Project 4:

Instructions:

- You are performing RNA-seq experiment: determine genes differentially expressed at different stages of development of an apical meristem
- collected samples (reads) at day 8 and day 16, extracted and sequenced cellular mRNA
- reference genome: athal_chr.fa
- reference gene annotations: athal_genes.gtf
- use default parameters unless otherwise specified
- commands.tar.gz: command files you can use to create your own pipeline

Quick review of what I gleaned from lecture:

- tophat maps the rna reads to the genome
- cufflinks assembles reads into transcripts
- cuffmerge puts all the transcripts together
- quantify isoform expression, see if they're different => cuffdiff

Questions 1-10: Align both RNA-seq datasets (Day16.fastq, Day8.fastq) to the reference genome (athal_chr.fa) using tophat

```
# first create the bowtie index and put it in its own directory
# bowtie2-build [options]* <reference_in> <bt2_index_base>
bowtie2-build athal_chr.fasta athal_index/athal
```

com.tophat

```
#!/bin/bash

# Set the working directory and Bowtie2 index
WORKDIR=./tophat_results/ # directory for tophat results
BT2INDEX=./athal_index/athal  # the bowtie index base name

# Create directories if they don't exist
mkdir -p "$WORKDIR"
mkdir -p "$WORKDIR/day_8"
mkdir -p "$WORKDIR/day_16"

# Tophat command for day 8
tophat -o "$WORKDIR/day_8" \  # output in results, day 8
  -p 10 \  # use 10 threads
  "$BT2INDEX" \  # bowtie index base name
  Day8.fastq  # reference reads

# Tophat command for day 16
```

```
tophat -o "$WORKDIR/day_16" \  # output in results, day 16
  -p 10 \  # use 10 threads
  "$BT2INDEX" \  # bowtie index base name
  Day16.fastq  # reference reads
```

1. **How many alignments were produced for the ‘Day8’ RNA-seq data set?**

(base) [root@12018efd4a72 day_8]# samtools view accepted_hits.bam | wc -l

63845

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

(base) [root@12018efd4a72 day_16]# samtools view accepted_hits.bam | wc -l

58398

3. **How many reads were mapped in ‘Day8’ RNA-seq data set?**

(base) [root@12018efd4a72 day_8]# more 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.

4. **How many reads were mapped in ‘Day16’ RNA-seq data set?**

(base) [root@12018efd4a72 day_16]# more 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.

5. **How many reads were uniquely aligned in ‘Day8’ RNA-seq data set?**

(base) [root@12018efd4a72 day_16]# expr 63489 - 356
63133

6. **How many reads were uniquely aligned in ‘Day16’ RNA-seq data set?**

(base) [root@12018efd4a72 day_16]# expr 57951 - 447
57504

7. **How many spliced alignments (contains introns) were reported for ‘Day8’ RNA-seq data set?**

(base) [root@12018efd4a72 day_8]# samtools view accepted_hits.bam | cut -f6 | grep -c "N"
8596

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

(base) [root@12018efd4a72 day_16]# samtools view accepted_hits.bam | cut -f6 | grep -c "N"
10695

9. **How many reads were left unmapped from ‘Day8’ RNA-seq data set?**

(base) [root@12018efd4a72 day_16]# expr 63573 - 63489
84

10. **How many reads were left unmapped from ‘Day16’ RNA-seq data set?**

(base) [root@12018efd4a72 day_16]# expr 57985 - 57951
34

Questions 11-20: Assemble RNA-seq read alignments into genes and transcripts using cufflinks. Use labels ‘Day8’ and ‘Day16’ for the identifiers.

com.cufflinks

```
#!/bin/bash

THDIR=/data/project4/tophat_results/
WORKDIR=/data/project4/cufflinks_results/

mkdir -p $WORKDIR/day_8
cd $WORKDIR/day_8
cufflinks -L Day8 -p 8 $THDIR/day_8/accepted_hits.bam
```

```
mkdir -p $WORKDIR/day_16
cd $WORKDIR/day_16
cufflinks -L Day16 -p 8 $THDIR/day_16/accepted_hits.bam
```

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

```
(base) [root@12018efd4a72 day_8]# more transcripts.gtf | cut -f9 | cut -d' ' -f2 | sort | uniq -c | wc -l
186
```

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

```
(base) [root@12018efd4a72 day_16]# more transcripts.gtf | cut -f9 | cut -d' ' -f2 | sort | uniq -c | wc -l
80
```

13. **How many transcripts were reported for Day8?**

```
(base) [root@12018efd4a72 day_8]# more transcripts.gtf | cut -f9 | cut -d' ' -f4 | sort | uniq -c | wc -l
192
```

14. **How many transcripts were reported for Day16?**

```
(base) [root@12018efd4a72 day_16]# more transcripts.gtf | cut -f9 | cut -d' ' -f4 | sort | uniq -c | wc -l
92
```

15. **How many single-exon transcripts on day 8?**
16. **How many single-exon transcripts on day 16?**

```
gene_id "Day16.3"; transcript_id "Day16.3.1"; exon_number "1"; FPKM "7930.9505871248"; frac "1.000000"; conf_lo "6147.517012"; conf_hi "9714.384162"; cov "22.980326";
```

- exon_number tells you how many exons in the transcript

```
(base) [root@12018efd4a72 day_8]# more transcripts.gtf | cut -f9 | cut -d';' -f1,2 | uniq | cut -d";" -f1 | uniq -c | awk '$1==1' | wc -l
180
```

```
(base) [root@12018efd4a72 day_16]# more transcripts.gtf | cut -f9 | cut -d';' -f1,2 | uniq |
cut -d";" -f1 | uniq -c | awk '$1==1' | wc -l
69
```

17. **How many single-exon transcripts were in the Day8 set?**
18. **How many single-exon transcripts were in the Day16 set?**

```
(base) [root@12018efd4a72 day_8]# more transcripts.gtf | cut -f9 | cut -d';' -f2,3 | grep
"exon_number" | cut -d ';' -f1 | uniq -c | awk '$1==1' | wc -l
119

(base) [root@12018efd4a72 day_16]# more transcripts.gtf | cut -f9 | cut -d';' -f2,3 | grep
"exon_number" | cut -d ';' -f1 | uniq -c | awk '$1==1' | wc -l
24
```

19. **How many multi-exon transcripts were in the Day8 set?**
20. **How many multi-exon transcripts were in the Day16 set?**

Why WRONG: -f2,3 should have been selected instead for the transcripts:

```
more transcripts.gtf | grep “exon number” | cut -f9 |cut -d';' -f2,3 | uniq |  cut -d";" -f1 |
uniq -c | awk ‘$1 > 1’ | wc -l  # would count transcript with exon number over 1.

(base) [root@12018efd4a72 day_8]# more transcripts.gtf | cut -f9 | cut -d';' -f1,2 | uniq |
cut -d";" -f1 | uniq -c | wc -l
186
(base) [root@12018efd4a72 day_8]# expr 186 - 180
6

(base) [root@12018efd4a72 day_16]# more transcripts.gtf | cut -f9 | cut -d';' -f1,2 | uniq |
cut -d";" -f1 | uniq -c | wc -l
80
(base) [root@12018efd4a72 day_16]# expr 80 - 69
11
```

Question 21-30: Use cuffcompare on the cufflinks transcripts, using the reference gene annotations, use option -R, create .tmap files

```
cuffcompare [-r <reference_mrna.gtf>] [-R] [-T] [-V] [-s <seq_path>]
   [-o <outprefix>] [-p <cprefix>]
   {-i <input_gtf_list> | <input1.gtf> [<input2.gtf> .. <inputN.gtf>]}
```

```
#!/bin/bash

ROOTDIR=/data/project4
CUFFLINKS=$ROOTDIR/cufflinks_results
WORKINGDIR=$ROOTDIR/cuffcompare_results

mkdir -p $WORKINGDIR/day_8
cuffcompare -r $ROOTDIR/athal_genes.gtf \
-R $CUFFLINKS/day_8/transcripts.gtf \
-o $WORKINGDIR/Day8

mkdir -p $WORKINGDIR/day_16
cuffcompare -r $ROOTDIR/athal_genes.gtf \
-R $CUFFLINKS/day_16/transcripts.gtf \
-o $WORKINGDIR/Day16
```

21. **How many cufflinks transcripts fully reconstruct annotation transcripts in Day8?**
22. **How many cufflinks transcripts fully reconstruct annotation transcripts in Day16?**

```
(base) [root@12018efd4a72 day_8]# more Day8.transcripts.gtf.tmap | cut -f3 | grep -c
"="
16

(base) [root@12018efd4a72 day_16]# more Day16.transcripts.gtf.tmap | cut -f3 | grep -c
"="
36
```

23. **How many splice variants does the gene AT4G20240 have in the Day8 sample?**
24. **How many splice variants does the gene AT4G20240 have in the Day16 sample?**

```
(base) [root@12018efd4a72 day_8]# more Day8.transcripts.gtf.tmap | grep
"AT4G20240"
AT4G20240   AT4G20240.1         o       Day8.157      Day8.157.1    100     4289.867816
3045.166614 5534.569017 13.617971     380     Day8.157.1    1608
```

AT4G20240 AT4G20240.1 j Day8.159 Day8.159.1 100 6112.378433 5517.096482 6707.660384 19.403440 1294 Day8.159.1 1608

2

(base) [root@12018efd4a72 day_16]# more Day16.transcripts.gtf.tmap | grep "AT4G20240"

0

25. **How many cufflinks transcripts are partial reconstructions of reference transcripts (‘contained’)? (Day8)**
26. **How many cufflinks transcripts are partial reconstructions of reference transcripts (‘contained’)? (Day16)**

(base) [root@12018efd4a72 day_8]# more Day8.transcripts.gtf.tmap | cut -f3 | grep -c "c"
134

-1 because header has class_code: 133

(base) [root@12018efd4a72 day_16]# more Day16.transcripts.gtf.tmap | cut -f3 | grep -c "c"
22

-1: 21

27. **How many cufflinks transcripts are novel splice variants of reference genes? (Day8)**
28. **How many cufflinks transcripts are novel splice variants of reference genes? (Day16)**

(base) [root@12018efd4a72 day_8]# more Day8.transcripts.gtf.tmap | cut -f3 | grep -c "j"
14

(base) [root@12018efd4a72 day_16]# more Day16.transcripts.gtf.tmap | cut -f3 | grep -c "j"
22

29. **How many cufflinks transcripts were formed in the introns of reference genes? (Day8)**
30. **How many cufflinks transcripts were formed in the introns of reference genes? (Day16)**

(base) [root@12018efd4a72 day_8]# more Day8.transcripts.gtf.tmap | cut -f3 | grep -c "i"
4

(base) [root@12018efd4a72 day_16]# more Day16.transcripts.gtf.tmap | cut -f3 | grep -c "i"
1

Question 31-35: Perform differential gene expression analysis with cuffmerge and cuffdiff. cuffmerge should use the provided annotation to merge and reconcile two sets of cufflinks transcripts to produce merged.gtf. cuffdiff should then perform the differential expression analysis.

GTFs.txt file:
/data/project4/cufflinks_results/day_8/transcripts.gtf
/data/project4/cufflinks_results/day_16/transcripts.gtf

```
WORKDIR=/data/project4
THDIR=/data/project4/tophat_results
ANNOT=/data/project4/athal_genes.gtf

mkdir -p $WORKDIR/cuffmerge_results
cuffmerge -g $ANNOT -p 8 -o $WORKDIR/cuffmerge_results /data/project4/GTFs.txt

mkdir -p $WORKDIR/cuffdiff_results
cuffdiff -o $WORKDIR/cuffdiff_results -p 10 $WORKDIR/cuffmerge_results/merged.gtf \
        $THDIR/day_8/accepted_hits.bam, $THDIR/day_16/accepted_hits.bam
```

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

4 Cufflinks exon 110 722 . + . gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; gene_name "AT4G19200"; oId "AT4G19200.1"; nearest_ref "AT4G19200.1"; class_code "="; tss_id "TSS1";

```
(base) [root@12018efd4a72 cuffmerge_results]# more merged.gtf | cut -f9| cut -d' ' -f2 |
sort | uniq | wc -l
129
```

32. How many transcripts?

```
(base) [root@12018efd4a72 cuffmerge_results]# more merged.gtf | cut -f9| cut -d' ' -f4 |
sort | uniq | wc -l
200
```

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

```
(base) [root@12018efd4a72 cuffdiff_results]# more gene_exp.diff | cut -f2 | sort | uniq -c
| wc -l
130
```

129 because ‘gene_id’ header was also there.

34. How many genes differentially expressed?

```
(base) [root@12018efd4a72 cuffdiff_results]# more gene_exp.diff | cut -f14 | grep -c
"yes"
4
```

35. How many transcripts differentially expressed?

```
(base) [root@12018efd4a72 cuffdiff_results]# more isoform_exp.diff | grep "yes"
TCONS_00000099  XLOC_000055     AT4G20350  4:488036-492159  q1     q2
OK     1356.25     0       -inf     nan     0.0003     0.0111    yes
TCONS_00000123  XLOC_000072     AT4G19430  4:98444-99151    q1     q2
OK     0       1934.57     inf      nan     0.0007     0.01554      yes
TCONS_00000153  XLOC_000095     AT4G19810  4:263933-265753  q1     q2
OK     5162.35     0       -inf     nan     5e-05  0.002775     yes
TCONS_00000154  XLOC_000096     AT4G19820  4:267435-268614  q1     q2
OK     976.483     0       -inf     nan     0.0007     0.01554      yes
TCONS_00000195  XLOC_000125     AT4G20240  4:431492-434212  q1     q2
OK     6465.53     0       -inf     nan     5e-05  0.002775     yes
5
```