Hands-on introduction to RNA-seq Data Processing
(PHC 6934: Applied Computational Genomics)
Justin Gibbons
USF Omics Hub, USF Genomics program

The aim is to:

- Understand the nature of RNA-seq data
- Familiarize yourself with working on a computer cluster
- Perform the initial quality control and data processing steps of an RNA-seq study

1. **Obtaining the data**
   - **a.** The data for this project is adapted from Shaw et al 2015. The paper has been uploaded to Canvas as rna_seq_project_data_paper.pdf under Modules Lesson 6. The samples labeled “treatment” in the sample names correspond to samples that have been treated with the drug and the samples with “control” in the sample name correspond to the control samples.
   - **b.** The data is on the student cluster and can be copied to your home directory for analysis using the following command:

     ```
     cp -r /shares/biocomputing/RNA-seq_Project_Data/ .
     ```
2. **Set up for the analysis**
   - **a.** Create a new directory for your analysis

     ```
     mkdir Process_RNA-seq_Data_Project
     ```
   - **b.** Move the data FASTQ data into this new directory

     ```
     mv RNA-seq_Project_Data  Process_RNA-seq_Data_Project
     ```
   - **c.** Enter your project directory and create directories for your code, quality control results and your work directory

     ```
     mkdir  Code
     mkdir FASTQ_QC
     mkdir Work
     ```

3. **Copy the reference scripts in the Genomics_Training/RNA-seq_Training_Scripts to your new Code directory (Process_RNA-seq_Data_Project/Code)**
   - **a.** Make sure you are in your home directory (~ is a short-cut for home directory)

     ```
     cd ~
     ```
   - **b.** Copy the reference code into your new code directory (* means all files in directory)

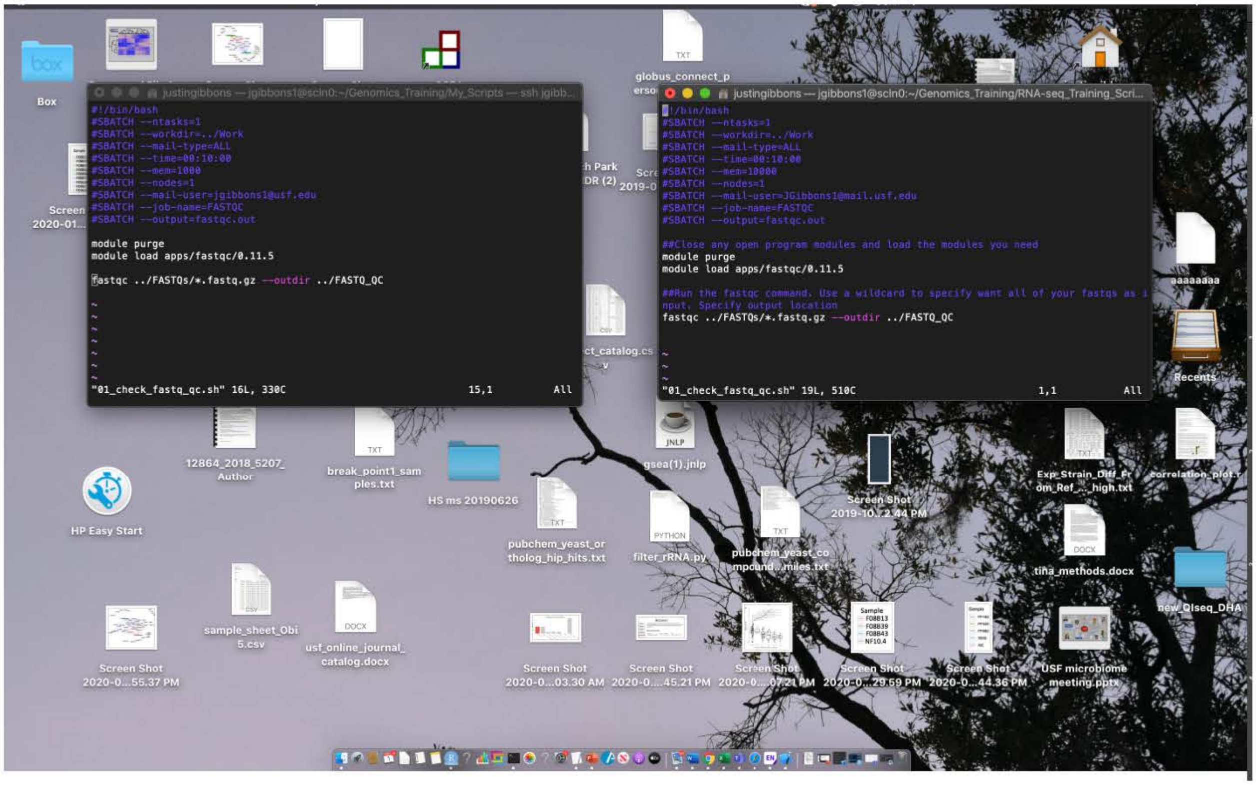
     ```
     cp  Genomics_Training/RNA-seq_Training_Scripts/* Process_RNA-seq_Data_Project/Code
     ```

4. **Modify the reference scripts to work with your new sample data. Each of the reference files starts with a number. Perform the analysis in the order indicated by the numbers. Remember to use your email for all of the submission scripts**
   - a. 01_check_fastq_qc.sh: The input file path needs to be changed to point to your new fastq files. These files can be downloaded using filezilla and viewed using a web browser.

b. 02_build_hisat_index.sh: This script will work as is (just remember to use your email).
c. 03_run_hisat.sh: The input file paths need to be changed to point your new fastqs and the output file names should be changed to reflect the new sample names (control_rep1, control_rep2, treatment_rep1, treatment_rep2).
d. 04_run_cufflinks.sh: The input files need to be renamed to point to your new sample results (the bam files from 03_run_hisat.sh) and the sample output names should be changed to reflect your new sample names.
e. 05_run_cuffnorm.sh: The input bam files need to be changed to point to your new samples and the group labels should be changed to reflect the new groups (Control and Treatment).
f. 06_run_featureCounts.sh: The input files need to be changed to point to your new samples and the outfile should be changed to reflect your new groups.

5. **Files to turn in:**
   a. The transcripts.gtf file for the control_rep1 sample. Provide a brief explanation of what a gtf file is and what it contains.
   b. The genes.fpkm_table from Cuffnorm_Output. Provide an example of what this type of data can be used for.
   c. The counts data from featureCounts. Provide a brief explanation of why we need raw counts for differential expression analysis (i.e DEseq, DEseq2, or EdgeR) and not normalized counts.

The new analysis can by just changing what the variables in scripts 3-6 are pointing to. If that is what you want to do that is fine, but you will get the best practice working on a cluster and running analysis scripts by rewriting the commands yourself. The easiest way to refer to the reference scripts while creating your new scripts is to be signed onto the cluster using 2 different terminals:



```
#!/bin/bash
#SBATCH --ntasks=1
#SBATCH --workdir=../Work
#SBATCH --mail-type=ALL
#SBATCH --time=00:10:00
#SBATCH --mem=1000
#SBATCH --nodes=1
#SBATCH --mail-user=jgibbons1@usf.edu
#SBATCH --job-name=FASTQC
#SBATCH --output=fastqc.out

module purge
module load apps/fastqc/0.11.5

fastqc ../FASTQs/*.fastq.gz --outdir ../FASTQ_QC

~
~
~
~
~
~
~
"01_check_fastq_qc.sh" 16L, 330C                     15,1          All
```

```
#!/bin/bash
#SBATCH --ntasks=1
#SBATCH --workdir=../Work
#SBATCH --mail-type=ALL
#SBATCH --time=00:10:00
#SBATCH --mem=10000
#SBATCH --nodes=1
#SBATCH --mail-user=JGibbons1@mail.usf.edu
#SBATCH --job-name=FASTQC
#SBATCH --output=fastqc.out

##Close any open program modules and load the modules you need
module purge
module load apps/fastqc/0.11.5

##Run the fastqc command. Use a wildcard to specify want all of your fastqs as i
nput. Specify output location
fastqc ../FASTQs/*.fastq.gz --outdir ../FASTQ_QC


~
~
~
"01_check_fastq_qc.sh" 19L, 510C                     1,1           All
```