RNA-seq Homework: Module 1

For your homework, you will be tasked with using featureCounts (part of the Subread package) to generate a counts matrix to be used for DESeq2 analysis. This will involve reading through the featureCounts documentation (http://bioinf.wehi.edu.au/subread-package/SubreadUsersGuide.pdf) in the read summarization step. This will involve writing your own shell script and submitting it to the hotel queue.

In your homework folder create a subfolder (RNA\_seq\_Homework) and place your answer file in there in a pdf format (1 pt): ~/Homework/**RNA\_seq\_Homework**/**RNA\_Homework\_Lastname\_Firstname.pdf**

After placing your file there, make it accessible to us so we can collect it. Run:

chmod 777 **RNA\_Homework\_Lastname\_Firstname.pdf (**1pt)

**featureCounts questions**:

1. Write a shell script called featureCounts.sh that will read in either the .bam files you aligned yourself OR the .bam files that have been provided in the shared folder (/biom262\_2019/Module\_1/all\_bams). Run all samples together. For the job, you will require 1 node and 2 processors per node. You can allot 1 hour of walltime. Carefully review the program usage arguments and include those that you decide are necessary for out dataset. Remember the technical details of our experiment (library preparation method, sequencing run type, etc.). (use ~/anaconda2/bin/ in front of featureCounts if necessary) (3 pts)
2. Successfully submit your job to the cluster. Make sure that your job runs and finishes successfully. If it does not run to completion, assess your error files, correct your mistakes and resubmit the job. Put the output somewhere meaningful and provide us with the **full path** to the directory that contains your final featureCounts output and the path to the location of your script well as the location of your .out and .err files. (2 pts)
3. In a bulleted list, describe the flags that you chose to include, as well as the reason that you decided to include them. Pay attention to the experiment used to generate the data to guide your decisions (5 pts).

**General questions:**

1. Describe the main steps of an RNA-seq analysis from the moment when you get your raw fastq files up to and including performing featureCounts. For each of the main steps, explain the purpose of what is done computationally. (5 pts)
2. What are the similarities and differences between RPKM and TPM (describe the differences between how the methods work)? Which one might be considered preferable and why? (2 pts)

**19 points total**