1. **Install programs needed to run the RNA-seq data pre-processing**

a. #Install pip under TSCC environment
cd
mkdir programs
cd programs
curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py
python get-pip.py --user

b. #Next, make a virtual environment to run MultiQC
cd
pip install virtualenv --user
virtualenv multqc_env
source multqc_env/bin/activate
pip install multiqc

c. # Exit the environment
deactivate

d. # Load Fastqc -QC tool
module load fastqc

e. # Install cutadapt -cut out adapters from the sequences
pip install cutadapt --user

f. # Check that cutadapt is installed
cutadapt --version
1.18

g. # Check that FastQC is installed
fastqc -v

h. #Install Trimgalore -tool to remove low quality reads
curl -fsSL https://github.com/FelixKrueger/TrimGalore/archive/0.6.5.tar.gz -o trim_galore.tar.gz
tar xvzf trim_galore.tar.gz

i. # Download and install Kallisto - Quantifies counts in the transcriptome
wget https://github.com/pachterlab/kallisto/releases/download/v0.46.1/kallisto_linux-v0.46.1.tar.gz
tar -xzvf kallisto_linux-v0.46.1.tar.gz

2. Run FastQC on raw FASTQ samples

#Run fastqc for the FASTQ files
(base) [prvaldes@tscc-login2 ~]$ qsub -I -q condo -l walltime=8:00:00 -l nodes=2:ppn=24
qsub: waiting for job 27215286.tscc-mgr7.local to start
qsub: job 27215286.tscc-mgr7.local ready
(base) [prvaldes@tscc-13-37 Chen_Foundation_FASTQ]$ fastqc *.fastq.gz
Started analysis of CN19009_sc15_b2_AD_S68_L004_R1_001.fastq.gz
Approx 5% complete for CN19009_sc15_b2_AD_S68_L004_R1_001.fastq.gz
Approx 10% complete for CN19009_sc15_b2_AD_S68_L004_R1_001.fastq.gz
[example]
Output = .fastqc.html files and .fastqc.zip files

3. Run MultiQC on raw fastqc samples

#Run MultiQC for the raw FastQC files
(base) [prvaldes@tscc-login2 ~]$ source multqc_env/bin/activate
(multqc_env) (base) [prvaldes@tscc-login12 FastQC]$ multiqc /home/prvaldes/scratch/Chen_Foundation_FASTQ/FastQC
[WARNING] multiqc : MultiQC Version v1.11 now available!
[INFO] multiqc : This is MultiQC v1.8
[INFO] multiqc : Template : default
[WARNING] multiqc : You are running MultiQC with Python 2.7.5
[WARNING] multiqc : Please upgrade! MultiQC will soon drop support for Python < 3.6
[INFO] multiqc : Searching : /home/prvaldes/scratch/Chen_Foundation_FASTQ/FastQC
[INFO] fastqc : Found 54 reports
[INFO] multiqc : Compressing plot data
[INFO] multiqc : Report : multiqc_report.html
[INFO] multiqc : Data : multiqc_data
[INFO] multiqc : MultiQC complete

4. Run Trimgalore!
(base) [prvaldes@tscc-login11 ~]$ qsub -I -q condo -l walltime=8:00:00 -l nodes=2:ppn=24
qsub: waiting for job 27224483.tscc-mgr7.local to start
qsub: job 27224483.tscc-mgr7.local ready

(base) [prvaldes@tscc-0-49 Trimgalore]$ bash Trimgalore_RNA_EOAD.NDC.sh

5. Run MultiQC on Trimmed, Validated Files

(base) [prvaldes@tscc-login2 ~]$ (base) [prvaldes@tscc-login11 clean_Chen_Foundation_FastQC_ALL]$ source /home/prvaldes/multqc_env/bin/activate
(multqc_env) (base) [prvaldes@tscc-login11 clean_Chen_Foundation_FastQC_ALL]$ multiqc /home/prvaldes/scratch/Trimgalore/clean_Chen_Foundation_FastQC_ALL

6. Make transcriptome index with a kmer length of k=31

#Get cDNA file
(base) [prvaldes@tscc-login11 homo_sapiens_104]$ wget ftp://ftp.ensembl.org/pub/release-104/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz

#Get non-coding RNA file
(base) [prvaldes@tscc-login11 homo_sapiens_104]$ wget ftp://ftp.ensembl.org/pub/release-104/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz

#Concatenate both files together into one file
#Notes from here: https://www.biostars.org/p/81924/
(base) [prvaldes@tscc-login11 homo_sapiens_104]$ cat Homo_sapiens.GRCh38.cdna.all.fa.gz Homo_sapiens.GRCh38.ncrna.fa.gz > Homo_sapiens.GRCh38.cdna.all.ncrna.fa.gz

#Start the screen in a new window
screen
#Submit interactive jobs to the home-shankar
[prvaldes@tscc-login2 ~]$ qsub -I -q condo -l walltime=8:00:00 -l nodes=2:ppn=24
qsub: waiting for job 27223067.tscc-mgr7.local to start
qsub: job 27223067.tscc-mgr7.local ready

#Build the transcriptome index using kmer count of 31 using Kallisto
#kallisto index builds an index from a FASTA formatted file of target sequences.
(base) [prvaldes@tscc-0-49 homo_sapiens_104]$ /home/prvaldes/programs/kallisto/kallisto index -k 31 -i Homo_sapiens.GRCh38.cdna.all.release-104_k31.idx /home/prvaldes/programs/homo_sapiens_104/Homo_sapiens.GRCh38.cdna.all.ncrna.fa.gz

7. Run Kallisto

(base) [prvaldes@tscc-login11 ~]$ qsub -I -q home-shankar -l walltime=48:00:00 -l nodes=1:ppn=24

(base) [prvaldes@tscc-2-13 Kallisto]$ bash KallistoScript_RNA_EOAD.NDC.sh

8. **Run MultiQC on Kallisto Files**

Note: runs on kallisto.log files
(base) [prvaldes@tscc-login2 ~]$ (base) [prvaldes@tscc-login11 clean_Chen_Foundation_FastQC_ALL]$ source /home/prvaldes/multqc_env/bin/activate
(multqc_env) (base) [prvaldes@tscc-login11 clean_Chen_Foundation_FastQC_ALL]$ multiqc /home/prvaldes/scratch/KallistoOut_RNA_Chen

9. **Proceed with downstream quantified transcript counts from Kallisto RNA- using EOAD.RNAseq.Kimma.nVenn.Analysis.Rmd script file.**