RNA-seq data processing with STAR and FeatureCounts

Project/Data: ISH Tau 12m Ribotag

Written by: JB

Last Updated: 03/11/2022

This doc contains the commands and explanations for the analysis of bulk RNA-sequencing using STAR and FeatureCounts up to DEG analysis using R/Rstudio

A few notes before gettting stared:

1. All data and analysis should be completed in the `/data` directory of our Lab Server due to space constraints in your individual home directories
2. Keep your files organized - this will make your life a whole lot easier. I recommend something like this:
   - /data/projectname/rawdata
     - store your raw fasta files here
     - also consider storing a list of your sample names here
   - /data/projectname/star/
     - make subdirectories for alignments, sorted-bam files, etc.
     - keep your scripts in the main directory here
3. Review the Linux/Bash links on the Github Page for assistance with navigating the server

Download Data (fastq files). Below are two examples of how to download data:

```
wget http://link/to/data/.fastq.gz>
```

```
wget -m ftp://username:password@link/to/data
```

Install Conda, other packages, and set up environment

-You can find information about Anaconda and Miniconda here -Right click on the download button and click "copy link address". Paste this link after wget command (see example below)

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
```

```
bash Miniconda3-latest-Linux-x86_64.sh
```

It will ask you a bunch of questions. Make sure that you use the defaults for all of them (press enter for all). Start pressing ENTER through all the license agreements. Type 'yes' when it asks if you agree to everything. One of the important questions it will ask during installation is if you want to add this to the PATH variable in your ~/.bashrc Check that the anaconda bin was added to your path by looking at your bashrc file:

```
less ~/.bashrc
```

There should be a line in there that looks something like this located BELOW the line "#User specific aliases and functions":

```
export PATH=/home/USERNAME/miniconda3/bin:$PATH
```

Then install necessary packges. You may need to make and activate a new conda environment in order to install new packages. More information about environment can be found here More information about Anaconda installation can be found here

```
conda install -c bioconda STAR fastqc subread
```

Remember to activate conda environment that has STAR installed

```
conda activate [environment name]
conda info --envs #to see environments
```

Building STAR reference index:

You need reference genome (fasta files) and gene annotation (GTF). These can be found on UCSC Genome Browser and Gencode Right click on the files and click "copy link address". Download these in a directory named /references/name_of_reference_w_version_ID. Unzip and merge fasta files into one file.

```
cd /data/Isabel/STAR/references/mm39
wget https://hgdownload.soe.ucsc.edu/goldenPath/mm39/bigZips/mm39.chromFa.tar.gz
tar -xvf chromFa.tar.gz
cat *.fa > mm39.fasta
wget https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_39/gencode.v39.annotat
gunzip gencode.vM28.annotation.gtf.gz
```

Create STAR reference index

```
cd /data/Isabel/STAR/Tau12mo/references/mm39/star_genome_index

STAR --runThreadN 8 \
--runMode genomeGenerate \
--genomeDir /data/Isabel/STAR/references/mm39/star_genome_index/ \
--genomeFastaFiles /data/Isabel/STAR/references/mm39/mm39.fasta \
--sjdbGTFfile /data/Isabel/STAR/references/mm39/gencode.vM28.annotation.gtf
```

Mapping your samples: Create a script to loop through your samples. The example below is for paired samples which are split into to txt files (namesR1 and namesR2).

```
vi star.sh #to create script
```

Your script should look something like this:

```
#!/bin/bash

while read -r f1 f2; do
    STAR --readFilesCommand zcat \
    --genomeDir /data/Isabel/STAR/references/mm39/star_genome_index \
    --runThreadN 8 \
    --readFilesIn /data/Isabel/raw_data/Tau12m/${f1} /data/Isabel/raw_data/Tau12m/${f2} \
    --outFileNamePrefix /data/Isabel/STAR/Tau12mo/Alignments/$f1 \
    --genomeLoad LoadAndKeep
done < <(paste namesR1.txt namesR2.txt)
```

To run the script, first change the permissions and then use nohup to run without interruption. It's best to run 1-2 samples first to ensure everything is working correctly.

```
chmod 777 star.sh
nohup ./star.sh &
tail -f nohup.out
```

Convert sam files to sorted bam files

```
ls | grep .*sam > sam_names.txt
mkdir /data/Isabel/STAR/Tau12mo/sorted_bam

while read -r i; do
    samtools view -bS /data/Isabel/STAR/Tau12mo/Alignments/$i | samtools sort -o /data/Isabel
done < sam_names.txt
```

Summarize counts with FeatureCounts from the Subread package

```
featureCounts -T 16 \
-a /data/Isabel/STAR/references/mm39/gencode.vM28.annotation.gtf \
-p \
-s 2 \
-g gene_id \
-o /data/Isabel/STAR/Tau12mo/counts.txt \
/data/Isabel/STAR/Tau12mo/sorted_bam/*.sam
```

Download counts.txt to your local computer or use RStudio on server for DEG analysis. Use FileZilla or command line (from local computer)

```
scp -r YourUsername@mnla1.salk.edu:~/path/to/output/files /path/to/local/directory
```