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.

Bookmark: <https://github.com/erichanse/RNASeqCode>

This is a link to all of the code chunks.

Download :

[“R”](https://www.r-project.org/)

[RStudio](https://www.rstudio.com/products/rstudio/)

[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)

[GSEA](https://www.gsea-msigdb.org/gsea/index.jsp)



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.

First thing we must do is trim the adapters that are used for the sequencing reaction. Check with the core to identify which adapter package you will need. For the Novaseq and MiSeq instruments they usually use the common Illumina adapters that Trimmomatic knows how to look for. Most of the time you should not need to trim you sequences but its good practice.

Code for this is found here: [**Trimmomatic**](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/Trimming.py)

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 files of each sequencing read and breakdown of the quality before we proceed with the genome alignment.

Next we must build the reference genome. Download the version of the genome you will be using to align your reads. Human, mouse etc. I have found that Gencode has the best sequences for this. For example, the genome file for human used most recently was called:

GRCh38.p13.genome.fa

And the annotation file used was:

Gencode.v42.basic.annotation.gtf

Put these files onto the server using the directions below:

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. You have to do some extra steps with ssh to generate keys for duo login. This is how they handle two factor authentication in shell.

The link to do that is here: [ssh login](https://rcic.uci.edu/hpc3/ssh.html)

Change the directory to /pub/$UserID Make sure you are using the public directory because the memory allotment is much higher than what you get for your personal file. The difference is that this memory is wiped after a few months so it is really for temporary use and that is what we will use it for here.

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 hg38 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. Login on shell using:

ssh [USERID@hpc3.rcic.uci.edu](mailto:USERID@hpc3.rcic.uci.edu)

next allocate memory and onto a node in the supercomputer,

For genome alignment I used the following parameters:

**srun --mem=100G -p free --pty /bin/bash -i**

The code to run this step is found here: [Genome\_Alignment.sh](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/Genome_Assembly_Slurm.sh)

This task took about 60 minutes. You may have to increase the amount of files you can have open at once during STAR. If you get an error about not being able to write a file after many files have already been written in the Genome Directory file, try this code to increase the amount of files you can have open (this only happened to me with human samples)

ulimit -n 16384

default is set to 8000 and that hits a limit late in the process.

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 depending on your internet speed. 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 for mouse and 40 minutes for human. Ask for 100MB 4 cores and run 8 threads.

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.

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)

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%20Master.Rmd)

7. Now 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 your working directory on your shell to where your files are and type jupyter-notebook into the command line. It will open 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.

To prepare for the GSEA: Use the following code. You will need to pull out the raw counts as log base 2. This code will accomplish that, you may repeat some of the steps as before so don’t be alarmed:

[GSEA from DESEQ2](https://github.com/erichanse/RNASeqCode/blob/main/Kong_RNASeq/DESeq2_to_GSEA.Rmd)

Next you have to prepare the output file from this for GSEA. This requires formatting the first couple of rows of the spreadsheet as outlined in the instructions found here

[Making GCT files for GSEA](https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats)

You also need to make a phenotype file that includes the conditions for your experiment using the specific instructions found here

[Making CLS files for GSEA](https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats)

Upload these files into the GSEA app.

Select the ENSEMBL\_Geneid reference section, (most recent version)

Run the GSEA. The output files will be under the /users directory of your computer