Assignment 11

How to complete this assignment

- Read through the background information
- Go through each question in order and complete any tasks that are described in the question. **Note that some information in this file may not be available in the Quercus quiz.**
- As you complete the questions, mark your answer to each question.
- Questions will be either:
  - multiple-choice questions that require you to provide either a single answer or to select multiple answers.
  - questions that require a short text answer
- Open the associated assignment quiz on Quercus and enter your answers to each question.
- You may only submit this quiz once, so be sure you answer all questions before submitting the quiz.

Before you begin

- Open a new terminal session from your JupyterHub (New > Terminal)
- Set the PWD to `/home/jovyan/Week.11/Assignment.11`

Mark breakdown

Part 1 – 4 questions – 4 marks
Part 2 – 8 questions – 8 marks
Part 3 – 4 questions – 4 marks
Part 4 – 6 questions – 6 marks

BACKGROUND

For this assignment, you will be using the same RNA-seq data that you have used for the week 11 lectures and tutorial. It is associated with the following paper:

> Wei, G., Luo, H., Sun, Y., Li, J. et al. Transcriptome profiling of esophageal squamous cell carcinoma reveals a long noncoding RNA acting as a tumor suppressor. (2015) *Oncotarget* **6(19)**: 17065-80

In this paper, the authors looked at tumour samples (specifically esophageal squamous cell carcinoma) from three patients and compared these samples to adjacent healthy tissue from each patient. RNA-sequencing was performed on each sample, and the raw reads are stored in 6 different FASTQ files, one per sample. Upon analyzing these data, the authors found that PTK6 acted as a tumour suppressor.

File	**Description**
`genome_annotation.gtf`	GTF file contain features in the human genome
`16T_hisat2.sam`	SAM file containing the results of aligning reads from sample 16T to the human genome
`countMatrix.tsv`	Count matrix containing read counts for genes across samples 16N, 18N, 19N, 16T, 18T, and 19T

Data Sources

The `genome_annotation.gtf` file was downloaded from RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.26_GRCh38/). The `16T_hisat2.sam` file was generated from FASTQ files from Wei, G., Luo, H., Sun, Y., Li, J. et al. The `countMatrix.tsv` was generated as described in lecture from SAM files for samples 16N, 18N, 19N, 16T, 18T, and 19T.

PART 1: The PTK6 Gene (4 marks)

Use the file `genome_annotation.gtf` to answer the following questions.

Question 1

(1 mark)

How many lines of the GTF file are associated with the gene “PTK6”?

a. 38
b. 45
c. 24
d. 62

Question 2

(1 mark)

What is the name of the contig that contains the gene “PTK6”?

a. chr20
b. NW_018654715.1
c. chr18
d. NT_187374.1

Question 3

(1 mark)

How many exons are associated with the gene “PTK6”?

a. 17
b. 4
c. 21
d. 12

Question 4

(1 mark)

What are the base pair coordinates of the gene “PTK6” on its respective contig?

a. 63537085 - 63537376
b. 63528001 - 63537376
c. 63537085 - 63537314
d. 63530921 - 63530923

PART 2: The RNA-seq Alignment Process (8 marks)

The following questions address various steps in the RNA-seq alignment process that was covered this week.

Question 5

(1 mark)
Why is cDNA sequenced instead of RNA?

a. It is the standard to record thymine instead of uracil in any sequence
b. It makes the experimental protocol cheaper
c. More accurate read counts are collected
d. RNA is not stable enough to be sequenced

Question 6

(1 mark)
What modification may be done to RNA-seq data before it is aligned to the genome?

a. If the sequences are paired, the second mate may be removed
b. Adaptors may be removed if they are present
c. Reads may be sorted by read length
d. Low quality bases may be removed from the ends of reads
e. A and B
f. B and D
g. C and D

Question 7

(1 mark)
What does an “N” in the reference genome FASTA file mean?

a. It represents the base “N”
b. Missing information - can be any base
c. It is a gap in the sequence
d. Either an A or T

Question 8

(1 mark)
Which of the following steps in the alignment happens first?

a. The number of counts associated with each gene are calculated
b. The reference genome is indexed
c. Reads are aligned to the reference genome
d. FASTQ files are sorted by transcript length

Question 9
(1 mark)
Which of the following is NOT a difference between FASTQ and FASTA files?

a. FASTQ files are larger than FASTA files
b. FASTQ files store base quality information
c. Each sequence in a FASTQ file uses 3 lines, whereas in a FASTA file it uses 2 lines
d. FASTA files can also store reference genome sequences

Question 10
(1 mark)
What can help reduce the problem of batch effects?

a. Adding replicates to the experiment
b. Reducing the number of technical zeros with additional software
c. Adding more experimental groups to the analysis
d. Filtering out low-quality reads more stringently

Question 11
(1 mark)
Which of the following files is NOT used when aligning reads to the genome with HISAT2?

a. The reference genome
b. The reference genome index
c. The FASTQ file(s) you are aligning
d. The name of the SAM alignment file

Question 12
(1 mark)
Which of the following is NOT true about the output of HISAT2?

a. The output from HISAT2 can be directly converted to a count matrix by FeatureCounts
b. If an output is not specified for HISAT2, the results will print to the screen
c. The output of HISAT2 is an alignment file that specifies where each read has aligned to the genome
d. If paired FASTQ files are used, the SAM file will store the gaps between each read

PART 3: Interpreting SAM Files (4 marks)

Look at the following line from the file `16T_hisat2.sam`:

```
16T_RNA.sra.106       256     chr14_KI270846v1_alt    1128362 1
38M     *       0       0
TACCACCCTGAACGCGCCCGATCTCGTCTGATCTCGGA
CCCCCCCCABCCCCCCCCCCC@BCCC@>B@BCCC@CBCC  AS:i:0  ZS:i:0  XN:i:0
XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:38 YT:Z:UU NH:i:5
```

Question 13
(1 mark)
What is the name of the contig that this read has aligned to?

a. chr3
b. chr14_KI270846v1_alt
c. chr14
d. 256

Question 14
(1 mark)
What is the starting base pair coordinate of the read on this contig?

a. 1
b. 256
c. 0
d. 1128362

Question 15
(1 mark)
Looking at `genome_annotation.gtf`, what gene does this read map to?

a. ATP11A
b. IGHVII-60-1
c. RNA5SP389
d. IL1B
e. SNX30

Question 16
(1 mark)
After which base pair is there a mismatch?

a. 0
b. 5
c. 38
d. 257
e. 1128362

PART 4: Count Matrices (6 marks)

Use the file `countMatrix.tsv` to answer the following questions.

Question 17
(1 mark)
Which of the following samples is not a tumour sample?

a. 19N
b. 16T
c. 18T
d. 19T

Question 18
(1 mark)
How many genes are detected in the count matrix?

a. 43,388
b. 43,389
c. 43,390
d. 4,339

Question 19
(1 mark)
What is a technical zero?

a. A zero in the count matrix indicating that a gene was not transcribed
b. A zero that has replaced a positive number in the count matrix due to errors in the alignment
c. A zero in the count matrix even though the gene was transcribed
d. An indicator in the count matrix that the expression of that gene should not be trusted

Question 20
(1 mark)
What does each SAM file become in the count matrix?

a. A row
b. A column
c. An expression value
d. A set of columns

Question 21
(1 mark)
What is the sum of counts detected for PTK6 in tumour and normal samples, respectively?

a. 706 and 7006
b. 7006 and 706
c. 654 and 8921
d. 8291 and 654

Question 22

(1 mark)

Given the answer to the last question, which of the following statements is most correct?

a. PTK6 is overexpressed in normal samples
b. PTK6 is underexpressed in normal samples
c. PTK6 is overexpressed in tumour samples
d. PTK6 is underexpressed in tumour samples