BENG 183 Applied Genomic Technologies (Fall 2019)
Due: November 5, 2019 by 11:59pm

Homework 3

1. **RNA-seq Analysis**
    a. Why is normalization important in RNAseq analysis?

    b. One method to normalize RNAseq results is RPKM (Reads Per Kilobase per Million mapped reads). Say we are looking at a gene sequence that is 400bp long. After performing alignment on a library of 10^8 total single-end reads, we find that 500 reads have mapped to the gene of interest. Calculate the RPKM for this case.

    c. A more popular normalization method nowadays is TPM (Transcripts Per Million). How is TPM calculated? Why is TPM preferred over RPKM?

2. **Online Tutorial 3: Mapping and Quantification**

Please go through Tutorial 3 to learn how to map RNAseq data to a reference genome and quantify each gene's expression level, and complete the following tasks:

a. After mapping the reads to the genome, look at the contents of one of the **Log.final.out* files. (You can use the `cat` function to print the file contents in the terminal.) Notice that it lists the statistics for 4 types of reads: unique, multi-mapping, unmapped, and chimeric.
    i. How does your percentage of unmapped reads compare to the percentage of mapped reads?
    ii. What are some causes for unmapped reads? (Give at least two reasons.)
    iii. What is the meaning of a "multi-mapping read" and why does it occur?

b. Next, visualize your **BAM files** using the IGV software. To do this, first create an index (.bai file) for each of your BAM files with the following command:

```
samtools index my.sorted.bam
```

Then, go to https://igv.org/app/. Under the "Genome" tab on the top left, choose "D. melanogaster (dm6)". Under the "Tracks" tab, click on "Local Files" and upload your BAM files and .bai index files (you should have 8 files total). Zoom in on the tracks until you can see the coverage data. The track at the very bottom (labeled "Refseq Genes") shows the genes on the reference genome. You can drag left/right to move to different positions along the genome.
    i. Find the "CG3106" gene using the search box near the top (next to the magnifying glass). Attach screenshots of your coverage data results at the location of this gene.
    ii. You should see a noticeable difference in coverage between the head samples and the midgut samples. What could these differences suggest? Can you make any definitive conclusions about gene expression just by comparing these coverage plots? Why or why not?

c. After quantifying gene expression level using FeatureCount, you will end up with a **count.txt.summary* file for each of the four samples (head1, head2, midgut1, midgut2). Use `cat` to print out the content of each summary file and attach screenshots of the results.

d. Kallisto uses an alignment-free method to map reads. Explain, in general, how an alignment-free method works. What are some advantages of alignment-free methods?