Alexandria University
Faculty of Engineering

Bioinformatics
Spring 2024
Lab 7
Assigned: 25/4/2024
Due: 3/5/2024

Lab 7: Gene expression patterns in human liver cancers (60 Points)

Objective

To analyze microarray data comprehensively, preprocess the data, conduct principal component analysis (PCA), apply regression analysis, perform clustering, and utilize classification techniques to distinguish between cancer and non-cancer samples.

Requirements

- You are obligated to attempt all tasks
- You **MUST** use R Markdown to submit your report
- State your name and ID at the first cell in your markdown report
- For all the tasks, you should use the complete BrainCancer Dataset and not the minimized version

Libraries

```
library(tidyverse)
library(ggplot2)
library(limma)
library(clValid)
library(scatterplot3d)
library(e1071)
library(gridExtra)
library(caret)
```

Part 1: Data Wrangling (25 points)

Task 1.1: Data Acquisition (5 Points)

- Read the dataset CSV file into R. For the remainder of this lab, you are supposed to work with this dataframe and it will be referred to as ‘your dataframe’.
- Notes about your data:
    - Make sure you have 130 samples x 54673 genes
    - You have two extra columns, one for the phenotype and one for the sample number
    - Notice the orientation of your dataframe where rows are samples and columns are genes
- Remove the sample id column and put it as rownames. Use the rownames() function
- Extract the expression data into a separate dataframe which will be called expression.data
- By this point you should have two dataframes one with expression data only and the second one with expression data + phenotypes.

Task 1.2: PCA Before QC (5 Points)

- Remove NAs by filling them with the means of their respective genes. (mean of gene across all samples)
- Compute PCA using prcomp function. You are not allowed to use any other alternatives like princomp.
- You should think about how you should do your PCA. Should you use the genes as rows or as columns? How does the transformation matrix of each differ? **Comment** your findings.
- For the remainder of this lab, perform PCA using prcomp and the genes as columns. You should have a transformation matrix 130x130. We will work with this matrix so do not use the predict function.
- Make a dataframe of your principal components. This dataframe will be referred to as pcs.
- Use the head function to view the data of your pcs.

Task 1.3: PCA Before QC [Visualization] (5 Points)

- Create plots comparing PC1 vs PC2, PC1 vs PC3, and PC2 vs PC3.
- You must use ggplot for each plot. Color according to the phenotype of each sample
- You must view the three plots together in a figure of size 15x15. Use grid.arrange function to place them on 3 rows and 1 column. Think about how to change the figure size in markdown.

Task 1.4: Data Cleaning (5 Points)

- For each gene, check for outliers (above/below three standard deviations)
- If outliers are present, remove them and place NAs instead.
- Remove NAs by filling them with the means of their respective genes. (mean of gene across all samples)
- Make sure that the data is in the right orientation (samples as rows and genes as columns)
- Perform normalization between genes using the quantile method. Use the normalizeBetweenArrays() function. **Comment** on what is this function and what the quantile method means.

Task 1.5: Data Inspection (5 Points)

- Repeat tasks 1.2 and 1.3 but this is for the data after being cleaned.
- **Commen**t on changes in data before and after QC

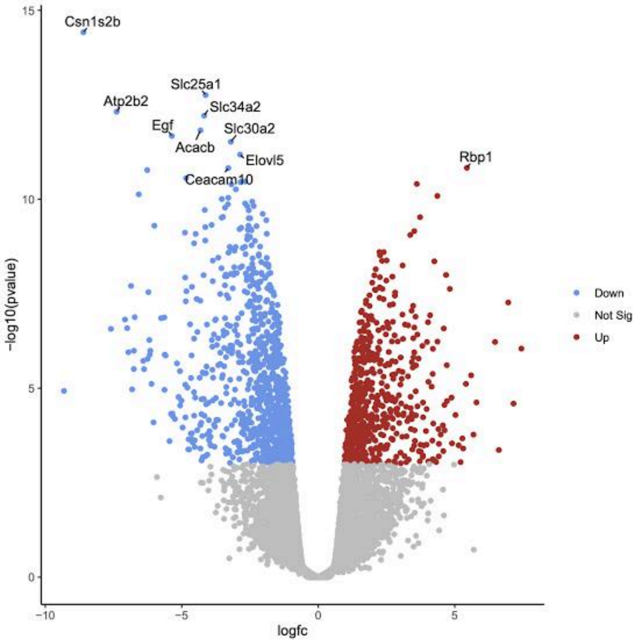
Part 2: Analysis (20 points)

Task 2.1: Regression Analysis (10 points)

- Perform logistic regression analysis using glm function and family binomial
- You have to encode your phenotype before running the regression where positive for Tumor and negative for Normal.
- Using only PC1 vs PC2 plots, check if there is any unexplained grouping by the PCA (after QC steps). Determine the PC that shows this variation and select it as a covariate in the model
- Run the regression model for class of tumor (positive/negative) against the gene expression: Class ~ Gene_i. There are almost 54000 genes, so specify only the 5000 genes found in this file [5000 genes]
- Determine all the significant Genes. (Use p-value threshold = 0.05)

Task 2.2: Visualization (10 points)

- Plot the heatmap (using heatmap function in R or ggplot2) of the gene expression data of the top 20 significant genes
- Plot the volcano plot (-log10(p-value) on the y-axis and log fold change on x =-axis), and color the top 20 significant genes on that plot (using red for upregulation (cancer higher than control) and green for downregulation (opposite))



Part 4: Annotation (10 points)

- In the brain cancer dataset, all genes are in the Affymetrix format
- Use david tools to convert the top 20 significant gene names to normal gene names.
- Use the EnrichR tool to get enrichment results of which pathways the top 20 significant genes are present in.
- You are requested to extract Kegg pathways annotation in tables and graphs produced by EnrichR. Comment on the results based on the enrichment analysis p-values

Reporting (5 points)

- You are required to write a report including the commands, and file formats and include any screenshots of the output of such commands

References

- BrainCancer Dataset
- About the BrainCancer Dataset
- ggplot2 Documentation
- Grid.arrange
- David Tools
- EnrichR

- Markdown Cheatsheet