Assignment #7 (Total: 150 points) Revision 8:19pm, Dec 6th, 2017

Due: 4:00pm, Monday, Dec 11th, 2017

1. Download the two fastq files (SRR2567795 and SRR2567786) of 12 runs for Bioproject (Accession#: PRJNA297893). Use FASTQC to check the read quality only for the SRR2567795 fastq file(s – if you downloaded paired-end files that I recommend).
    a. Copy and paste "Per base sequence quality" from FASTQC output below.
    b. What does the "yellow box" in the "Per base sequence quality" represent?

2. To conduct reference-based RNA-seq analysis using the sequencing read data you downloaded, you will need their corresponding genome (Schizosaccharomyces pombe) sequences and genome annotation file (gtf or gff3). Identify a source to obtain these files and download its reference genome sequence (fasta) and genome annotation file. Download the genome and gene annotations from PomBase (https://www.pombase.org/) database website. Useful link is ftp://ftp.pombase.org/pombe/genome_sequence_and_features/fasta/ . Be careful that our system cannot access the FTP site. So, you can copy the files from my shared directory at hopper.slu.edu: /public/ahnt/courses/bcb5200/HW7
    a. Search and provide the citation of the paper for the assembly.

3. Use Tophat2 to map the each data set to the reference genome, guided by gff3. Provide tophat command for SRR2567795 you used to map the reads to genome. I recommend you to use few multi-threads option.
   ** Please note that the values in the first column of the provided GFF file (column which indicates the chromosome or contig on which the feature is located), must match the name of the reference sequence in the Bowtie index you are using with TopHat. You can get a list of the sequence names in a Bowtie index by typing:
   $ bowtie-inspect --names your_index
   So, if the sequence names are different between gff3 and fasta files, then change the fasta file IDs to match.

4. Based on align_summary.txt in your tophat output directory, report overall read mapping rates.

5. Use cufflinks to assemble transcriptomes for the 2 alignment files generated by Tophat. Based on your cufflinks output files, please report how many genes (loci) and transcripts are assembled by cufflinks from each RNA-library.

6. Use cuffmerge to merge the 2 transcripts generated by cufflinks into a single merged transcript annotation file. Based on your cuffmerge output file, please find out the total number of transcripts are produced.

7. Use cuffcompare to compare your merged annotation file with reference annotation. Report the stat output and find the total number of novel transcripts that have novel isoform. You should shortly describe how you find this.

8. Now, you want to run the edgeR to report differential expression. How to convert your mapped results to counts? You need to run the ht-seq count for it. Check the page: http://htseq.readthedocs.io/en/master/count.html . (Somehow, I am in trouble to install the HTseq in hopper with local (user) option. So, I installed it into my workstation (biohpc01.slu.edu). You can just hop to the biohpc01 from hopper with the command $ssh biohpc01. Then, you will see your home directory that is exactly same as the hopper by network shared drive)
   Finally, report the total number of differentially expressed genes at FDR< 0:05 and report plot from function plotSmear like the mini-homework.

9. Optional (+10 points). Optionally, you can report the Gene Ontology of differentially expressed genes and do enrichment analysis using any tool.