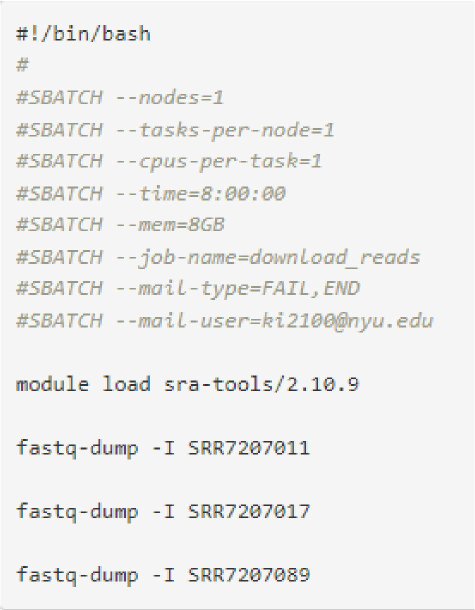
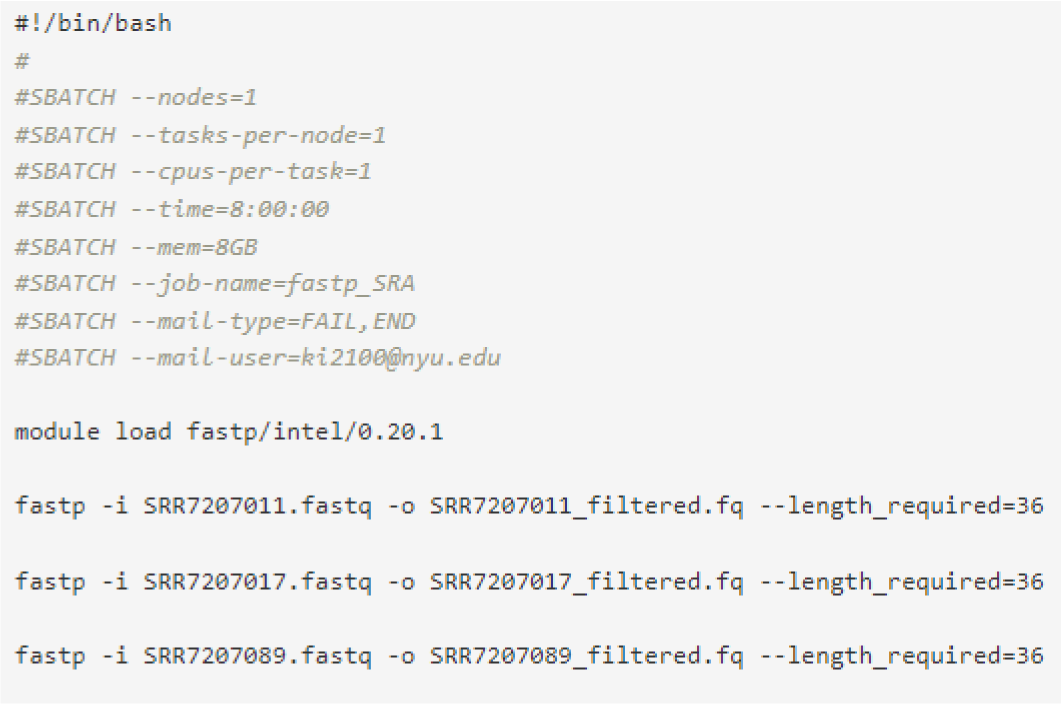
Khan Inan
BIGY 7653 NGS
Week 11 Assignment

Q1.1 Log in to Greene and create a slurm script that downloads the data. The script should load the most recent sra-tools module then download the three fastq file. Execute your script and report the contents and the names of the fastq files in your /scratch. [ 1 point ].



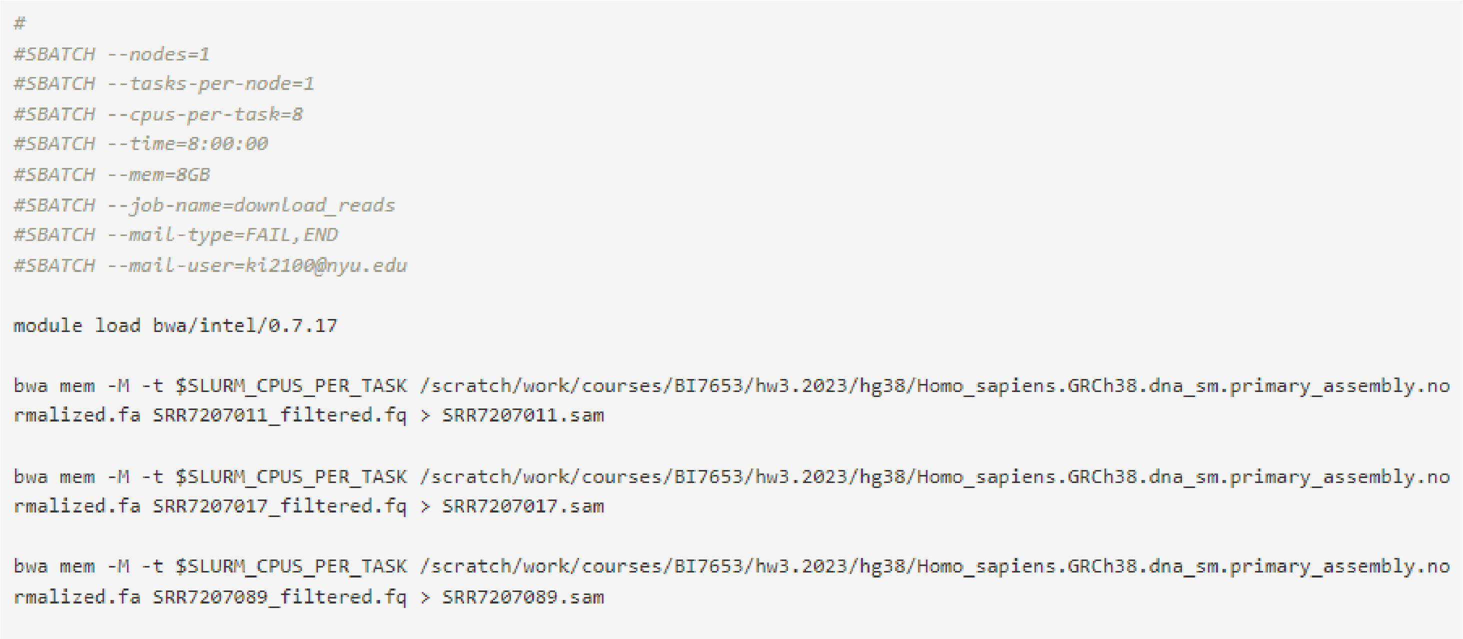
```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=download_reads
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu

module load sra-tools/2.10.9

fastq-dump -I SRR7207011

fastq-dump -I SRR7207017

fastq-dump -I SRR7207089
```

Q2.1 Create a slurm script that will process the single-end reads downloaded in Task 1 using fastp. Execute your script and report its contents for your answer [ 1 point ].



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

module load fastp/intel/0.20.1

fastp -i SRR7207011.fastq -o SRR7207011_filtered.fq --length_required=36

fastp -i SRR7207017.fastq -o SRR7207017_filtered.fq --length_required=36

fastp -i SRR7207089.fastq -o SRR7207089_filtered.fq --length_required=36
```

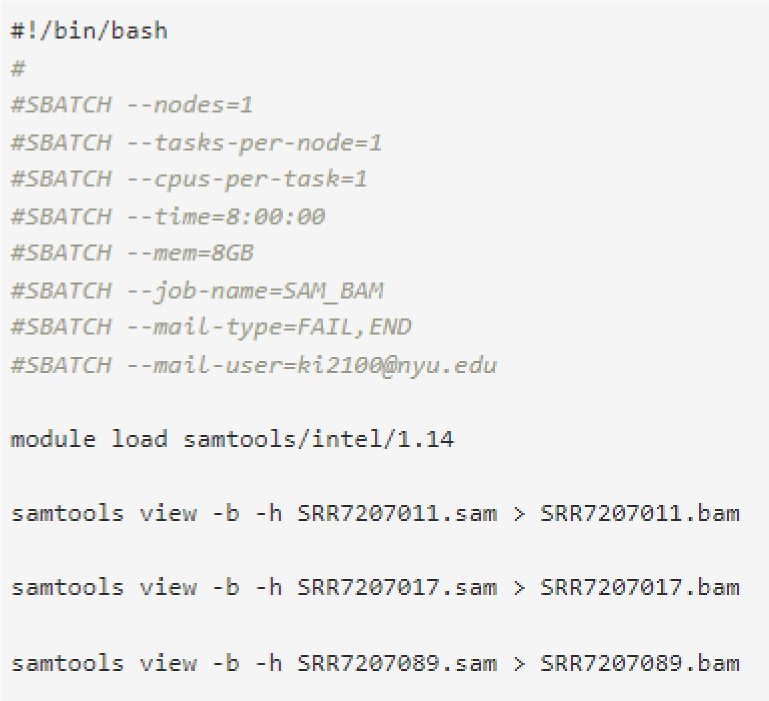
Q3.1. Execute your script and report its contents for your answer [ 1 point ].



```
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=download_reads
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu

module load bwa/intel/0.7.17

bwa mem -M -t $SLURM_CPUS_PER_TASK /scratch/work/courses/BI7653/hw3.2023/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.no
rmalized.fa SRR7207011_filtered.fq > SRR7207011.sam

bwa mem -M -t $SLURM_CPUS_PER_TASK /scratch/work/courses/BI7653/hw3.2023/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.no
rmalized.fa SRR7207017_filtered.fq > SRR7207017.sam

bwa mem -M -t $SLURM_CPUS_PER_TASK /scratch/work/courses/BI7653/hw3.2023/hg38/Homo_sapiens.GRCh38.dna_sm.primary_assembly.no
rmalized.fa SRR7207089_filtered.fq > SRR7207089.sam
```

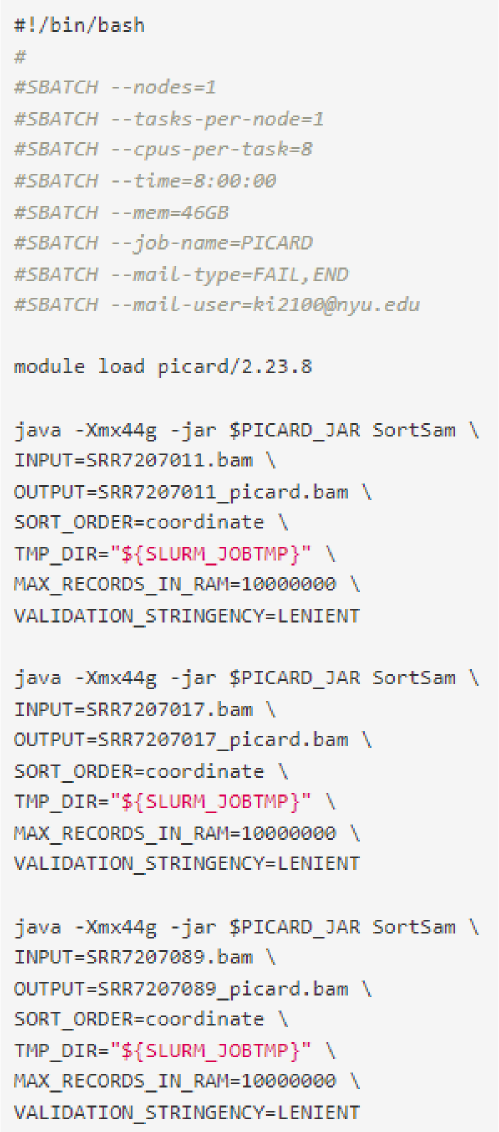
Q3.2. When your alignments from Q3.1 have completed, create a script that will conduct the following 3 steps to prepare the BAMs for downstream analysis. The expected output is 3 coordinate-sorted BAM files and 3 BAM index files. (note: the simplest approach is to just run each of the required commands below three times, once for SAM). Obviously, you will need to use appropriate output filenames at each step.

1st step:



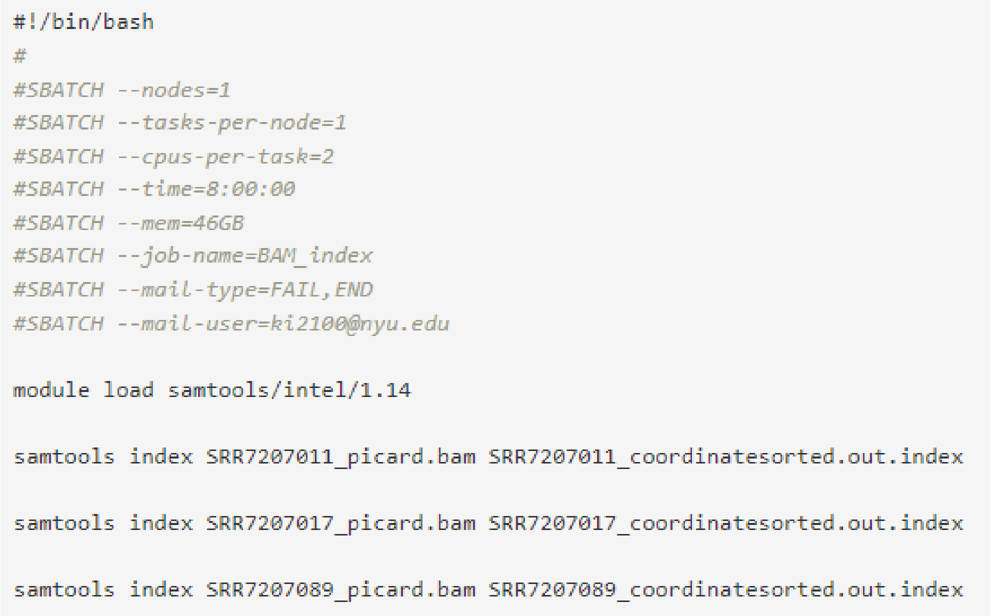
```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=SAM_BAM
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu

module load samtools/intel/1.14

samtools view -b -h SRR7207011.sam > SRR7207011.bam

samtools view -b -h SRR7207017.sam > SRR7207017.bam

samtools view -b -h SRR7207089.sam > SRR7207089.bam
```

2nd step



```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --time=8:00:00
#SBATCH --mem=46GB
#SBATCH --job-name=PICARD
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu

module load picard/2.23.8

java -Xmx44g -jar $PICARD_JAR SortSam \
INPUT=SRR7207011.bam \
OUTPUT=SRR7207011_picard.bam \
SORT_ORDER=coordinate \
TMP_DIR="${SLURM_JOBTMP}" \
MAX_RECORDS_IN_RAM=10000000 \
VALIDATION_STRINGENCY=LENIENT

java -Xmx44g -jar $PICARD_JAR SortSam \
INPUT=SRR7207017.bam \
OUTPUT=SRR7207017_picard.bam \
SORT_ORDER=coordinate \
TMP_DIR="${SLURM_JOBTMP}" \
MAX_RECORDS_IN_RAM=10000000 \
VALIDATION_STRINGENCY=LENIENT

java -Xmx44g -jar $PICARD_JAR SortSam \
INPUT=SRR7207089.bam \
OUTPUT=SRR7207089_picard.bam \
SORT_ORDER=coordinate \
TMP_DIR="${SLURM_JOBTMP}" \
MAX_RECORDS_IN_RAM=10000000 \
VALIDATION_STRINGENCY=LENIENT
```

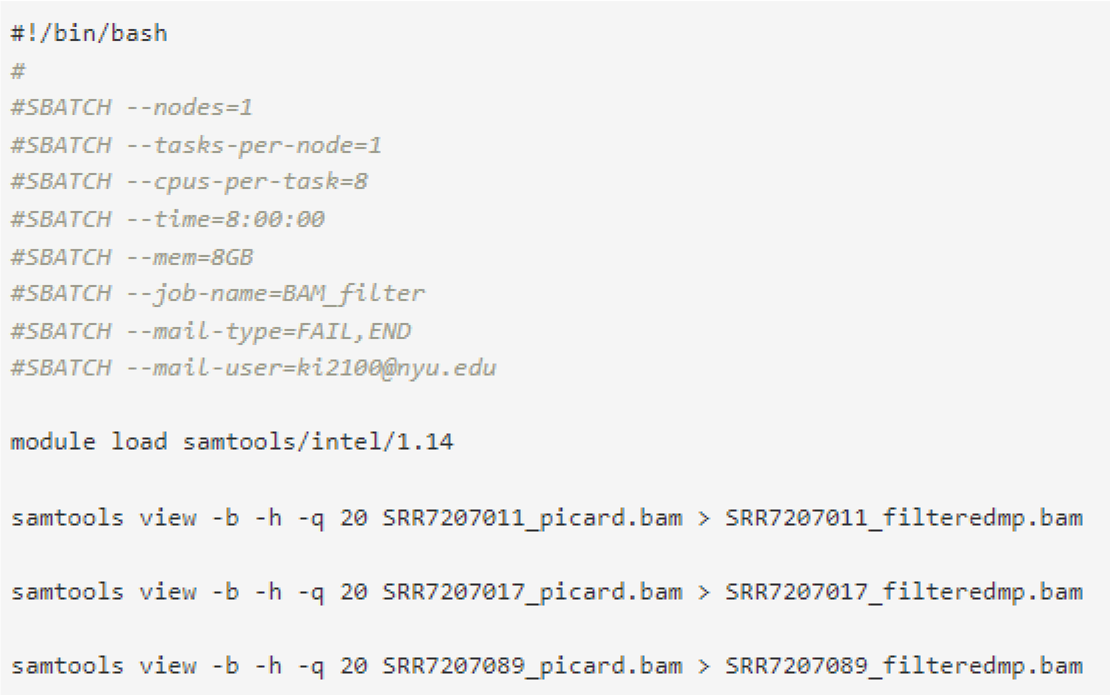
3rd step:



```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=2
#SBATCH --time=8:00:00
#SBATCH --mem=46GB
#SBATCH --job-name=BAM_index
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu

module load samtools/intel/1.14

samtools index SRR7207011_picard.bam SRR7207011_coordinatesorted.out.index

samtools index SRR7207017_picard.bam SRR7207017_coordinatesorted.out.index

samtools index SRR7207089_picard.bam SRR7207089_coordinatesorted.out.index
```

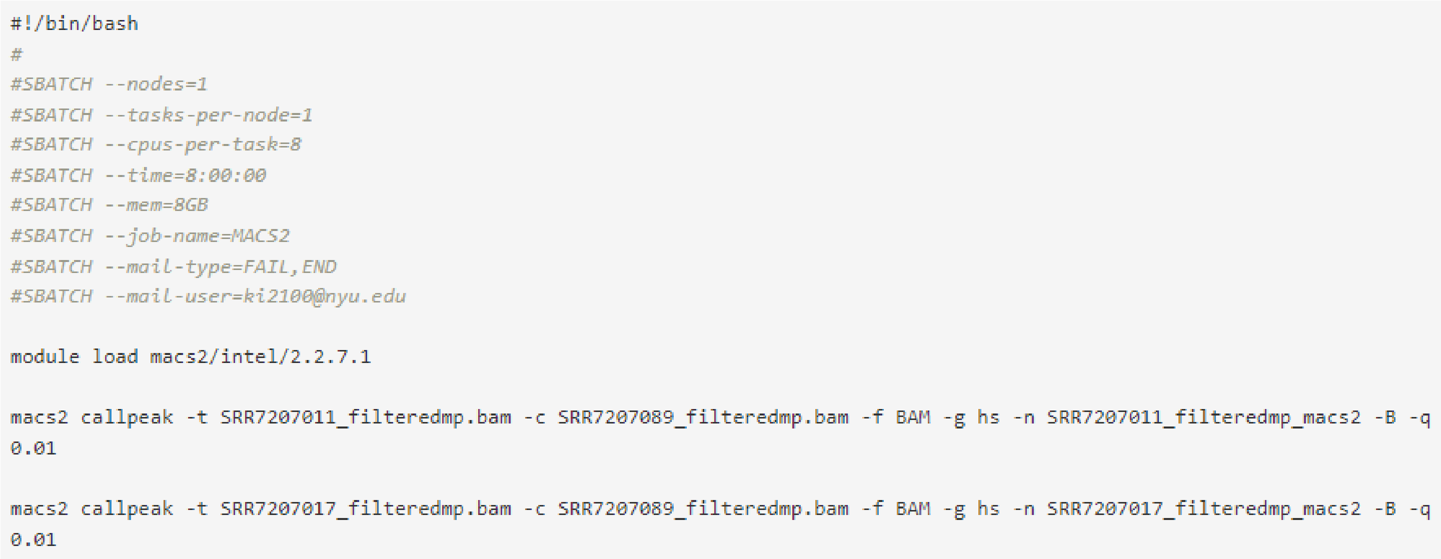
Q3.3 An important issue in ChIP-seq is whether to remove multiply-mapped reads. Allowing multiply-mapped reads in the analysis increases the number of reads and may increase the sensitivity of peak detection, but may also have a cost such as an increase in false positives. It is common to remove multiply-mapped reads in ChIP-seq analysis.

Execute your commands separately **on all three BAMs** in a slurm script to produce new BAMs with low mapping quality reads removed and their index files. Report your script for your answer [ 1 point ].



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

module load samtools/intel/1.14

samtools view -b -h -q 20 SRR7207011_picard.bam > SRR7207011_filteredmp.bam

samtools view -b -h -q 20 SRR7207017_picard.bam > SRR7207017_filteredmp.bam

samtools view -b -h -q 20 SRR7207089_picard.bam > SRR7207089_filteredmp.bam
```

Q4.1 Execute your script and report its contents for your answer. Retain your outputs for Week 12 assignment [ 1 point ].



```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=8
#SBATCH --time=8:00:00
#SBATCH --mem=8GB
#SBATCH --job-name=MACS2
#SBATCH --mail-type=FAIL,END
#SBATCH --mail-user=ki2100@nyu.edu

module load macs2/intel/2.2.7.1

macs2 callpeak -t SRR7207011_filteredmp.bam -c SRR7207089_filteredmp.bam -f BAM -g hs -n SRR7207011_filteredmp_macs2 -B -q
0.01

macs2 callpeak -t SRR7207017_filteredmp.bam -c SRR7207089_filteredmp.bam -f BAM -g hs -n SRR7207017_filteredmp_macs2 -B -q
0.01
```

Q4.2 Describe in your own words in three sentences or less the primary purpose of a transcription factor ChIP-seq experiment such as the one you are conducting this week and next [ 1 point ].

From what I understand, conducting a transcription chip-seq experiment is good for identifying, grouping and also analyzing the sites for binding on the genome. Additionally this experiment can be applied to DNA as well, and their protein counterparts. This type of experiment can also find the primary sequences of DNA for which the factor can bind to the sequence.

Q4.3. The human genome is highly repetitive. There are segmental duplications, low complexity sequence and transposable elements all of which may also harbor transcription factor binding sites. How reads map to these regions should always be a concern in NGS applications. In ChIP-seq it is common practice to remove multiply-mapped reads ('multireads'), but this is not always advisable and in fact Chung et al. 2011 argue for not removing multiply-mapped reads. Review Chung et al. 2011 https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002111 and select all statements below that are true. (note: most answers can be found in the Abstratct and Introduction) [ 1 point ].

a. **the primary problem with multiply mapped reads is that you dont know where in the genome the reads actually map**
b. **discarding multiply mapped reads could introduce false negatives (failure to detect true peaks)**
c. **many multiply mapped reads map to segmental duplication regions**
d. **including multiply-mapped reads in moderate to highly mappable regions can improve peak identification**
e. multiply mapped reads are not important because peaks should all be in non-repetitive parts of the genome

**The correct answer choices are A,B,C,D**