The following is a guide for analyzing RNASeq data as received from the core. It has been optimized for Mac but should be able to be translated to Windows with some changes to the software used to access the terminal or remote server.

Download :

[“R”](https://www.r-project.org/)

[RStudio](https://www.rstudio.com/products/rstudio/) (the free version is all we need)

[iTerm2](https://iterm2.com/)

[Filezilla](https://filezilla-project.org/)

[Contact hpc3 and get a login account for the supercomputer. Instructions are here.](https://rcic.uci.edu/hpc3/hpc3-reference.html)

Fastq files are received from the sequencing core as txt.gz

These are zipped files and are large, leave them zipped. The code will unzip them as they use them.

1. First thing we must do is trim the adapters that are used for the sequencing reaction.

Code for this is found here:

[**Trimmomatic**](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/Trimming.py)

1. Next we will check the trimmed sequences for quality using the code found here:

[**Fastqc**](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/FastQC.R)

This will give us html breakdown of the sequencing quality before we proceed with the genome alignment.

1. Build the reference genome. Download the version of the genome you will be using to align your reads. Human, mouse etc. I found the version of the mouse genome at [UCSC](https://genome.ucsc.edu/cgi-bin/hgGateway).
2. For alignment we need to build the reference genome and load our samples onto the server. To do this, I use Filezilla and connect to the host: sftp://hpc3.rcic.uci.edu

Login using your uci credentials

Change the directory to /pub/$UserID

Make a new file here called Genome\_Directory that will take the assembled genome

Make another new file called GTF\_Dir and upload the annotation file

Make another new file called mm39 or whatever you want to call your genome sequence and move the genome sequence reference file there.

Next, you need to submit a Slurm session to the supercomputer which is basically using the processing power of the University’s computer system to run this task. More details regarding this are in the code found here: [Slurm\_STAR.sh](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/Slurm_STAR.sh)

This task took about 40 minutes

The outputs here in Genome\_Directory will be used to align our trimmed reads.

Make a new directory called Trimmed\_unzipped. Move your trimmed and unzipped .fastq files here. I did this one or two at a time because the size of the files is time limiting. It takes about 10 minutes to move one read file into the directory. Now you can run the mapping found here: [Slurm\_STAR2.sh](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/Slurm_STAR2.sh)

After uploading the sequence the alignment took about 5-9 minutes.

The output here are various logs and a BAM file which we will use in the next step to generate read counts. Move the files from the server and logout.

5. Generating read counts:

Here we are going to use the mapped files to count and compare the gene expression between given conditions.

Code can be found here: [featureCounts](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/featureCounts.sh)

6. Running first stats on the list with DESeq2 at the end will generate the list of significantly regulated genes

Code can be found here: [DESeq2](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/DESeq2.R)

7. Now the tricky part. We have a list of ENSEMBL IDs and we have to import the gene Ids from the annotation file, match them up with the ENSEMBL ID and then add them in a new column. We will use python for this using jupyter notebook. Open your terminal and install [jupyter notebook](https://jupyter.org/install). You will also need to install the [pandas](https://pandas.pydata.org/docs/getting_started/install.html) package to move these big files around. Set you working directory on your shell to where your files are and type jupyter-notebook into the command line. It will open up a command line through python running webpage. Open my jupyter notebook from github [found here](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/Adding%20gene%20names.ipynb).

At the end you will have a csv file you can open and manipulate in excel.