Devya Gurung

Gen 811: Final Project

Dr. Miller

**Project: RNA seq analysis in Mouse Breast Cancer**

**Background**:

Cancer is a byproduct of many dysregulations, both within the cell and in the tumor microenvironment. It can occur due to mutations, or environmental factors. Breast cancer is one of the most common cancers found in women. In the United States alone, 255,000 cases of breast cancer are diagnosed in women and 2300 in men (CDC). Breast cancer is caused when cells in breast grow uncontrollably. In the study where I gathered raw fastq data from studied the breast cancer in genetically engineered mouse models. The mouse models carried Trp53-mutated breast cancer, along with Brca1 and Brca2 associated breast cancer and E-cadherin (Cdh1) mutated lobular breast cancer.

This study was done in mice model to further understand the mechanism in human breast cancer. Cancer compromises the functions of the immune system, therefore understanding how it regulates cells and the microenvironment is vital in developing cancer immunotherapy.

**Citations**

1. <https://www.cdc.gov/cancer/breast/basic_info/index.htm>
2. Varela I, Klijn C, Stephens PJ, et al. Somatic structural rearrangements in genetically engineered mouse mammary tumors. Genome Biology. 2010 ;11(10):R100. DOI: 10.1186/gb-2010-11-10-r100. PMID: 20942901; PMCID: PMC3218656

**Various Tools and Packages** were installed while performing RNA seq analysis.

STAR, fastqc, featureCounts, and R.

**Methods:**

RNA-Seq Analysis Protocol

STAR Alignment

Raw data: Paired wise (2per/Sample)

Fastq files, Ref. Genome(Idx), Samples

Bam files (1per/Sample)

Samtools used on bam files.

1

Features (GTF)

featureCounts

CountMatrix

Column Data

Goal: RLD: normalized exp. Data

Res: Sig diff exp genes

DESeq2

#Work through the pipelines

Methods Steps:

1. 1st get the fastq raw data (RNA seq) from public databases.

2. Then run fastqc to check the quality of your reads.

3. Fastqc files were checked in Filezilla. Filezilla was installed in computer

4. Next, perform trimming.

5. Trimming was done via trim\_scriptV2 which has trimming scripts and parameters inside

6. Next, Fastqc were performed in trimming files. The trimmed fastqc files were checked in Filezilla

7. Then, Genome Index were created using STAR

8. STAR was also used to do alignments of the sequence. Ran pair-end reads

9. After STAR, output files such as .bam were generated which was used in next step.

10. FeatureCounts was performed on .bam files output

11. Samtools was used on bam files to generate bam stats files, to create bam plot files.

12. The output from featureCounts were used in R to see files.

**Fastqc** were run on raw fastq files.

Then the ouput files were viewed in Filezilla to see the sequence quality.

Filezilla is a tool that allows you to visualize files, but you need to connect to Ron.

Raw Trimmed

Graphical user interface, chart

Description automatically generated

Raw Trimmed

Chart, bar chart

Description automatically generated

**Untrimmed file FastQc Report**

Chart

Description automatically generated

**Trimmed file FastQc Report**

Chart

Description automatically generated

To generate genome index for Mouse or Human, use Gencode.

For this project, Mouse genome was used.

<https://www.gencodegenes.org/> options to filter: Use GTF file of (regions: CHR) Comprehensive gene annotation and Fasta files (Genome sequence, Regions: ALL).

Graphical user interface, application

Description automatically generated

Once the fasta files and gtf files were downloaded.

Star was performed for index and alignment.

Star: Generating genome index

Text

Description automatically generated

Alignment Output files: 1st Pair

Text

Description automatically generated with medium confidence

Alignment Output files: 2nd Pair

Text

Description automatically generated

less ERR015623Log.final.out

Text

Description automatically generated

less ERR015618Log.final.out

Text

Description automatically generated

Output of FeatureCounts

Text

Description automatically generated

Text

Description automatically generated

**Samtools output files from bam files such as png files were viewed in Filezilla.**

**Fig below:** Indels: Insertion/deletions stats

Insertions/deletions ratio in 2 samples look slightly different. The insertion/deletions ratio in sample2 **ERR015623** has higher ratio difference than sample1 **ERR015618**.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

**Fig below: indel cycles**

**The indel cycles:**

insertions/deletions in both forward and reverse of 2 samples looked similar.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

**Fig below:** GC depth

The GC content in sample1 **ERR015618** fell in-between 30-40th percentile.

But, in sample2 **ERR015623**, the GC content was below 10th percentile of mapped sequence.

Chart, line chart

Description automatically generated Chart

Description automatically generated

**Fig below:**

Quality per read (Reverse and Forward) in both samples.

Chart

Description automatically generated

**Analysis in R: The featureCounts output file were used in R.**

Summary of ERR015623Aligned.out.bam

Text

Description automatically generated

**Summary** of ERR015618Aligned.out.bam

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

**Working in R with GTEX data**

Due to challenges that I have withing analyzing my data in R after featureCounts, I worked with GTEX files that had RNA seq data already loaded.

**Analysis of GTEX files in R.**

# Running Libraries

library(ggplot2)

library(dplyr)

library(tidyr)

# Setting working directory

setwd("~/GTEx")

File samples

samples <- read.csv("data/samples.csv")

View(samples)

Text

Description automatically generated with medium confidence

Letter

Description automatically generated with medium confidence

# this is a code for ggplot with color

ggplot(samples, aes(x = SMTS, color = SEX)) +

geom\_bar(stat = "count") +

coord\_flip()

Chart

Description automatically generated

# this is a quick summary of your data

ggplot(samples, aes(x = SMTS, fill = DTHHRDY)) +

geom\_bar(stat = "count") +

coord\_flip() +

facet\_wrap(~SEX)

Chart, timeline

Description automatically generated with medium confidence

# Boxplot # (SMCENTER = seq center), facet\_wrap (~ needs tilda)

# color coded, aes is in 1st line specified, but everything below has them aes

ggplot(samples, aes(x = SMCENTER, y = SMRIN, color = DTHHRDY)) +

geom\_boxplot() +

geom\_jitter() +

facet\_wrap(~SEX)

Chart, scatter chart

Description automatically generated

# color coded, aes is in 1st line specified, but everything below has them aes

ggplot(samples, aes(x = SMCENTER, y = SMRIN)) +

geom\_boxplot() +

geom\_jitter(aes(color = SMRIN))+

facet\_wrap(~SEX)

Chart, scatter chart

Description automatically generated

# importing data, file is in .tsv format

results <- read.table("data/GTEx\_Heart\_20-29\_vs\_50-59.tsv")

head(results)

**Fig. Below. Differentially expressed genes based on adjusted P. Values**

# filtering steps # giving P values

# these are genes that are diff expressed

resultsDEGs <- filter(results, adj.P.Val < 0.05,

logFC > 1 | logFC < -1) %>%

arrange(desc(adj.P.Val)) # arranging in descending order

resultsDEGs

View(resultsDEGs)

# other way to do it

resultsDEGs <- results %>%

filter(adj.P.Val < 0.05,

logFC > 1 | logFC < -1) %>%

arrange(desc(adj.P.Val))

resultsDEGs

Graphical user interface, application, table

Description automatically generated