Stoneman, Charli

cs660@exeter.ac.uk

Abstract

A step-by-step guide, with links to all scripts, to carrying out RNA sequencing analysis once results have been received from the sequencing service.

RNA Sequencing Analysis Pipline

April 2020

Table of Contents

[RNA Sequencing Analysis protocol 2](#_Toc42163785)

[*Helpful commands* 2](#_Toc42163786)

[1. Downloading sequencing data onto server 2](#_Toc42163787)

[2. Quality Control of sequencing data 2](#_Toc42163788)

[3. Additional Trimming using Trimmomatic 3](#_Toc42163789)

[4. Quality Control of FastQ files 3](#_Toc42163790)

[5. Alignment of RNA sequencing data to reference genome 3](#_Toc42163791)

[6. Visualise the alignment files 3](#_Toc42163792)

[7. Quality Control of the alignment files 3](#_Toc42163793)

[8. Create index file of the of the bam files 4](#_Toc42163794)

[9. IGV Viewing of the alignment files 4](#_Toc42163795)

[10. QC of the alignment files using QoRTs 4](#_Toc42163796)

[11. Gene Quantification 4](#_Toc42163797)

[12. Visualising featureCounts results 5](#_Toc42163798)

[13. Differential Gene Analysis 5](#_Toc42163799)

[14. Gene Ontology Analysis 5](#_Toc42163800)

# RNA Sequencing Analysis protocol

## *Helpful commands*

Make sure any scripts are saved on Linux environment as shell extensions (**.sh**)

To change permissions on files so they are executable, readable and writeable: chmod 775 ./file\_name

Execute a script in the queue: msub script\_name.sh

Check a file: less file\_name.txt

## Downloading sequencing data onto server

Once you have received email from Exeter Sequencing Service, you can download the results onto server using the following command

wget -r -p <ftp://Username:Password@ftp1.sequencing.exeter.ac.uk>

* To run this in the background so if you log off it still continues to download add in the -b option. This will produce a wget-log file with information on where it is in the downloading.
* To look into the wget-log file use: less wget-log

## Quality Control of sequencing data

1. Exeter Sequencing Service will have already performed QC analysis on the RNA raw data using FastQC and will have produced QC plots for you to look at using MultiQC. On your download portal you can find these results under the Additional tab

H0253\_3064\_sequencing\_qc.html

1. Open up this file and view your QC results. In many cases, they have already performed trimming to remove the adapter sequences from the reads using the *Cutadapt version 1.13 tool*.

* Paul O’Neil reported that the trimmed date removes the adapter sequences and any low quality bases from the 3’ end.
* Sequencing adapters are removed using sequencing defined here: <http://emea.support.illumina.com/bulletins/2016/12/what-sequences-do-i-use-for-adapter-trimming.html>
* Low quality bases (<Q22) were trimmed.
* Reads shorter than 25 were discarded.

1. In the QC file: under the General Statistics section, check that % Dups is around 50%, %GC is around 50% and length is similar across samples. The total number of sequences for each sample can also be seen in the final column.
2. Under FastQC: Sequence Quality Histograms check that all bases have a Phred score > 28 so are in the green section of the graph. If the first 5 bases are not some additional trimming may be needed.
3. Per Sequence GC content should have a normal distribution with a peak just under 50%
4. Per Base N Content should be very close to 0
5. It is fine if you have a sudden rise at the end of the sequence length distribution plot as samples have been trimmed by sequencing service to be the same length.
6. A little peak in sequence duplications peak is seen commonly so do not be concerned about this
7. There should be no over-represented sequences
8. There should be no adapter content as these should have already been removed by Cutadapt by the sequencing team.

## Additional Trimming using Trimmomatic

Additional trimming can be carried out if the QC from the raw reads is not acceptable – sometimes you may need to remove the first 5 bases from each read as these may have a poor Phred score (<28).

Please see script *03\_trimming\_using\_trimmomatic.txt*

## Quality Control of FastQ files

Re-run FastQC to perform quality control on the trimmed data.

See script *04\_fastqc.txt*

## Alignment of RNA sequencing data to reference genome

Carried out using STAR.

See script 05\_STAR\_alignment

Once alignment has been run, check alignment scores in the .log.final.out files.

Looking for:

* 1. Uniquely mapped reads > 80%
  2. Reads mapped to multiple loci
  3. Reads unmapped
  4. Average mapping length should be the length of the sequence

## Visualise the alignment files

Carried out using R. See pdf document for workthrough and code 06\_visualising\_STAR\_alignments.pdf

## Quality Control of the alignment files

Run multiQC on the alignment files

module load MultiQC/1.2-intel-2017b-Python-2.7.14

multiqc ./alignment\_STAR/ --dirs --interactive -o ./multiQC/

## Create index file of the of the bam files

Most downstream tools will require an index file (.bam.bai) of the bam files produced after alignment.

See script 08\_create\_bam\_index.txt

## IGV Viewing of the alignment files

To do this, can either load each individual bam file into IGV and view the region you desire or you can merge together the bam files for each cell line. To view in IGV. an index file is required.

* + - 1. To merge bam files and create index see script 09\_merge\_bam\_for\_IGV.txt
      2. View in IGV

- For more information on IGV and how to download,look here: http://software.broadinstitute.org/software/igv/download. There are really good online tutorials into how to best use IGV.

a. Load the corresponding genome e.g. hg19 from the left hand side drop down list on the upper left of the IGV window

b. Load in the bam file which you want to view

* 1. Load from file
  2. Select bam file. IGV will then automatically detect and look for the index file of this bam file which we created earlier.
  3. Loading a BAM file creates up to 3 associated tracks:

Alignment Track to view individual aligned reads

Coverage Track to view depth of coverage

Splice Junction Track which provides an alternative view of reads spanning splice junctions

* 1. You can then zoom into the gene of interest by either typing the gene name into the search bar or by entering the coordinates of the gene
  2. Investigate whether the region expected has been deleted i.e no reads are mapping to this area
  3. It is also important to have a look at a few housekeeping genes to ensure alignment has worked and that there is good coverage and read depth across the genes.

## QC of the alignment files using QoRTs

This step is often not carried out if the alignment stats look good but can be done for sanity check and to get more figures that can be used to show alignment looks good.

See script 10\_QoRTs

## Gene Quantification

There are two ways to do this;

1. Assign all the reads to a given gene
2. Infer the quantity of individual transcripts

There are also several different programs which can be used to carry out gene quantification. For this purpose, to count the number of reads for each gene, we will use featureCounts package. See script 11\_featureCounts

featureCounts outputs several different files;

1. Count table
   1. This includes annotation columns (`Geneid', `Chr', `Start', `End', `Strand' and `Length') and data columns (read counts for genes for each library)
2. Counting summary file
   1. This file includes information on the total number of alignments that were successfully assigned to genes and the number of alignments which failed to be assigned due to multiple reasons.
   2. Note that the counting summary includes the number of alignments, not the number of reads. Number of alignments will be higher than the number of reads when multi-mapping reads are included since each multi-mapping read contains more than one alignment.
3. Screen ouput
   1. Contains information similar to the summary file with number and percentages of successfully assigned alignments

## Visualising featureCounts results

It is quite good practice to then plot the statistics produced from featureCounts so you can assess if the quantification worked and can compare the number of alignments that have been assigned vs those that have not.

See 12\_visualising\_featureCounts.pdf for R code and run through example

## Differential Gene Analysis

This is carried out in R using the DeSeq2 package. For run through example and R script, see 13\_Differential\_Gene\_Expression\_Analysis.pdf

## Gene Ontology Analysis

Once you have a list of differentially expressed genes, gene ontology enrichment analysis can be carried out using DAVID (https://david.ncifcrf.gov ) or Enrichr (https://amp.pharm.mssm.edu/Enrichr/) – both give similar results. Just plug in a list of gene names to both online tools.

The main categories to look for enriched terms if KEGG pathways, GO biological pathways, GO molecular function, GO Cellular Component, OMIM Disease.