Khan Inan
BI-GY 7653 NGS
Week 8 hw assignment

Q1.1. Report the contents of your array job script and the job id on Greene [3 points].

```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=16GB
#SBATCH --job-name=RNAseq_align
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu
#SBATCH --array=1-8

module load samtools/intel/1.14
module load star/intel/2.7.6a

table=/scratch/work/courses/BI7653/hw8.2023/fastqs.txt

line="$(head -n $SLURM_ARRAY_TASK_ID $table | tail -n 1)"
sample="$(printf "%s" "${line}" | cut -f1)"
fq1="$(printf "%s" "${line}" | cut -f2)"
fq2="$(printf "%s" "${line}" | cut -f3)"

mkdir $sample
cd $sample



STAR --runThreadN $SLURM_CPUS_PER_TASK \
     --genomeDir /scratch/work/courses/BI7653/hw8.2023/STAR.genome \
     --readFilesIn /scratch/work/courses/BI7653/hw8.2023/fastqs/${sample}_1PE.fastq.gz /scratch/work/courses/BI7653/hw8.202
3/fastqs/${sample}_2PE.fastq.gz \
     --outFileNamePrefix $sample \
     --outSAMtype BAM SortedByCoordinate \
     --readFilesCommand zcat \
     --limitBAMsortRAM 20000000000 \
     --outTmpDir ${SLURM_TMPDIR}/${SLURM_ARRAY_JOB_ID}_${SLURM_ARRAY_TASK_ID}

samtools index ${sample}Aligned.sortedByCoord.out.bam
```

```
        16762189_8        cm RNAseq_a   bl2477  R       0:13      1 cm003
        16762189_7        cm RNAseq_a   bl2477  R       0:13      1 cm002
        16762189_6        cm RNAseq_a   bl2477  R       0:13      1 cm001
        16762189_5        cm RNAseq_a   bl2477  R       0:13      1 cm029
        16762189_4        cm RNAseq_a   bl2477  R       0:13      1 cm029
        16762189_3        cm RNAseq_a   bl2477  R       0:13      1 cm029
        16762189_2        cm RNAseq_a   bl2477  R       0:13      1 cm028
        16762189_1        cm RNAseq_a   bl2477  R       0:13      1 cm028
#jobID on Greene
```

Q1.2a Review the file “Log.final.out” for sample PDAC253 and report the following [1 point]:

1. The number of uniquely mapped reads
2. The percentage of uniquely mapped reads
3. The total number of input reads

```
1. The number of uniquely mapped reads= 5091972
2. The percentage of uniquely mapped reads=14.89%
3. The total number of input reads= 34202682
```

Q1.2b. If you want to make sure STAR output only uniquely mapped reads, how might you do this? What is the default mapping quality assigned in the SAM alignment records for uniquely mapped reads?

In order to make the output only unique reads, the command -outFilterMultimapNmax 1 can be used. The 1 is there because the default is 10 loci, but changing it to 1 displays only uniquely mapped reads. The default quality for mapping used in star for unique reads is the value of 255.

Q1.2c The number and percentage of reads mapped to too many loci is very high for this library. Provide a hypothesis for this observation and how you might go about evaluating it. [1 point]

Since the # and % of reads to the extra loci is high in the PDAC253 sample, it suggests that there are some rRNA that is negatively affecting the preparation of the library. This is most likely due to insufficient depletion. rRNA has many paralogs that are very alike, so as a result some rRNA may not be filtered out or removed during the preparation of the library, and as a result reads can become mapped to multiple locations. In order to get around this issue and determine if reads are being mapped to multiple locations due to not enough rRNA depletions, sequences are that mapped to different locations can be BLAST searched according to an rRNA database

Q1.3. Report the first 20 lines of the header for one output BAM (using samtools view). Then answer is your BAM coordinate-sorted? Please include your samtools view command in your answer for full credit. [1 point]?

```
module load samtools/intel/1.14
samtools view -H PDAC253Aligned.sortedByCoord.out.bam | head -n 20
# samtools view command to view the first 20 lines of the header for PDAC253 outout BAM

#First 20 lines of the header for PDAC253 output BAM:
@HD VN:1.4  SO:coordinate
@SQ SN:chr1 LN:40814151
@SQ SN:chr2 LN:29301675
@SQ SN:chr3 LN:24755689
@SQ SN:chr4 LN:33281721
@SQ SN:chr5 LN:18619412
@SQ SN:chr6 LN:18596258
@SQ SN:chr7 LN:16639383
@SQ SN:chr8 LN:31698078
@SQ SN:chr9 LN:22757669
@SQ SN:chr10    LN:15825318
@SQ SN:chr11    LN:29487722
@SQ SN:chr12    LN:14769854
@SQ SN:chr13    LN:12891333
@SQ SN:chr14    LN:24628924
@SQ SN:chr15    LN:12030914
@SQ SN:chr16    LN:13553361
@SQ SN:chr17    LN:16126437
@SQ SN:chr18    LN:9812533
@SQ SN:000007F  LN:4728343
```

So according to the first line, this file is indeed coordinate-sorted

Q1.4a What mapping quality scores are present in the alignment for PDAC253 (note: you may need to convert BAM to SAM)? [1 point]

```
samtools view -h PDAC253Aligned.sortedByCoord.out.bam > PDAC253Aligned.sortedByCoord.out.sam
#converting BAM file to SAM file using samtools
```

I saw that scores of 0, 1, 3 and 255 are present in the alignment

Q1.4b. According to the documentation, the uniquely mapped reads are defined by –outSAMmapqUnique and the mapping qualities for multiply mapped reads are: int(-10*log10(1-1/Nmap)). For each of the mapping qualities in Q1.4a, how many places do the reads with each of the other mapping qualities map to? Which mapping quality is assigned to uniquely mapped reads? [1 point]

So the quality score 255 suggests that the reads are unique while a score of 3 suggests that the reads have 2 loci. The score of 1 suggests that the reads have around 4-9 loci and lastly the score of 0 is the default score, which means it has 10 or more loci. I did not see a score of 2 which was missing but it essentially maps to 3 loci.

Q1.5. Imagine that you are working on a pair of recently duplicated genes and want to independently test for differential gene expression for the duplicated genes with the RNA-seq data in this assignment. Do you think this is possible? What factor(s) should be considered in order to do so? [1 point]

This is indeed possible and it can be done by counting the # of reads in the calculation/analyze step using a tool like htseq count. This tool creates a matric of the # of counts for each gene. In our casem is the duplicated genes are not exactly alike, then the two genes can be considered as separate genes and a matric can be created for the reads of each individual form. Afterwards, statistical analyzing can be done on the matrix to determine the differences in gene expression. If, however, the the copied genes are exactly alike, then we need to check if reads are mapped to multiple locations and labeled as multiply-mapped reads in STAR. Then the process is exactly the same where we can use the htseq count tool to generate matrixes and then determine the gene expression differential