Final project - RNA-seq analysis
Step-by-Step Procedure

Shane Schroeder and William Gregor

For our project we wanted to explore RNA-seq data using two different analysis pipelines. For our composite project to work correctly, we ran the kallisto-DESeq2 notebooks **first** as they download extra software that will help with the other pipeline.

Please uncompress the tgz file in the root directory.

Part 1 - Kallisto-DESeq2 notebooks

This part will deal with the following files provided in the tgz file:

- notebook-0-import-data.ipynb
- notebook-1-examining-sequence.ipynb
- notebook-2-trimming-sequence.ipynb
- notebook-3-pseudoalignment.ipynb
- notebook-4-visualize-alignment.ipynb
- notebook-5-featureCounts.ipynb
- notebook-6-DESeq2.ipynb
- SraRunTable.txt
- loadAnaconda.sh
- installPkgs.sh

Step 1 - Install and load programs using the following command

```
source installPkgs.sh
source loadAnaconda.sh
```

Step 2 - Create working directory and move files into working directory

```
mkdir -p test/notebooks
mv *.ipynb test/notebooks
mv SraRunTable.txt test/
```

Step 3 - Start jupyter notebooks

```
cd test/
jupyter notebook
```

Step 4 - There are more directions to follow along inside the notebooks

Step 5 - When you are finished with the notebooks, please delete all data inside “~/test/” except for the directory “~/test/DESeq2” because RSEM needs more space as it uses uncompressed paired fastq data, failure to delete them will cause RSEM to not complete. For context, RSEM uses ~275GB of the 500GB quota.

Important notes:
notebook-3 and notebook-4 have been modified to write slurm scripts for their main function. This is because Trimming all the files from notebook-3 and pseudo alignment from notebook-4 are both extremely intensive on the CPU and should not be run on the head node (Dr. Mueller respectfully told me not to do that).

Also if you do not finish the pipeline in one session of jupyter notebooks, you will need to do “source loadAnaconda.sh” again to reload the programs.

Part 2 - RSEM-EBSEQ slurm script

This uses some of the programs installed from the previous part so please run this second! We will only be using the files:

- RSEM-EBSEQ-ANALYSIS.sh
- get6MostDE_genes.R

This script will do all the work of creating the directories and the computation itself, it is not interactive like the previous part is.

Step 1 - Submit slurm script for execution

```
sbatch RSEM-EBSEQ-ANALYSIS.sh
```

Since this pipeline is set up to handle uncompressed paired end reads, it takes considerable time to run. Expect it to take 8-10 hours to finish.

Part 3 - Compare The Top 6 Most Differentially Between Kallisto-DESeq2 results and RSEM-EBSEQ results

We compared the genes reported from the following files to get to our conclusions:

- ~/test/DESeq2/plots/6mostSigGenes.png
- ~/rsem/exp/6mostSigGenes.csv