11 RNA-sequencing & Alignment – Tutorial

At the end of this tutorial you should be able to:

- Describe the steps required for RNA-sequencing alignment
- Create command line arguments for tools that are used to process RNA-seq data
- Navigate different file types, including: FASTQ, GTF, and SAM

How to complete this tutorial

- Go through each question in order and complete any tasks that are described in the question.
- 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 quiz on Quercus and enter your answers to each question to verify that you completed the tutorial questions correctly.
- Alternatively, open the Quercus quiz when you start the tutorial and verify your answers as you complete the tutorial. **Note that there may be some information that is in this file that is not in the Quercus quiz!**
- The answers will be released at the end of the week.

Before you begin

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

Background

You may need to refer to the HISAT2 manual (http://daehwankimlab.github.io/hisat2/manual/) to answer some of these questions. For the purpose of this tutorial, assume we are working with the following input files (you don't necessarily have all these files, but these will be the names you will use when designing commands):

1. The reference genome FASTA file: `genome.fna`
2. The reference genome GTF file: `genome_annotation.gtf`
3. FASTQ files of RNA-seq reads: `16N_RNA.fastq, 18N_RNA.fastq, 19N_RNA.fastq, 16T_RNA.fastq, 18T_RNA.fastq, 19T_RNA.fastq`

You will be exploring some of these files to answer questions about their content, as well as use the names of these files to create commands that you would use to align the RNA-seq data to the reference genome.

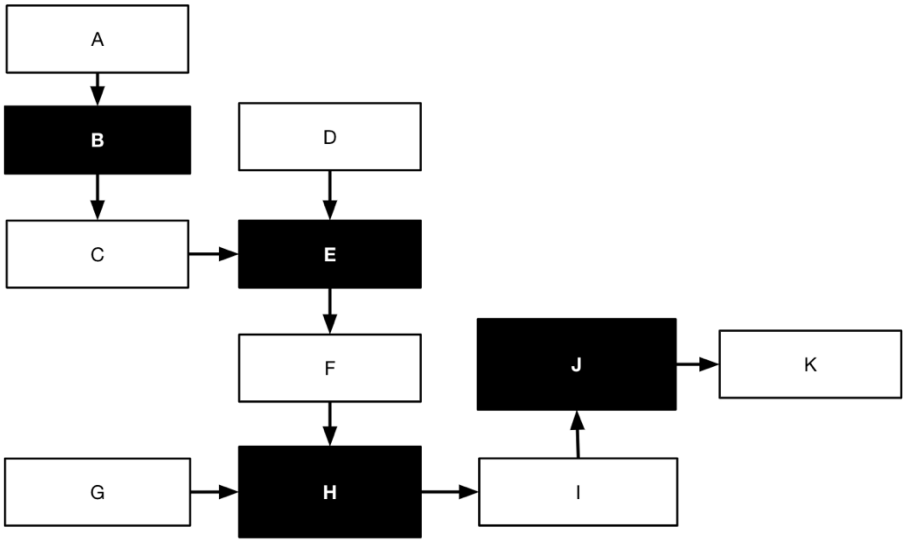
The `genome.fna` & `genome_annotation.gtf` file were downloaded from RefSeq (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.26_GRCh38/). FASTQ files are from:

> 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

11.1.1 Overview of the RNA-sequencing Analysis Pipeline

Question 1

First, let's take a step back and make sure we understand the RNA-sequencing pipeline. The following diagram is a flow chart of the general steps involved in RNA-seq alignment that takes you from the raw FASTQ files to the count matrix. Match the correct steps to the letters.



Letter	Step
	Reference genome GTF file
	Isolate specific columns for count matrix using "cut"
	Reference genome FASTA
	Return gene counts with FeatureCounts
	Alignment with HISAT2
	Intermediate count matrix
	Reference genome index
	Index genome with hisat2-build
	Final count matrix
	Raw FASTQ reads
	SAM file of aligned reads

Question 2

Which of the following would NOT potentially change the results of the RNA-seq alignment pipeline?

a. Using different technology to perform the RNA-seq experiment
b. Using a different alignment tool
c. Using an updated version of the reference GTF or genome FASTA files
d. Running the alignment pipeline on a different computer

Question 3

At what resolution is RNA-seq data?

a. Intracellular level
b. Cellular level
c. Tissue level
d. Organ level

11.1.2 Raw RNA-sequencing Reads & 11.1.3 FASTQ Files

Question 4

Now let's come back to the RNA-seq reads which are stored as FASTQ files.
Of the following symbols, which would represent the highest quality base in a FASTQ file?

a. H
b. #
c. 2
d. @

Question 5

How long are the reads in the FASTQ file `16T_RNA_100_lines.fastq`? You can do this by isolating a line of the file that is a raw base pair sequence (e.g. `sed -n '1p'` would extract the first line of a file) and counting the number of characters (e.g. with the command `wc`).

a. 37
b. 38
c. 39
d. 40

11.1.4 Genome FASTA Files and GTF Files

Question 6

Now let's take a look at the files associated with the reference genome we will be using. These files include the genome FASTA file (`genome.fna`) and the GTF file (`genome_annotation.gtf`).
What is the difference between a chromosome and a contig?

a. A contig is the computational term for a chromosome
b. A biological genome is divided into chromosomes while the genome sequence in the FASTA file is divided into contigs
c. Contigs are shorter than chromosomes
d. Contigs are additional segments of the genome that aren't located on a chromosome

Question 7

Use `grep -c -P "\tgene\t"` on the reference genome GTF file. This looks for the word "gene" that is surrounded by tabs, which would only occur for lines that are describing a gene. The `-c` argument counts the number of lines that are returned, rather than returning the lines, themselves. How many genes are there?

a. 28,943
b. 35,087
c. 24,066
d. 59,282

Question 8

What are the different features that are present in the GTF file? You can figure this out by using `cut` to extract the "feature type" column of the GTF file, then `sort` and `uniq` to get a list of the unique features.

a. gene, exon, CDS, start_codon, stop_codon
b. gene, mRNA, exon, CDS
c. gene, mRNA, start_codon, stop_codon
d. gene, mRNA, exon, CDS, start_codon, stop_codon

Question 9

Which of the following features would NOT be found in ANY GTF file? (I.e., not a real feature type).

a. mRNA
b. base
c. start_codon
d. CDS

11.2.1 How Alignment Works & 11.2.2 Running the Alignment

We will now go over how to use the files we just explored to (1) index the reference genome and (2) align the RNA-seq reads to the reference genome. Let's start by indexing the reference genome, as that must be done first.

Question 10

Write the full command using `hisat-build` to create an index with the base name of `genome_index`.

```
hisat2-build genome.fna genome_index
```

How many files are outputted by the indexing process?

a. 1
b. 2
c. 4
d. 8

Question 11
Why does HISAT2 use a reference genome index instead of the reference genome FASTA file for the alignment?

a. The index helps the alignment occur more accurately
b. The index helps the alignment go faster
c. The reference genome FASTA file is too big to be processed by HISAT2
d. The reference genome FASTA contains redundant information

Question 12
Next, we use HISAT2 to align the RNA-seq reads to the reference genome. Do we do this for all of the files at once, or for one file at a time?

a. All files at once
b. One file at a time

Question 13
What is the output of the HISAT2 alignment?

a. A BAM file
b. A SAM file
c. A FASTA file
d. A GTF file

Question 14
What will happen if we don't specify the name of the SAM file to use as output when aligning the reads with HISAT2? You may need to look at the HISAT2 manual to answer this question.

a. A SAM file with a default name will be created.
b. The output will be printed to the screen.
c. An error will be triggered.
d. A BAM file will be output instead.

Question 15
How do we correctly specify the indexed reference genome in the HISAT2 command? Hint: look at the <ht2-index> description.

a. `-x genome_index`
b. `-x genome_index.X.ht2`
c. `-x genome_index.1.ht2 -x genome_index.2.ht2 -x genome_index.3.ht2 -x genome_index.4.ht2 -x genome_index.5.ht2 -x genome_index.6.ht2 -x genome_index.7.ht2 -x genome_index.8.ht2`
d. `-x genome_index.1.ht2 genome_index.2.ht2 genome_index.3.ht2 genome_index.4.ht2 genome_index.5.ht2 genome_index.6.ht2 genome_index.7.ht2 genome_index.8.ht2`

Question 16

What should our FASTQ file argument look like when aligning the FASTQ file `16T_RNA.fastq`?

a. `-1 16T_RNA.fastq`
b. `-2 16T_RNA.fastq`
c. `-U 16T_RNA.fastq`
d. `-1 16T_RNA.fastq -2 16T_RNA.fastq`

Question 17

Using the pieces of information from the preceding questions, write out the HISAT2 command used to align the FASTQ file `16T_RNA.fastq` to the reference genome. The output file for this command should be called `16T_hisat2.sam`

Question 18

After running the alignment, it is possible that some of the RNA-seq reads may not be mapped to the reference genome. Which of the following is NOT a potential cause of unmapped reads?

a. Contamination of the sample
b. Some of the transcripts in the sample are from viruses or bacteria
c. The aligner ran out of memory in the middle of the job
d. Some of the transcripts are from regions of the genome that haven't yet been sequenced or annotated

---

11.2.3 BAM/SAM Files

Question 19

After running the alignment with HISAT2, we are left with a SAM file associated with each of our samples (e.g. `16T.sam`).

What is the main benefit of a BAM file over a SAM file?

a. BAM files are human-readable
b. BAM files are more compressed so they take up less space
c. BAM files are always easier to generate than SAM files
d. BAM files are used more universally than SAM files

Question 20

The following is a line from the SAM file 16T.sam:

```
16T_RNA.12        272     chr6    4428035 1       38M     *
0       0       TAGTACTTGGATGGGAGACCGCCTGGGAATACCGGGTG
?;7?7>A>=?=0ACABB?CCCCCCCCCCCCCCCCCCCC  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
```

What is the name of the contig that this read aligned to?

a. 272
b. chr6
c. 38M
d. 16T_RNA.12

Question 21

What does the number 4428035 indicate?

a. The mapping quality of the read
b. The number of reads that mapped to the same position as this read
c. The coordinate of the contig that the first base mapped to
d. The length of the contig that the read mapped to

11.3.1 Count Matrices & 11.3.2 Converting Alignment Files to Counts

---

Question 22

After running the HISAT2 alignment on all 6 FASTQ files, `FeatureCounts` is used to convert the 6 SAM files to the intermediate count matrix file. `FeatureCounts` also requires the GTF file of the reference genome to connect the contig locations listed in the SAM file to the associated gene names written in the GTF file.

Write the command you would use to run `FeatureCounts`, naming the output file `countMatrixIntermediate.tsv`.

a. `featureCounts -a genome_annotation.gtf -o countMatrixIntermediate 16T.sam`
b. `featureCounts -a genome_annotation.gtf 16N.sam 18N.sam 19N.sam 16T.sam 18T.sam 19T.sam`
c. `featureCounts -d 38 -o countMatrixIntermediate.tsv 16N.sam 18N.sam 19N.sam 16T.sam 18T.sam 19T.sam`
d. `featureCounts -d 38 -a genome_annotation.gtf -o countMatrixIntermediate.tsv 16N.sam 18N.sam 19N.sam 16T.sam 18T.sam 19T.sam`

Question 23

This creates an intermediate file that still contains unnecessary information; we are only interested in creating a count matrix populated with the raw counts of every gene for every sample. Instead, we currently have a TSV file that has extra information in columns 2-6 of a 12-column matrix. Which command would isolate only the information we want from this file?

a. `cut -f7-12 countMatrixIntermediate.tsv > countMatrix.tsv`
b. `grep "16N|18N|19N|16T|18T|19T" countMatrixIntermediate.tsv > countMatrix.tsv`
c. `cut -f1,7-12 countMatrixIntermediate.tsv > countMatrix.tsv`
d. `cut -f2-6 countMatrixIntermediate.tsv > countMatrix.tsv`