Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

✔ Congratulations! You passed!

Next Item

✔ 1 / 1 points

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

63845

Correct Response

Step 1: Suppose we are in the directory /media/fs_gencommand_proj4/, which is where the supporting data for Exam 4 are stored on the virtual box provided for this course. Let's call this our 'base'. Create a sub-directory Tophat, and inside it two sub-directories, 'Day8' and 'Day16', where alignments for the two data sets will be stored.

```
% mkdir Tophat
% mkdir Tophat/Day8
% mkdir Tophat/Day16
```

Note: You can substitute your own directory name for the 'base', if you are using a different setup.Step 2: Create a bowtie index of the genome using bowtie2-build, with the prefix 'athal'. Include a copy of the reference genome ('athal_chr.fa') with the name 'athal.fa' in the index directory.

```
% mkdir athal
% bowtie2-build athal.fa athal/athal
% cp athal_chr.fa athal/
```

Step 3: Run tophat on each data set, using the genome index created and all default parameters, storing the output in the 'Tophat/Day{8,16}' directories:

```
% tophat -o Tophat/Day8/ athal/athal Day8.fastq
% tophat -o Tophat/Day16 athal/athal Day16.fastq
```

These will create the 'accepted_hits.bam' files containing the alignments, the 'align_summary.txt' files containing summary stats on the mapped reads, the 'unmapped.bam' files containing

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

the records of unmapped reads, as well as other derived files, in each directory.As an observation, since these are all single-end reads, no information on the insert size distribution needs to be specified. IMPORTANT NOTE: Throughout this assignment, you can also modify the scripts provided to implement the workflows.

✔ 1 / 1 points

2.

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

58398

Correct Response

Step 1: Suppose we are in the directory /media/fs_gencommand_proj4/, which is where the supporting data for Exam 4 are stored on the virtual box provided for this course. Let's call this our 'base'. Create a sub-directory Tophat, and inside it two sub-directories, 'Day8' and 'Day16', where alignments for the two data sets will be stored.

```
% mkdir Tophat
% mkdir Tophat/Day8
% mkdir Tophat/Day16
```

Note: You can substitute your own directory name for the 'base', if you are using a different setup.Step 2: Create a bowtie index of the genome using bowtie2-build, with the prefix 'athal'. Include a copy of the reference genome ('athal_chr.fa') with the name 'athal.fa' in the index directory.

```
% mkdir athal
% bowtie2-build athal.fa athal/athal
% cp athal_chr.fa athal/
```

Step 3: Run tophat on each data set, using the genome index created and all default parameters, storing the output in the 'Tophat/Day{8,16}' directories:

```
% tophat -o Tophat/Day8/ athal/athal Day8.fastq
% tophat -o Tophat/Day16 athal/athal Day16.fastq
```

These will create the 'accepted_hits.bam' files containing the alignments, the 'align_summary.txt' files containing summary stats on the mapped reads, the 'unmapped.bam' files containing the records of unmapped reads, as well as other derived files, in each directory.As an observation, since these are all single-end reads, no information on the insert size distribution needs to be

specified. IMPORTANT NOTE: Throughout this assignment, you can also modify the scripts provided to implement the workflows.

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

1 / 1 points

3.

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

63489

Correct Response

Inspect the files 'Tophat/Day*/align_summary.txt'. Look for the number of mapped reads (key word "Mapped").

1 / 1 points

4.

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

57951

Correct Response

Inspect the files 'Tophat/Day*/align_summary.txt'. Look for the number of mapped reads (key word "Mapped").

1 / 1 points

5.

How many reads were uniquely aligned in 'Day8' RNA-seq data set?

63133

Correct Response

In the 'align_summary.txt' files, subtract the number of reads reported to have multiple alignments from the number of mapped reads.

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

---

1 / 1 points

6.

How many reads were uniquely aligned in 'Day16' RNA-seq data set?

57504

Correct Response

In the 'align_summary.txt' files, subtract the number of reads reported to have multiple alignments from the number of mapped reads.

---

1 / 1 points

7.

How many spliced alignments were reported for 'Day8' RNA-seq data set?

8596

Correct Response

A spliced alignment would be marked with 'N' in the CIGAR field (column 6):

```
1  % samtools view Tophat/Day8/accepted_hits.bam | cut -f6 | grep -c 'N'
2  % samtools view Tophat/Day16/accepted_hits.bam | cut -f6 | grep -c 'N'
```

---

1 / 1 points

8.

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

10695

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

Correct Response

A spliced alignment would be marked with 'N' in the CIGAR field (column 6):

```
1  % samtools view Tophat/Day8/accepted_hits.bam | cut -f6 | grep -c 'N'
2  % samtools view Tophat/Day16/accepted_hits.bam | cut -f6 | grep -c 'N'
```

✔ 1 / 1 points

9.

How many reads were left unmapped from 'Day8' RNA-seq data set?

84

Correct Response

In the 'align_summary.txt' files, subtract the number of mapped reads from the total number of input reads.

✔ 1 / 1 points

10.

How many reads were left unmapped from 'Day16' RNA-seq data set?

34

Correct Response

In the 'align_summary.txt' files, subtract the number of mapped reads from the total number of input reads.

✔ 1 / 1 points

11.

How many genes were generated by cufflinks for Day8?

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

186

Correct Response

Step 1: Go to your 'base' directory, if not already there. Create a directory Cufflinks, with two sub-directories 'Day8' and Day16', where the assembled transcripts will be generated.

```
% mkdir Cufflinks
% mkdir Cufflinks/Day8
% mkdir Cufflinks/Day16

```

Step 2: Run cufflinks on each data set, directing the output to the directories above ('-o') and using the specified labels as prefix for naming the assembled transcripts ('-L'):

```
% cufflinks -o Cufflinks/Day8 -L Day8 Tophat/Day8/accepted_hits.bam
% cufflinks -o Cufflinks/Day16 -L Day16 Tophat/Day16/accepted_hits.bam

```

These will generate the files 'transcripts.gtf' containing the assembled transcripts, as well as files '*.fpkm_tracking' containing expression (FPKM) estimates for genes and transcripts. Visually inspect each file for consistency.Step 3: Determine the number of genes. This information is stored in column 9 of the 'transcripts.gtf' files:

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f2 | sort -u | wc -l
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f2 | sort -u | wc -l
```

1 / 1 points

12.

How many genes were generated by cufflinks for Day16?

80

Correct Response

Step 1: Go to your 'base' directory, if not already there. Create a directory Cufflinks, with two sub-directories 'Day8' and Day16', where the assembled transcripts will be generated.

```
% mkdir Cufflinks
% mkdir Cufflinks/Day8
% mkdir Cufflinks/Day16

```

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

Step 2: Run cufflinks on each data set, directing the output to the directories above ('-o') and using the specified labels as prefix for naming the assembled transcripts ('-L'):

```
% cufflinks -o Cufflinks/Day8 -L Day8 Tophat/Day8/accepted_hits.bam
% cufflinks -o Cufflinks/Day16 -L Day16 Tophat/Day16/accepted_hits.bam
```

These will generate the files 'transcripts.gtf' containing the assembled transcripts, as well as files '*.fpkm_tracking' containing expression (FPKM) estimates for genes and transcripts. Visually inspect each file for consistency.Step 3: Determine the number of genes. This information is stored in column 9 of the 'transcripts.gtf' files:

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f2 | sort -u | wc -l
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f2 | sort -u | wc -l
```

✔ 1 / 1 points

13.

How many transcripts were reported for Day8?

192

Correct Response

This information is stored in column 9 of the 'transcripts.gtf' files:

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f4 | sort -u | wc -l
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f4 | sort -u | wc -l
```

✔ 1 / 1 points

14.

How many transcripts were reported for Day16?

92

Correct Response

This information is stored in column 9 of the 'transcripts.gtf' files:

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f4 | sort -u | wc -l
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f4 | sort -u | wc -l
```

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

1 / 1 points

15.

How many single transcript genes were produced for Day8?

180

Correct Response

The rationale is similar to that employed in modules 1 and 2. Each gene can have one or more transcripts. We first create a listing of (gene,transcript) pairs and use it to determine the number of transcripts for each gene. For instance, a gene with a single transcript would appear only 1 time in column 1, a gene with 2 transcripts 2 times, etc. We calculate the number of occurrences in column 1 for each gene using 'uniq', and then select those that have count 1.

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f2,4 | sort -u | cut -d '
    ' -f1 | sort | uniq -c | grep -c " 1 "
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f2,4 | sort -u | cut -d
    ' ' -f1 | sort | uniq -c | grep -c " 1 "
```

1 / 1 points

16.

How many single transcript genes were produced for Day16?

69

Correct Response

The rationale is similar to that employed in modules 1 and 2. Each gene can have one or more transcripts. We first create a listing of (gene,transcript) pairs and use it to determine the number of transcripts for each gene. For instance, a gene with a single transcript would appear only 1 time in column 1, a gene with 2 transcripts 2 times, etc. We calculate the number of occurrences in column 1 for each gene using 'uniq', and then select those that have count 1.

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f2,4 | sort -u | cut -d '
    ' -f1 | sort | uniq -c | grep -c " 1 "
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f2,4 | sort -u | cut -d
    ' ' -f1 | sort | uniq -c | grep -c " 1 "
```

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

1 / 1 points

17.

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

119

Correct Response

Each transcript is represented in the GTF file with one 'transcript' (column 3) line and one or several 'exon' lines. Therefore, a single-exon transcript would appear listed on exactly 2 lines. We determine the number of lines in the GTF file for each transcript, then select those that have exactly 2 lines:

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f4 | sort | uniq -c |
    grep -c " 2 "
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f4 | sort | uniq -c |
    grep -c " 2 "
```

1 / 1 points

18.

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

24

Correct Response

Each transcript is represented in the GTF file with one 'transcript' (column 3) line and one or several 'exon' lines. Therefore, a single-exon transcript would appear listed on exactly 2 lines. We determine the number of lines in the GTF file for each transcript, then select those that have exactly 2 lines:

```
% cut -f9 Cufflinks/Day8/transcripts.gtf | cut -d ' ' -f4 | sort | uniq -c |
    grep -c " 2 "
% cut -f9 Cufflinks/Day16/transcripts.gtf | cut -d ' ' -f4 | sort | uniq -c |
    grep -c " 2 "
```

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

1 / 1 points

19.

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

73

Correct Response

Simply subtract the number of single-exon transcripts (Q9) from the total number of transcripts (Q7).

1 / 1 points

20.

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

68

Correct Response

Simply subtract the number of single-exon transcripts (Q9) from the total number of transcripts (Q7).

1 / 1 points

21.

How many cufflinks transcripts fully reconstruct annotation transcripts in Day8?

16

Correct Response

Step 1: Cuffcompare is the simplest in the cufflinks suite and was not demonstrated in the lectures. It compares the assembled transcripts against a set of reference gene annotations provided by the user, exon-by-exon, to determine which genes and transcripts in the sample are known, and which ones are likely novel. In the end, it assigns each predicted (cufflinks) transcript a 'class' code depending on how it relates to a reference transcript,

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

for example: it is the same as a reference transcript (‘=’), it is only a portion of one (‘c’), a new splice variant of a reference gene (‘j’), etc. To find the appropriate command line parameters for cuffcompare, run the command without parameters to obtain information about its usage, and save this to a file for easy inspection:

```
% cuffcompare >& cuffcompare.log

```

Note that you can use this any time you need to test new software.Step 2: Run cuffcompare against the provided annotation (‘-r’) and with the option ‘-R’ to exclude from statistics genes that do not appear to be represented in the sample:

```
% cd Cufflinks/Day8
% cuffcompare -r ../../athal_genes.gtf -R transcripts.gtf
% cd ../../Cufflinks/Day16
% cuffcompare -r ../../athal_genes.gtf -R transcripts.gtf
% cd ../../

```

These will create the files ‘cuffcmp.combined.gtf’ combining the reference and predicted annotations, ‘cuffcmp.transcripts.gtf.tmap’ containin a mapping between the assembled transcripts and the reference genes and transcripts, as well as other derived files, in the corresponding directories.Step 3: The ‘*.tmap’ files contain a line for each cufflinks transcript, and indicate the ‘class’ code in column 3.

```
% cut -f3 Cufflinks/Day8/cuffcmp.transcripts.gtf.tmap | sort | uniq -c
% cut -f3 Cufflinks/Day16/cuffcmp.transcripts.gtf.tmap | sort | uniq -c

```

These commands will list the number of transcripts for each class, which you will use to answer this and the following questions. The class code for transcripts that fully reconstruct reference transcripts is ‘=’, so look for the count corresponding to this class code.

✔ 1 / 1 points

22.

How many cufflinks transcripts fully reconstruct annotation transcripts in Day16?

36

Correct Response

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

Step 1: Cuffcompare is the simplest in the cufflinks suite and was not demonstrated in the lectures. It compares the assembled transcripts against a set of reference gene annotations provided by the user, exon-by-exon, to determine which genes and transcripts in the sample are known, and which ones are likely novel. In the end, it assigns each predicted (cufflinks) transcript a 'class' code depending on how it relates to a reference transcript, for example: it is the same as a reference transcript ('='), it is only a portion of one ('c'), a new splice variant of a reference gene ('j'), etc. To find the appropriate command line parameters for cuffcompare, run the command without parameters to obtain information about its usage, and save this to a file for easy inspection:

```
% cuffcompare >& cuffcompare.log
```

Note that you can use this any time you need to test new software.Step 2: Run cuffcompare against the provided annotation ('-r') and with the option '-R' to exclude from statistics genes that do not appear to be represented in the sample:

```
% cd Cufflinks/Day8
% cuffcompare -r ../../athal_genes.gtf -R transcripts.gtf
% cd ../../Cufflinks/Day16
% cuffcompare -r ../../athal_genes.gtf -R transcripts.gtf
% cd ../../
```

These will create the files 'cuffcmp.combined.gtf' combining the reference and predicted annotations, 'cuffcmp.transcripts.gtf.tmap' containin a mapping between the assembled transcripts and the reference genes and transcripts, as well as other derived files, in the corresponding directories.Step 3: The '*.tmap' files contain a line for each cufflinks transcript, and indicate the 'class' code in column 3.

```
% cut -f3 Cufflinks/Day8/cuffcmp.transcripts.gtf.tmap | sort | uniq -c
% cut -f3 Cufflinks/Day16/cuffcmp.transcripts.gtf.tmap | sort | uniq -c
```

These commands will list the number of transcripts for each class, which you will use to answer this and the following questions. The class code for transcripts that fully reconstruct reference transcripts is '=', so look for the count corresponding to this class code.

✔ 1 / 1 points

23.

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

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

2

Correct Response

Search for the gene name in the '*.tmap' file, which contains a listing of all cufflinks transcripts and their relationship to reference genes:

```
% grep AT4G20240 Cufflinks/Day8/cuffcmp.transcripts.gtf.tmap
% grep AT4G20240 Cufflinks/Day16/cuffcmp.transcripts.gtf.tmap
```

1 / 1 points

24.

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

0

Correct Response

Search for the gene name in the '*.tmap' file, which contains a listing of all cufflinks transcripts and their relationship to reference genes:

```
% grep AT4G20240 Cufflinks/Day8/cuffcmp.transcripts.gtf.tmap
% grep AT4G20240 Cufflinks/Day16/cuffcmp.transcripts.gtf.tmap
```

1 / 1 points

25.

How many cufflinks transcripts are partial reconstructions of reference transcripts ('contained')? (Day8)

133

Correct Response

The corresponding class code is 'c' (contained). Using the counts calculated at Q11-Step3, look for the count for class code 'c'.

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

1 / 1 points

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

21

Correct Response
The corresponding class code is 'c' (contained). Using the counts calculated at Q11-Step3, look for the count for class code 'c'.

1 / 1 points

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

14

Correct Response
The corresponding class code is 'j' (novel 'junction'). Using the counts calculated at Q11-Step3, look for the count for class code 'j'.

1 / 1 points

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

22

Correct Response

The corresponding class code is ‘j’ (novel ‘junction’). Using the counts calculated at Q11-Step3, look for the count for class code ‘j’.

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions

1 / 1 points

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

4

Correct Response
The corresponding class code is ‘i’ (intronic). Using the counts calculated at Q11-Step3, look for the count for class code ‘i’.

1 / 1 points

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

1

Correct Response
The corresponding class code is ‘i’ (intronic). Using the counts calculated at Q11-Step3, look for the count for class code ‘i’.

1 / 1 points

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

129

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

Correct Response

Step 1: Generate the file 'GTFs.txt' containing the list of GTF files to be merged (/media/sf_gencommand_proj4/Cufflinks/Day*/transcripts.gtf), one per line, with the full paths. Then run cuffmerge with the reference annotation ('-g'):

```
% cuffmerge -g athal_genes.gtf GTFs.txt
```

This will create a directory 'merged_asm' containing the file 'merged.gtf'.Step 2: Use a workflow similar to that in Q6-Step 3 to determine the number of genes:

```
% cut -f9 merged_asm/merged.gtf | cut -d ' ' -f2 | sort -u | wc -l
```

---

1 / 1 points

32.

How many transcripts?

200

Correct Response

Use a workflow similar to that in Q7 to determine the number of transcripts:

```
% cut -f9 merged_asm/merged.gtf | cut -d ' ' -f4 | sort -u | wc -l
```

---

1 / 1 points

33.

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

129

Correct Response

Module 4 Exam

Quiz, 35 questions

35/35 points (100%)

Step 1: Create a directory 'Cuffdiff' in the base directory. Then run cuffdiff with the 'merged.gtf' file as reference annotation, taking as input the two alignment files and directing the output to the directory Cuffdiff ('-o'):

```
% mkdir Cuffdiff
% cuffdiff -o Cuffdiff merged_asm/merged.gtf Tophat/Day8/accepted_hits.bam
    Tophat/Day16/accepted_hits.bam

```

This will create the file 'gene_exp.diff' containing test scores and results for the gene-level differential expression analysis, other '*.diff' files, as well as tracking files for genes, transcripts, splicing, CDS, TSS, etc.Step 2: To answer the question, simply count the number of lines in the 'gene_exp.diff' file and subtract 1 (the header):

```
% wc -l Cuffdiff/gene_exp.diff
```

✔ 1 / 1 points

34.

How many genes were detected as differentially expressed?

4

Correct Response

These lines are marked with 'yes' in column 13 ('significant') in the 'gene_exp.diff' file:

```
% grep -c yes Cuffdiff/gene_exp.diff
```

✔ 1 / 1 points

35.

How many transcripts were differentially expressed between the two samples?

5

Correct Response

Similar to Q19, only for the 'isoform_exp.diff' file:

```
% grep -c yes Cuffdiff/isoform_exp.diff
```

Module 4 Exam

35/35 points (100%)

Quiz, 35 questions