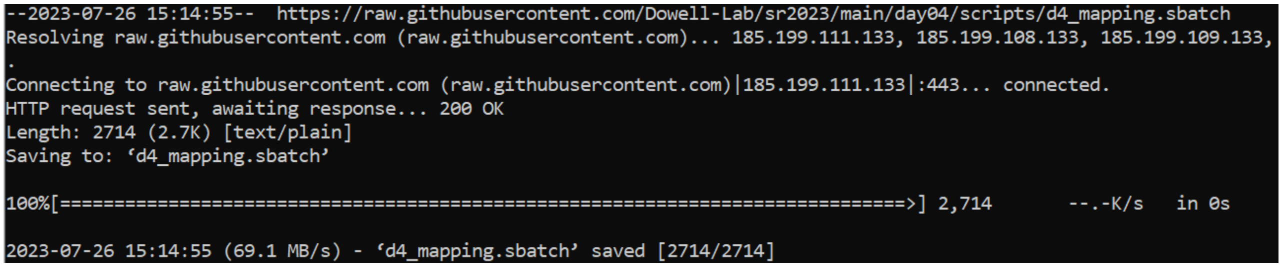
2. Create new directory (***mkdir***) named **hisat2**, under …**/day4/**, for the output directory for mapped reads.



```
[arer2562@ip-172-31-29-36 day4]$ mkdir hisat2
[arer2562@ip-172-31-29-36 day4]$ ls -lsh
total 16K
4.0K drwxrwxr-x 2 arer2562 arer2562 6.0K Jul 20 13:24 eofiles
4.0K drwxrwxr-x 2 arer2562 arer2562 6.0K Jul 20 16:11 hisat2
4.0K drwxrwxr-x 2 arer2562 arer2562 6.0K Jul 20 14:54 scripts
4.0K drwxrwxr-x 2 arer2562 arer2562 6.0K Jul 20 13:24 trimmomatic
[arer2562@ip-172-31-29-36 day4]$
```

3. IF your gitpull fails you can also wget the d4_mapping script from the sread2023 github to your scripts directory from raw.githubusercontent.com/Dowell-Lab/sr2023/main/day04/scripts/d4_mapping.sbatch

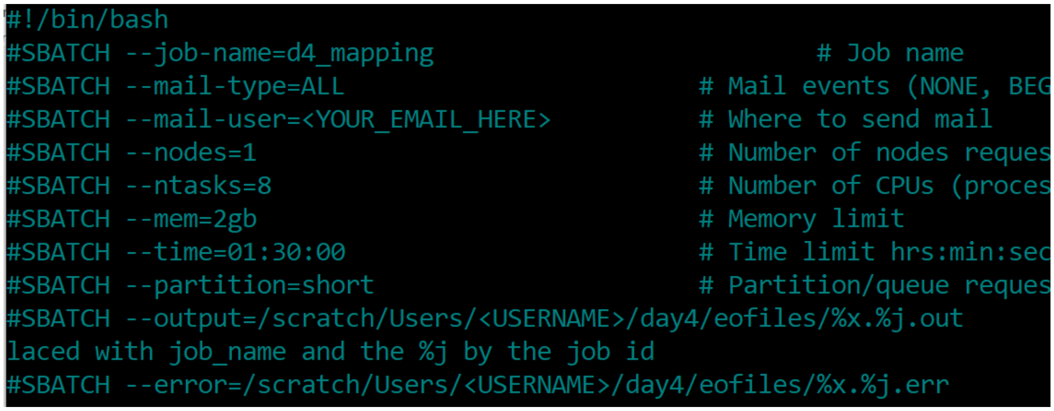
If successful screen should look like this:



```
--2023-07-26 15:14:55--  https://raw.githubusercontent.com/Dowell-Lab/sr2023/main/day04/scripts/d4_mapping.sbatch
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.133, 185.199.108.133, 185.199.109.133,
.
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 2714 (2.7K) [text/plain]
Saving to: ‘d4_mapping.sbatch’

100%[=========================================================================>] 2,714       --.-K/s   in 0s

2023-07-26 15:14:55 (69.1 MB/s) - ‘d4_mapping.sbatch’ saved [2714/2714]
```

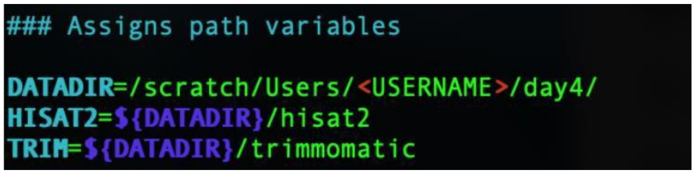
4. Edit the new “d4_mapping.sbatch” using the text editor **vim**. First, edit the SBATCH configuration to meet the needs of read mapping:

   a. Change the name of the job to something more useful, such as “hisat2_mapping”.
   b. Replace <EMAIL> with your own email address to which you want to receive any notifications.
   c. Replace <USERNAME> with your own username to complete the path directory to where to store the error and output files.
   d. Complete the following fields: nnodes, ntasks, mem and time. Hisat2 can use multiple processors per input file. So, 1 node, 8 tasks/processors/CPUs, 2 Gb for memory and 90 minutes for wall-time should be enough.



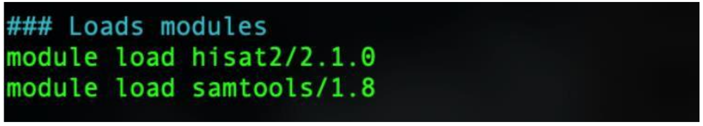
```
#!/bin/bash
#SBATCH --job-name=d4_mapping                            # Job name
#SBATCH --mail-type=ALL                          # Mail events (NONE, BEG
#SBATCH --mail-user=<YOUR_EMAIL_HERE>            # Where to send mail
#SBATCH --nodes=1                                # Number of nodes reques
#SBATCH --ntasks=8                               # Number of CPUs (proces
#SBATCH --mem=2gb                                # Memory limit
#SBATCH --time=01:30:00                          # Time limit hrs:min:sec
#SBATCH --partition=short                        # Partition/queue reques
#SBATCH --output=/scratch/Users/<USERNAME>/day4/eofiles/%x.%j.out
laced with job_name and the %j by the job id
#SBATCH --error=/scratch/Users/<USERNAME>/day4/eofiles/%x.%j.err
```

5.	Next, assign path variables. In this case, we will specify two directories, both under **DATADIR**. **TRIM** stores the directory path to trimmed reads. **HISAT2** stores the directory path to output mapped reads.



```
### Assigns path variables

DATADIR=/scratch/Users/<USERNAME>/day4/
HISAT2=${DATADIR}/hisat2
TRIM=${DATADIR}/trimmomatic
```

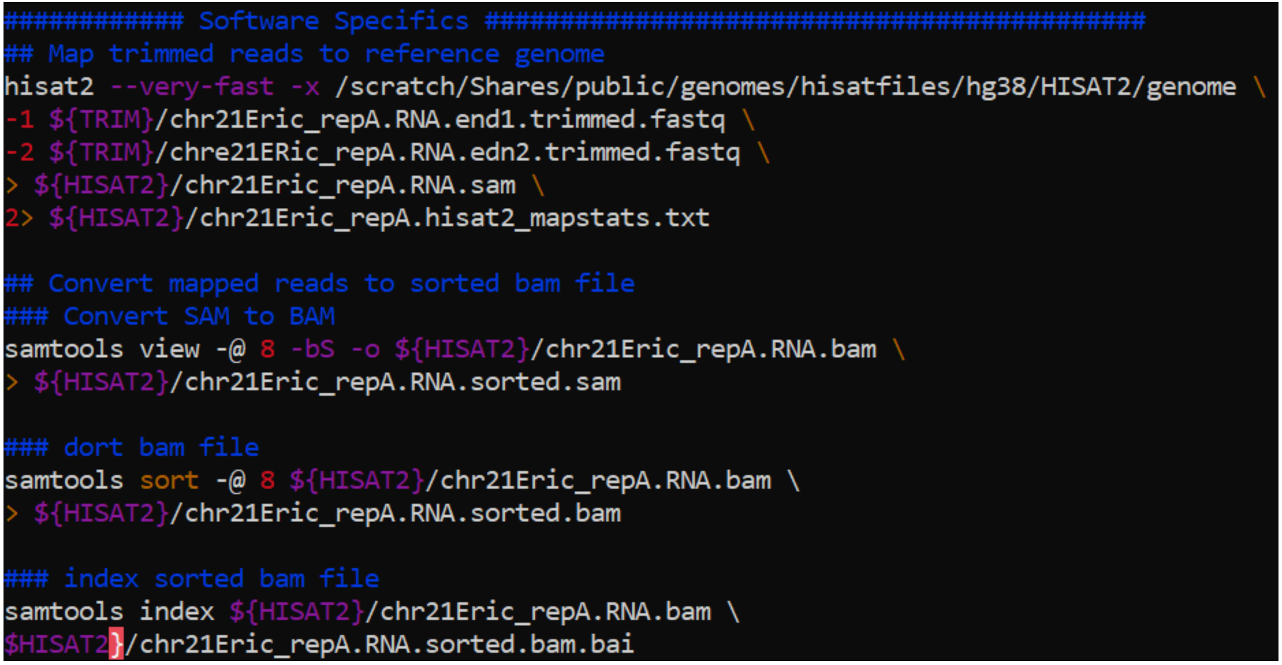
6. Next, load the modules/software needed for mapping reads and file conversion:



```
### Loads modules
module load hisat2/2.1.0
module load samtools/1.8
```

7. And finally, specify the read mapping and file conversion commands. Note that you could instead break up the command onto many lines using the character "\" at the end of every line. These \ characters are ignored by the computer, but will help you identify each part of the command more easily:

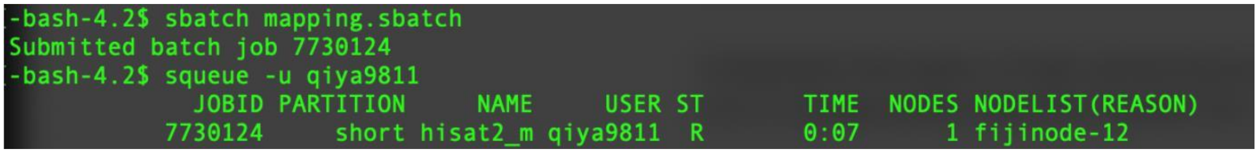
NOTE: The genome index is located at **/scratch/Shares/public/genomes/hisatfiles/hg38/HISAT2/genome**



```
############ Software Specifics ############################################
## Map trimmed reads to reference genome
hisat2 --very-fast -x /scratch/Shares/public/genomes/hisatfiles/hg38/HISAT2/genome \
-1 ${TRIM}/chr21Eric_repA.RNA.end1.trimmed.fastq \
-2 ${TRIM}/chre21ERic_repA.RNA.edn2.trimmed.fastq \
> ${HISAT2}/chr21Eric_repA.RNA.sam \
2> ${HISAT2}/chr21Eric_repA.hisat2_mapstats.txt

## Convert mapped reads to sorted bam file
### Convert SAM to BAM
samtools view -@ 8 -bS -o ${HISAT2}/chr21Eric_repA.RNA.bam \
> ${HISAT2}/chr21Eric_repA.RNA.sorted.sam

### dort bam file
samtools sort -@ 8 ${HISAT2}/chr21Eric_repA.RNA.bam \
> ${HISAT2}/chr21Eric_repA.RNA.sorted.bam

### index sorted bam file
samtools index ${HISAT2}/chr21Eric_repA.RNA.bam \
$HISAT2}/chr21Eric_repA.RNA.sorted.bam.bai
```

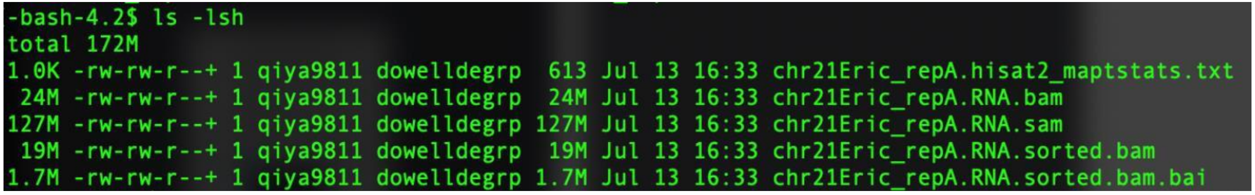
8. Before you close vim, make sure to save your edits by press Esc button to exit insertion mode, then type in ***:wq*** to save and quit vim.

9. Now that the job script is complete, submit the job by type in ***sbatch*** command. While waiting for the job to execute, you can check the job status using the command ***squeue -u <USERNAME>***:



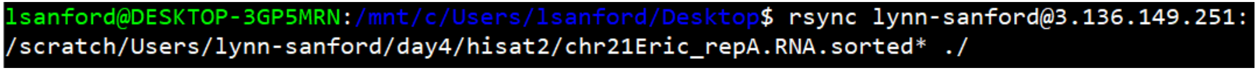
```
-bash-4.2$ sbatch mapping.sbatch
Submitted batch job 7730124
-bash-4.2$ squeue -u qiya9811
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           7730124     short hisat2_m qiya9811  R       0:07      1 fijinode-12
```

10. Finally, check the output directory …/day4/hisat2/ - there should be 5 different files:



```
-bash-4.2$ ls -lsh
total 172M
1.0K -rw-rw-r--+ 1 qiya9811 dowelldegrp  613 Jul 13 16:33 chr21Eric_repA.hisat2_maptstats.txt
 24M -rw-rw-r--+ 1 qiya9811 dowelldegrp  24M Jul 13 16:33 chr21Eric_repA.RNA.bam
127M -rw-rw-r--+ 1 qiya9811 dowelldegrp 127M Jul 13 16:33 chr21Eric_repA.RNA.sam
 19M -rw-rw-r--+ 1 qiya9811 dowelldegrp  19M Jul 13 16:33 chr21Eric_repA.RNA.sorted.bam
1.7M -rw-rw-r--+ 1 qiya9811 dowelldegrp 1.7M Jul 13 16:33 chr21Eric_repA.RNA.sorted.bam.bai
```

11. To visualize the mapped reads using IGV, you will need to transfer the sorted.bam and sorted.bam.bai files to your local machine. **rsync** the files from the AWS using a terminal on your local machine. Note that here, I've navigated to the directory for my desktop before rsyncing (Windows machine).



```
lsanford@DESKTOP-3GP5MRN:/mnt/c/Users/lsanford/Desktop$ rsync lynn-sanford@3.136.149.251:
/scratch/Users/lynn-sanford/day4/hisat2/chr21Eric_repA.RNA.sorted* ./
```