Breast Cancer Prognosis between 2 Populations

About

This task was carried out by me to complete the hackbio internship

Background

The objective of this task is to perform transcriptomic analysis between two population to get insights about the mutation which causes breast cancer. For the analysis I have selected two population Europe and USA, then performed differential expression analysis on these two datasets and find upregulated and downregulated genes.

Methodology

Data collection

The datasets are collected from the GEO (gene expression omnibus) which we can access through the NCBI, which is a major resource for bioinformatics tools and services. The downloaded dataset contain detailed information about the transcriptomic profile of breast cancer of two population USA and Europe.

Software packages

1. Rstudio (R language)
2. R packages used-
   a) `BiocManager`
   b) `Forcats`
   c) `Stringr`
   d) `ggplot2`
   e) `ggrepel`
   f) `readr`
   g) `tidyr`
   h) `survminer`
   i) `GEOquery`
   j) `Limma`
   k) `Pheatmap`

Analysis

For installing R packages-

```
install.packages("BiocManager")
install.packages("forcats")
install.packages("stringr")
install.packages("ggplot2")
install.packages("ggrepel")
install.packages("readr")
install.packages("tidyr")
install.packages("survminer")
BiocManager::install("GEOquery")
BiocManager::install("limma")
BiocManager::install("pheatmap")
BiocManager::install("org.Hs.eg.db")
```

Next I have imported the datasets-

```
library(GEOquery)

US_population_id <-
"GSE171957"
Europe_population_id <-
"GSE180186"
gse1 <-
getGEO(US_population_id)
gse2 <-
getGEO(Europe_population_id)
```

Now I have performed the differential gene expression analysis using Rstudio.

```
library(limma)
design1 <- model.matrix(~0+sampleInfo1$group)
design1
design2 <- model.matrix(~0+sampleInfo2$group)
design2
## the column names are a bit ugly, so we will rename
colnames(design1) <- c("Ductal","lobular")
colnames(design2) <- c("Ductal","lobular")
summary(exprs(gse1))
summary(exprs(gse2))
## calculate median expression level
cutoff1 <- median(exprs(gse1))
cutoff2 <- median(exprs(gse2))
## TRUE or FALSE for whether each gene is "expressed" in each sample
is_expressed1 <- exprs(gse1) > cutoff1
is_expressed2 <- exprs(gse2) > cutoff2
```

Result and Discussion-

From the gene expression analysis, the upregulated and downregulated genes have been found.

For the Europe sample-

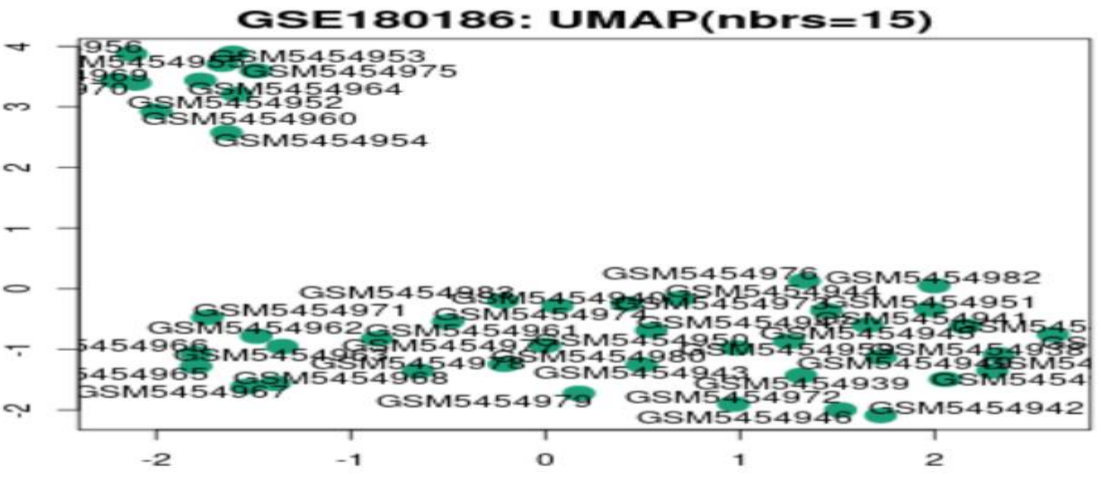

Figure1- shows upregulated and downregulared genes for Europe sample.

In the above given graph the gene IDs that above 0, that genes are upregulated genes. The genes IDs that are below 0, that are downregulated genes.

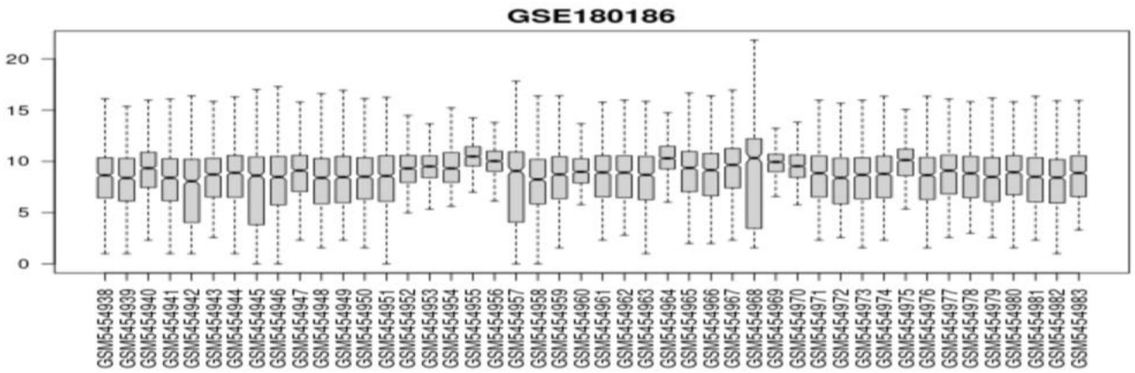

Figure2- boxplot of different genes in Europe sample.

For the USA sample-

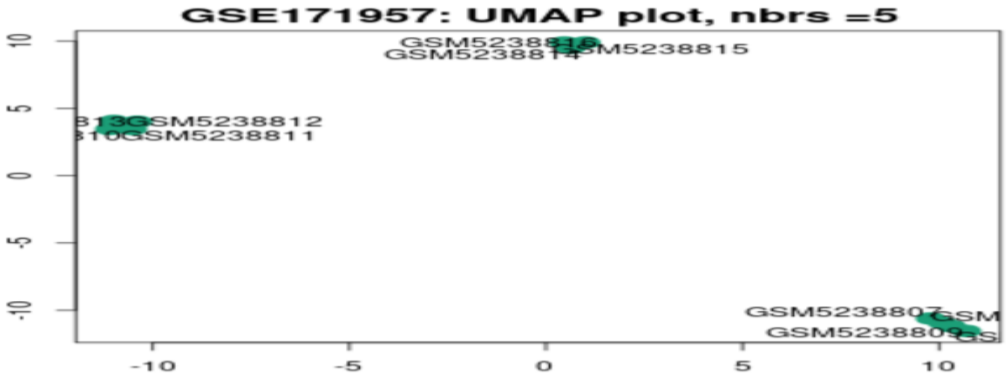

Figure3- shows upregulated and downregulared genes for USA sample.

In the above given graph the gene IDs that above 0, that genes are upregulated genes. The genes IDs that are below 0, that are downregulated genes.

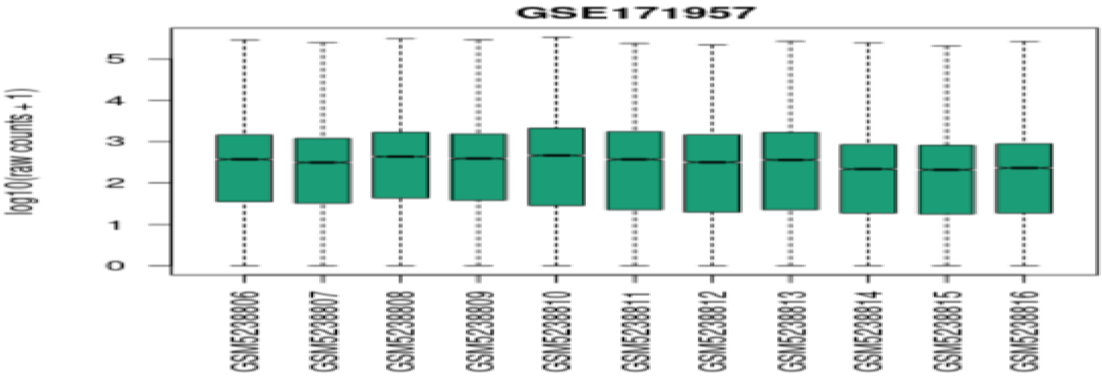

Figure4- boxplot of different genes in USA sample.

Conclusion-

The above analysis shows the gene expression analysis of breast cancer of two different population. This analysis shows the genes which are upregulated and downregulated in the breast cancer in two different population.

Code

The scripts for the above pipeline are available at https://github.com/DevanshiGupta481/HB_SUB/blob/a762cfdf4016a1349790cf893fe8d15d0b70d806/Final_project_code