RNA-seq data processing and analysis with Salmon

Project/Data: Aging Astrocyte Transcriptome Paper (doi: 10.1016/j.celrep.2017.12.039)

Written by: JB

Last Updated: 08/26/2022

This doc contains the commands and explanations for the analysis of bulk RNA-sequencing using Salmon 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/salmon/
        - ♠ make subdirectories for the output files
3. Review the Linux/Bash links on the Github Page for assistance with navigating the server
4. Review Salmon Documentation

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

wget -r -nH -nd -np http://link/to/data/.fastqz

from UCSD: wget -m ftp://username:password@link/to/data/.fastq.gz>
```

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 salmon
```

Remember to activate conda environment that has STAR installed

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

Building Salmon reference index:

You need reference genome (fasta files) of transcriptome (cDNA) 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.

```
mkdir /data/jburgado_data/salmon/mm39_index
cd /data/jburgado_data/salmon/mm39_index
wget https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M30/gencode.vM30.trans
salmon index -t Mus_musculus.GRCm39.cdna.all.fa.gz -i mm39_index
```

Running Salmon

```
ls | grep .*R1> R1_names.txt # create file names for R1 and R2 files separately. This is done
ls | grep .*R2> R2_names.txt

mkdir /data/jburgado_data/salmon/outputs/AgingAstrocyteTranscritpome_output #create directory
vi salmon_AgingAstrocyteTranscriptome.sh #create script
```

The script should look something like this:

#!/bin/bash while read -r f1 f2; do salmon quant -i /data/jburgado_data/salmon/mm39_index -l A
-1 <(gunzip -c /data/jburgado_data/rawdata/AgingAstrocyteTranscriptome/$)
-2 <(gunzip -c /data/jburgado_data/rawdata/AgingAstrocyteTranscriptome/$)
-p 8 --validateMappings
-o /data/jburgado_data/salmon/outputs/AgingAstrocyteTranscriptome_output/$f1 done < <(paste R1_names.txt R2_names.txt)

You're done!

Transfer all of the output files (easiest way is via FileZilla) to your local computer