CSCI 291T (Spring 2016) Programming assignment 7 (30 pts) Due: 05/03 (T)

RNA-Seq data analysis pipeline should include at least the following three processes:

1. exon region mask complement;
2. junction region preparation;
3. mapping on exon and junction regions, and gene-level read counting.

In this project, we manipulate process1 and process3, i.e., processing without considering junction region mappings. In practice, junction mapping covers about 6% of total mappings.

Task 1: exon region mask complement

As the first step of building RNA-Seq data analysis pipeline, we need to prepare collapsed exon regions. We limit our experiment with only HG19 chr1.

1. Prepare HG19 chr1 exon annotation:
   Download HG19 RefSeq exon annotation and trim it for chr1 only with 6-columns each.
   We already have it from previous activities.

2. Make a collapsed exon annotation:
   From the resulting annotation from Step1, collapse all overlapped regions and make a collapsed annotation with each collapsed exon name “x” and strand “+”.
   For example,

   ```
   chr1   6469   6628   NR_024540_exon_3_0_chr1_6470_r   0   -
   chr1   6469   6631   NR_028269_exon_3_0_chr1_6470_r   0   -
   ```

   becomes

   ```
   chr1   6469   6631   x   0   +
   ```

   You need to write a program for this.

3. Mask-complement the genome:
   Using the collapsed exon annotation (from Step2), mask all non-exon regions of HG19 chr1 with ‘N’s.
   You need to write a program for this.

Task 2: read mapping and read count

1. Using Bowtie, map a reads file (fastq format) onto the genome.
   The input reads file will be provided from instructor.

2. Using the original exon annotation file, count mapped reads on each exon.
   You need to write a program for this.

3. Convert the exon-level read counting to gene-level counting, i.e., gene expression level.
   You need to write a program for this, or you can include this operation in the step2.

Submission:

1. Include good documentations in each source code and submit the hard copy of the source codes. Documentation should include a global documentation (at least, program description – what it does, input/output description, methods/algorithms used, how to compile and run) and each function head documentation (what the function does, methods/algorithms used, input/output).

2. Make a zip file containing the following two files and send it to jpark@csufresno.edu.
   - collapsed exon annotation file (HG19 chr1 only);
   - a table showing the gene expression levels – 2 columns (geneID, read_count).