Quantitative Molecular Neurogenomics
(ME.440.825)
Johns Hopkins University

Module 6 (10/10/22) Problem set
Instructor: Taeyoung Hwang, PhD

1. Alignment without trimming

Our fastqc results show that our example RNA-seq data contains library adapter sequences in some sequencing reads. Therefore, we consider trimming FASTQ before moving on to an alignment step. But some alignment tools including salmon and STAR can handle this issue making a trimming step not necessary at least in the case of mRNA sequencing whose read lengths are long enough to be mapped. For example, STAR adopts so-called a soft clipping strategy that ignores portions of reads that are not aligned to the genome index during the mapping process. Run the STAR alignment steps without trimming and with trimming and compare and discuss the results. Do you see any difference in alignment rates?

2. Counting reads with consideration of strand specificity

Our counting software, featureCounts can take into account whether RNA-seq data is stranded or not. In RNA-seq library preparation, the strand information can be preserved in different ways depending on the library prep method. For example, Illumina stranded RNA-seq kit uses dUTP in the second strand synthesis to amplify the selected strand only. This holds the strand information of which DNA strand is the template of RNA. In order to take advantage of stranded RNA-seq data, we need to specify it in the option of featureCounts. But you must put the correct strand information. Depending on the kit, the intact sequence of read 1 or the reverse complement of read 1 matches the RNA sequence. What if you provide the wrong information to featureCounts ? Run the featureCounts with three possible options of strandedness: “-s 0”, “-s 1” and “-s 2”. Compare the assignment rates and discuss the results.

3. Differential expression analysis with covariates

In DESeq2, you need to specify a design formula. If you have covariates in addition to your main interest variable, it is advisable to specify them in your design formula. In our RNA-seq dataset, we have a covariate of “Replicate”. Compare the DESeq2 results without and with the “Replicate” covariate. Do you find any change in the number of significant genes? Describe the results.

4. Enrichment (Over-representation) analysis p-value

We use an R package “cluterProfiler” to perform enrichment (over-representation) analysis. Pick three of the significant gene ontology terms and check the number of genes in the gene ontology term and the number of significant genes associated with the gene ontology. Confirm the p-value with your manual calculation with the hypergeometric distribution. Use “phyper” function in base R.