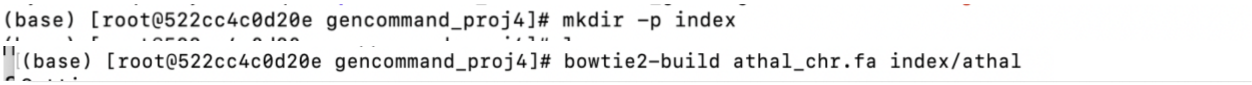
Project 4

For this project, it is recommended that you use the VMBox virtual environment (***Docker instead for MacOS with M1 or M2 chips) provided with the Course package and the tools therein. You may also use your own system and software, however make sure that appropriate versions are installed. The answers are compatible with the following versions of the software: samtools v.1.2, bowtie2 v.2.2.2, tophat v.2.0.14, and cufflinks/ cuffmerge/ cuffcompare/ cuffdiff v.2.2.1.

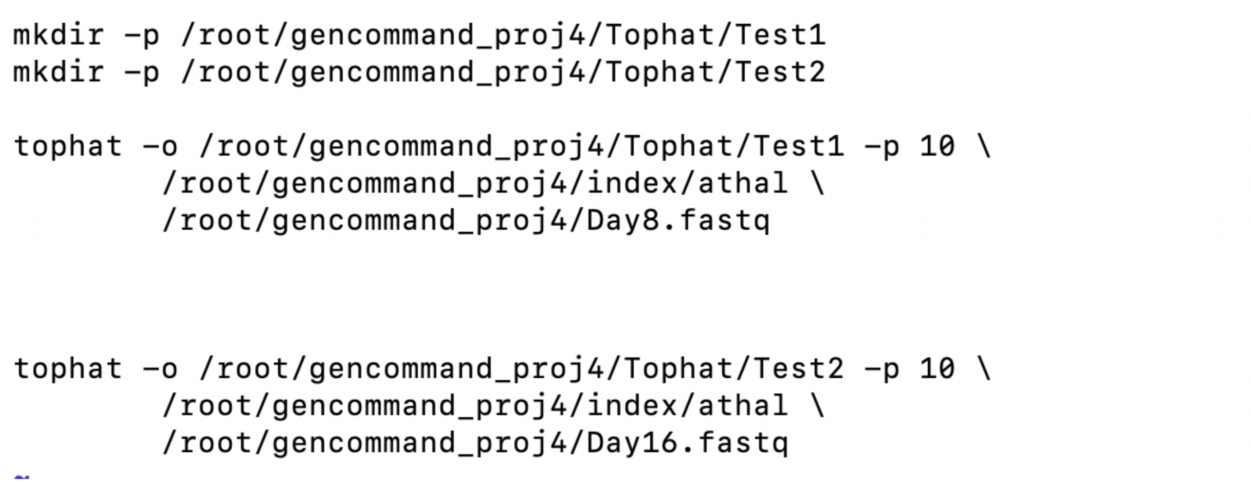
You are performing an RNA-seq experiment to determine genes that are differentially expressed at different stages in the development of Arabidopsis thaliana shoot apical meristem. You collected samples at day 8 and day 16 (files “Day8.fastq” and “Day16.fastq”), extracted and sequenced the cellular mRNA, and are now set to perform the bioinformatics analysis. The reference genome you will need for the analysis is “athal_chr.fa” and the reference gene annotations are in “athal_genes.gtf”. Use default parameters unless otherwise specified. Sample command files that you can modify to create your own pipeline are provided in the file “commands.tar.gz”. All files are provided in the archive gencommand_proj4.tar.gz.

NOTE: The genome and annotation data were obtained and modified from the Arabidopsis Information Resources (TAIR) Database, and the RNA-seq reads were extracted from GenBank’s Short Read Archive (SRA).

Create a bowtie index of the genome using bowtie2-build, with the prefix ‘athal’. Include a copy of the reference genome with the name “athal.fa” in the index directory.



```
(base) [root@522cc4c0d20e gencommand_proj4]# mkdir -p index
(base) [root@522cc4c0d20e gencommand_proj4]# bowtie2-build athal_chr.fa index/athal
```

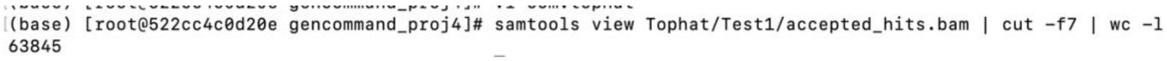


```
mkdir -p /root/gencommand_proj4/Tophat/Test1
mkdir -p /root/gencommand_proj4/Tophat/Test2

tophat -o /root/gencommand_proj4/Tophat/Test1 -p 10 \
        /root/gencommand_proj4/index/athal \
        /root/gencommand_proj4/Day8.fastq



tophat -o /root/gencommand_proj4/Tophat/Test2 -p 10 \
        /root/gencommand_proj4/index/athal \
        /root/gencommand_proj4/Day16.fastq
```

Apply to question 1-10.

Align both RNA-seq data sets to the reference genome using tophat. Analyze the results to answer the following questions.

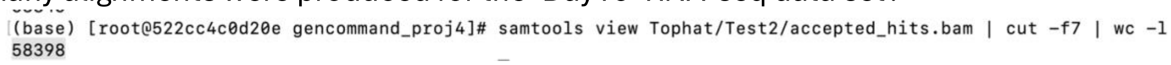
1. How many alignments were produced for the 'Day8' RNA-seq data set?
   a.



```
(base) [root@522cc4c0d20e gencommand_proj4]# samtools view Tophat/Test1/accepted_hits.bam | cut -f7 | wc -l
63845
```

   b.

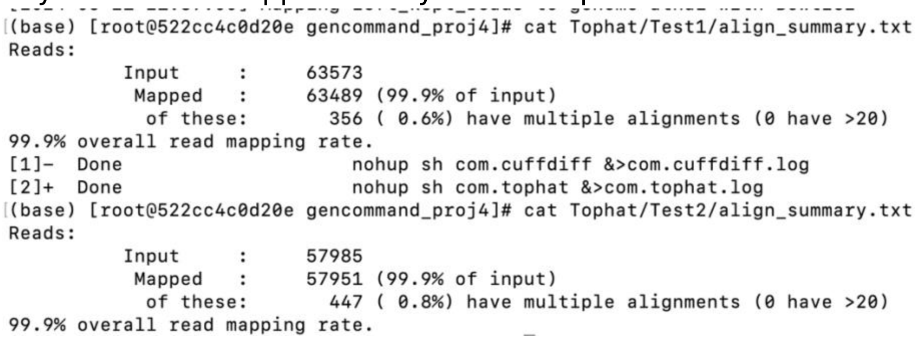
2. How many alignments were produced for the 'Day16' RNA-seq data set?



```
(base) [root@522cc4c0d20e gencommand_proj4]# samtools view Tophat/Test2/accepted_hits.bam | cut -f7 | wc -l
58398
```

   a.

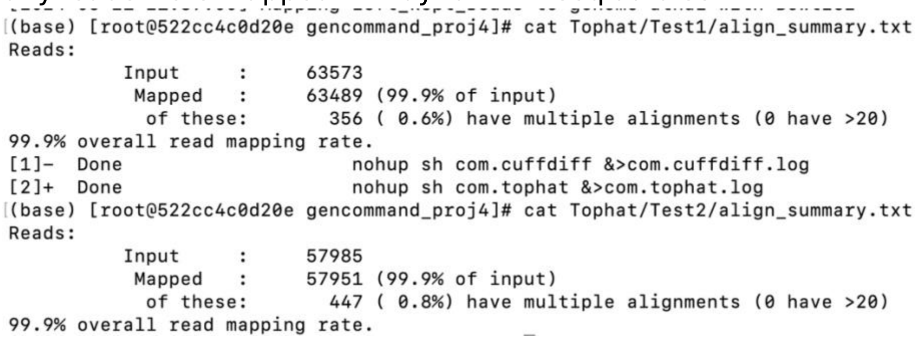
3. How many reads were mapped in 'Day8' RNA-seq data set?



```
(base) [root@522cc4c0d20e gencommand_proj4]# cat Tophat/Test1/align_summary.txt
Reads:
          Input     :     63573
           Mapped   :     63489 (99.9% of input)
            of these:       356 ( 0.6%) have multiple alignments (0 have >20)
99.9% overall read mapping rate.
[1]-  Done                    nohup sh com.cuffdiff &>com.cuffdiff.log
[2]+  Done                    nohup sh com.tophat &>com.tophat.log
(base) [root@522cc4c0d20e gencommand_proj4]# cat Tophat/Test2/align_summary.txt
Reads:
          Input     :     57985
           Mapped   :     57951 (99.9% of input)
            of these:       447 ( 0.8%) have multiple alignments (0 have >20)
99.9% overall read mapping rate.
```

   a.
   b. 63489

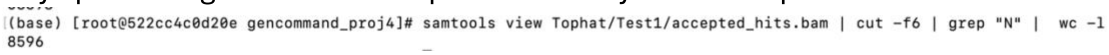
4. How many reads were mapped in 'Day16' RNA-seq data set?



```
(base) [root@522cc4c0d20e gencommand_proj4]# cat Tophat/Test1/align_summary.txt
Reads:
          Input     :     63573
           Mapped   :     63489 (99.9% of input)
            of these:       356 ( 0.6%) have multiple alignments (0 have >20)
99.9% overall read mapping rate.
[1]-  Done                    nohup sh com.cuffdiff &>com.cuffdiff.log
[2]+  Done                    nohup sh com.tophat &>com.tophat.log
(base) [root@522cc4c0d20e gencommand_proj4]# cat Tophat/Test2/align_summary.txt
Reads:
          Input     :     57985
           Mapped   :     57951 (99.9% of input)
            of these:       447 ( 0.8%) have multiple alignments (0 have >20)
99.9% overall read mapping rate.
```

   a.
   b. 57951

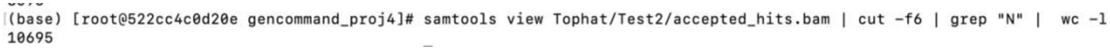
5. How many reads were uniquely aligned in 'Day8' RNA-seq data set?
   a. 63133
6. How many reads were uniquely aligned in 'Day16' RNA-seq data set?
   a. 57504
7. How many spliced alignments were reported for 'Day8' RNA-seq data set?



```
(base) [root@522cc4c0d20e gencommand_proj4]# samtools view Tophat/Test1/accepted_hits.bam | cut -f6 | grep "N" |  wc -l
8596
```

   a.

8. How many spliced alignments were reported for 'Day16' RNA-seq data set?



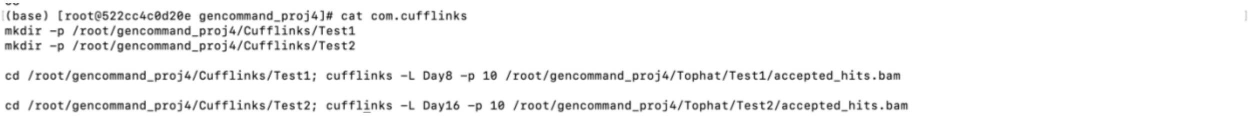
```
(base) [root@522cc4c0d20e gencommand_proj4]# samtools view Tophat/Test2/accepted_hits.bam | cut -f6 | grep "N" |  wc -l
10695
```

   a.

9. How many reads were left unmapped from 'Day8' RNA-seq data set?
   a. 84
10. How many reads were left unmapped from 'Day16' RNA-seq data set?
    a. 34

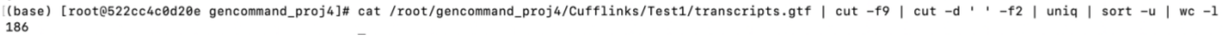
Apply to question 11-20.

Assemble the aligned RNA-seq reads into genes and transcripts using cufflinks. Use the labels ‘Day8’ and ‘Day16’, respectively, when creating identifiers. For this portion of the analysis, answer the following questions.



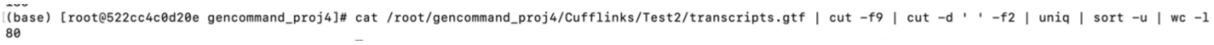
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat com.cufflinks
mkdir -p /root/gencommand_proj4/Cufflinks/Test1
mkdir -p /root/gencommand_proj4/Cufflinks/Test2

cd /root/gencommand_proj4/Cufflinks/Test1; cufflinks -L Day8 -p 10 /root/gencommand_proj4/Tophat/Test1/accepted_hits.bam

cd /root/gencommand_proj4/Cufflinks/Test2; cufflinks -L Day16 -p 10 /root/gencommand_proj4/Tophat/Test2/accepted_hits.bam
```

11. How many genes were generated by cufflinks for Day8?

    a. 

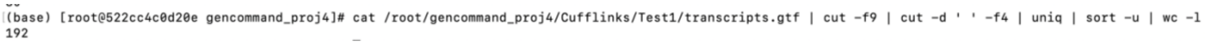
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/transcripts.gtf | cut -f9 | cut -d ' ' -f2 | uniq | sort -u | wc -l
186
```

12. How many genes were generated by cufflinks for Day16?

    a. 

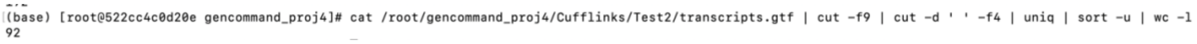
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/transcripts.gtf | cut -f9 | cut -d ' ' -f2 | uniq | sort -u | wc -l
80
```

13. How many transcripts were reported for Day8?

    a. 

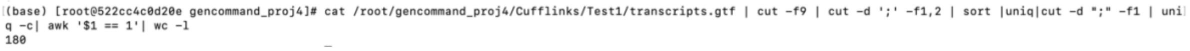
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/transcripts.gtf | cut -f9 | cut -d ' ' -f4 | uniq | sort -u | wc -l
192
```

14. How many transcripts were reported for Day16?

    a. 

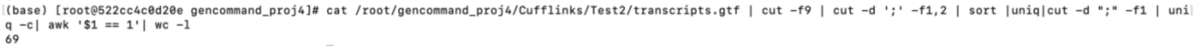
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/transcripts.gtf | cut -f9 | cut -d ' ' -f4 | uniq | sort -u | wc -l
92
```

15. How many single transcript genes were produced for Day8?

    a. 

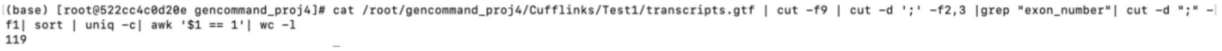
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/transcripts.gtf | cut -f9 | cut -d ';' -f1,2 | sort |uniq|cut -d ";" -f1 | uni
q -c| awk '$1 == 1'| wc -l
180
```

16. How many single transcript genes were produced for Day16?

    a. 

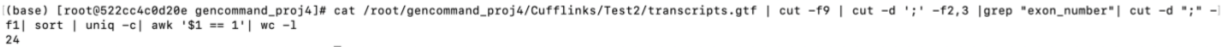
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/transcripts.gtf | cut -f9 | cut -d ';' -f1,2 | sort |uniq|cut -d ";" -f1 | uni
q -c| awk '$1 == 1'| wc -l
69
```

17. How many single-exon transcripts were in the Day8 set?

    a. 

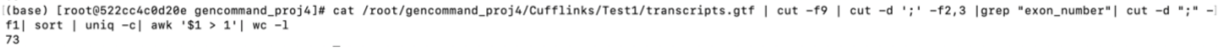
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/transcripts.gtf | cut -f9 | cut -d ';' -f2,3 |grep "exon_number"| cut -d ";" -
f1| sort | uniq -c| awk '$1 == 1'| wc -l
119
```

18. How many single-exon transcripts were in the Day16 set?

    a. 

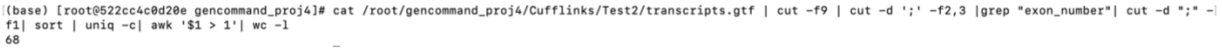
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/transcripts.gtf | cut -f9 | cut -d ';' -f2,3 |grep "exon_number"| cut -d ";" -
f1| sort | uniq -c| awk '$1 == 1'| wc -l
24
```

19. How many multi-exon transcripts were in the Day8 set?

    a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/transcripts.gtf | cut -f9 | cut -d ';' -f2,3 |grep "exon_number"| cut -d ";" -
f1| sort | uniq -c| awk '$1 > 1'| wc -l
73
```

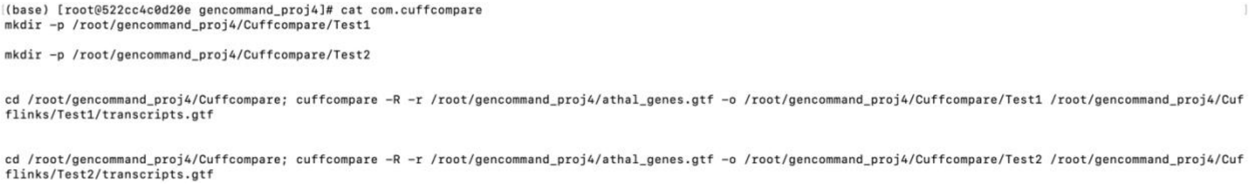
20. How many multi-exon transcripts were in the Day16 set?

    a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/transcripts.gtf | cut -f9 | cut -d ';' -f2,3 |grep "exon_number"| cut -d ";" -
f1| sort | uniq -c| awk '$1 > 1'| wc -l
68
```

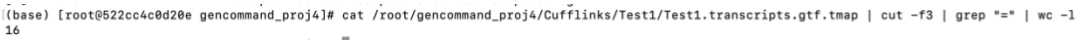
Apply to question 21-30.

Run cuffcompare on the resulting cufflinks transcripts, using the reference gene annotations provided and selecting the option '-R' to consider only the reference transcripts that overlap some input transfrag. For this step, using the *.tmap files answer the following, for both sets.



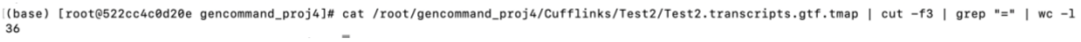
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat com.cuffcompare
mkdir -p /root/gencommand_proj4/Cuffcompare/Test1

mkdir -p /root/gencommand_proj4/Cuffcompare/Test2


cd /root/gencommand_proj4/Cuffcompare; cuffcompare -R -r /root/gencommand_proj4/athal_genes.gtf -o /root/gencommand_proj4/Cuffcompare/Test1 /root/gencommand_proj4/Cuf
flinks/Test1/transcripts.gtf


cd /root/gencommand_proj4/Cuffcompare; cuffcompare -R -r /root/gencommand_proj4/athal_genes.gtf -o /root/gencommand_proj4/Cuffcompare/Test2 /root/gencommand_proj4/Cuf
flinks/Test2/transcripts.gtf
```

21. How many cufflinks transcripts fully reconstruct annotation transcripts in Day8?

    a. 

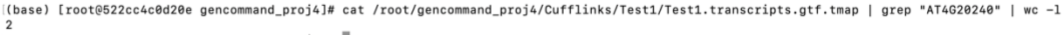
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/Test1.transcripts.gtf.tmap | cut -f3 | grep "=" | wc -l
16
```

22. How many cufflinks transcripts fully reconstruct annotation transcripts in Day16?

    a. 

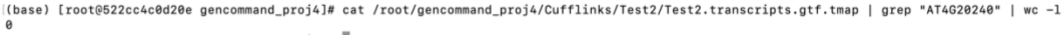
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/Test2.transcripts.gtf.tmap | cut -f3 | grep "=" | wc -l
36
```

23. How many splice variants does the gene AT4G20240 have in the Day8 sample?

a. 

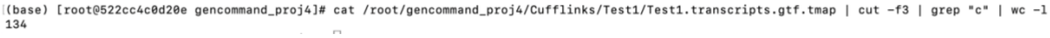
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/Test1.transcripts.gtf.tmap | grep "AT4G20240" | wc -l
2
```

24. How many splice variants does the gene AT4G20240 have in the Day16 sample?

a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/Test2.transcripts.gtf.tmap | grep "AT4G20240" | wc -l
0
```

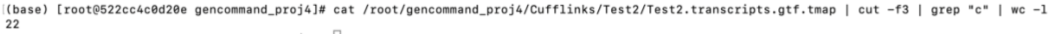
25. How many cufflinks transcripts are partial reconstructions of reference transcripts (‘contained’)? (Day8)

a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/Test1.transcripts.gtf.tmap | cut -f3 | grep "c" | wc -l
134
```

i. Substract -1 because of header = 133

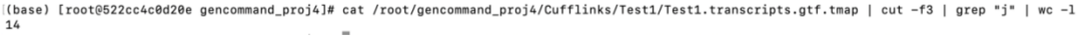
26. How many cufflinks transcripts are partial reconstructions of reference transcripts (‘contained’)? (Day16)

a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/Test2.transcripts.gtf.tmap | cut -f3 | grep "c" | wc -l
22
```

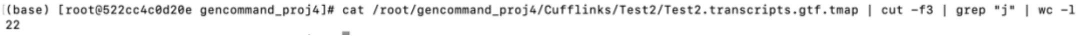
i. Substract -1 because of header = 21

27. How many cufflinks transcripts are novel splice variants of reference genes? (Day8)

a. 

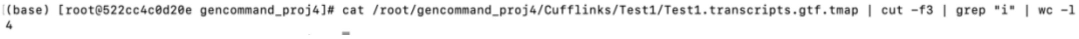
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/Test1.transcripts.gtf.tmap | cut -f3 | grep "j" | wc -l
14
```

28. How many cufflinks transcripts are novel splice variants of reference genes? (Day16)

a. 

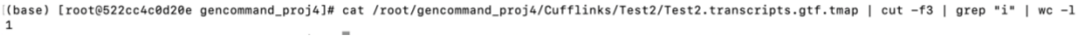
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/Test2.transcripts.gtf.tmap | cut -f3 | grep "j" | wc -l
22
```

29. How many cufflinks transcripts were formed in the introns of reference genes? (Day8)

a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test1/Test1.transcripts.gtf.tmap | cut -f3 | grep "i" | wc -l
4
```

30. How many cufflinks transcripts were formed in the introns of reference genes? (Day16)

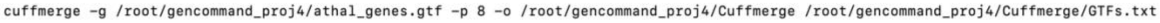
a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cufflinks/Test2/Test2.transcripts.gtf.tmap | cut -f3 | grep "i" | wc -l
1
```

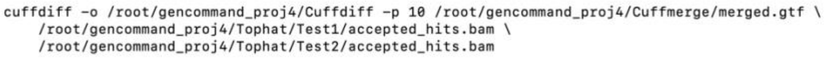
Apply to question 31-35.

Perform the differential gene expression analysis. For this, in a first stage run cuffmerge using the provided annotation to merge and reconcile the two sets of cufflinks transcripts. Make a note of the resulting file, ‘merged.gtf’. In a second step, use cufdiff to perform the differential expression analysis.

NOTE: Note that in general at least three replicates per condition are required to establish statistical significance. The single replicate example is provided here only to illustrate the analysis.

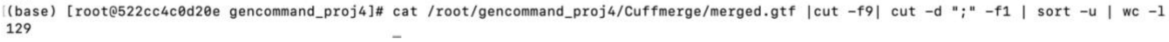


```
cuffmerge -g /root/gencommand_proj4/athal_genes.gtf -p 8 -o /root/gencommand_proj4/Cuffmerge /root/gencommand_proj4/Cuffmerge/GTFs.txt
```



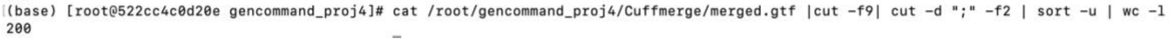
```
cuffdiff -o /root/gencommand_proj4/Cuffdiff -p 10 /root/gencommand_proj4/Cuffmerge/merged.gtf \
    /root/gencommand_proj4/Tophat/Test1/accepted_hits.bam \
    /root/gencommand_proj4/Tophat/Test2/accepted_hits.bam
```

31. How many genes (loci) were reported in the merged.gtf file?\

a. 

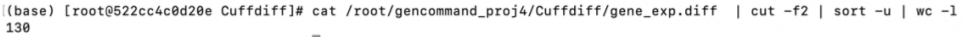
```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cuffmerge/merged.gtf |cut -f9| cut -d ";" -f1 | sort -u | wc -l
129
```

32. How many transcripts?

a. 

```
(base) [root@522cc4c0d20e gencommand_proj4]# cat /root/gencommand_proj4/Cuffmerge/merged.gtf |cut -f9| cut -d ";" -f2 | sort -u | wc -l
200
```

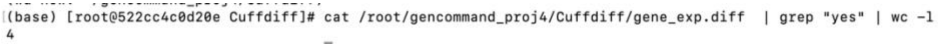
33. How many genes total were included in the gene expression report from cuffdiff?

a. 

```
(base) [root@522cc4c0d20e Cuffdiff]# cat /root/gencommand_proj4/Cuffdiff/gene_exp.diff  | cut -f2 | sort -u | wc -l
130
```

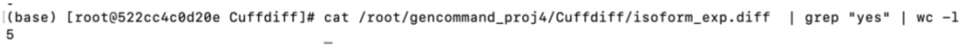
b. Subtract by 1 because of header

34. How many genes were detected as differentially expressed?

   a. 

```
(base) [root@522cc4c0d20e Cuffdiff]# cat /root/gencommand_proj4/Cuffdiff/gene_exp.diff  | grep "yes" | wc -l
4
```

35. How many transcripts were differentially expressed between the two samples?

   a. 

```
(base) [root@522cc4c0d20e Cuffdiff]# cat /root/gencommand_proj4/Cuffdiff/isoform_exp.diff  | grep "yes" | wc -l
5
```