NGS Bioinformatics Course
AFRICA 2021

Module 7: RNA-Seq Expression Analysis (Pathogens)

Practical Assignment Answers

Assigned Tasks

Task 1: Introducing The Tutorial Dataset

1. Why is there more than one FASTQ file per sample?
   There are 2 FASTQ files for each sample (e.g. N2_sub_R1.fq.gz and N2_sub_R2.fq.gz) because the data were generated using paired-end sequencing strategy.

   For further explanations see here:
   https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

2. How many reads were generated for the N2 sample?

   There are **521,532** reads for N2.

   Ideally, you should count the reads in both files. This is because we have singletons (reads with a mate).
   Our reads look like this:

   ```
   @HWI-D00466:59:C6UBKANXX:3:1101:9580:2134/1
   NGCACCGGACGGGGTGAACCGCAGCTCGGGGTCAGCGGTCAGATTTCCGACGATGGTGATGGTGGTGTCACCAGCCACAATTTCCTCCTGCGTCCGCCTGGGTTCGGCGGCTGGTCGTGTTCTTC
   +
   #<BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
   ```

   With the FASTQ format there are four lines per read. So, as long as our files are not truncated we can count the number of lines and divide them by four.

   ```
   zless N2_sub_R1.fq.gz | wc -l
   zless N2_sub_R2.fq.gz | wc -l
   ```

   Answer: ((1,043,064 +1,043,064)/4)

3. The three Resistant samples N2, N6 and N10 represent technical replicates. True or False? Comment on your answer?
False. It really depends on asking additional questions about how the RNA-seq libraries were generated. If one batch of each isolate was maintained and three independent samples extracted then these samples represent **biological replicates.** However, if one extraction was performed, the sample vial split into three different vials from which RNA-seq data was generated then these samples represent **technical replicates.**

For RNA-seq experiments, good study design depends on having as many **biological replicates** as possible.

Task 2: Estimate Transcript Abundance With Salmon

Take a look at the *quant.sf* file in the **N2** folder and answer these questions:

4. What does *TPM* stand for?

   Transcripts Per Million

5. How many reads in total are mapped to the tRNA genes?
   (*Hint: they start with "rna-"*)

   There are 3 tRNA genes, with 2 total reads mapped to them.

6. Do you think this level of tRNA is acceptable?

   Yes, there appears to be only a few reads mapped to the tRNA genes.

Task 3: Differential Expression Analysis with DESeq2

7. How many genes are up- and down-regulated between the resistant and sensitive isolates of the Mycobacterium tuberculosis?
   4 genes up-regulated
   0 genes down-regulated

8. How many genes are significantly differentially expressed (i.e., meet the LFC threshold and adjusted p-value cutoffs) between the resistant and sensitive isolates of the *Mycobacterium tuberculosis*? Name these genes.

   3 Genes: Rv0024, Rv0089 and Rv0090.

9. What are the p-values for the significantly differentially expressed genes?

   Rv0024: 1.31171e-04

   Rv0089: 1.07364e-03

   Rv0090: 7.20022e-07