The report of Breast Cancer Recurrence Dataset Analysis

**Group: M**

XU JIAQI, Student No.: MB955210

CHEN TIANTIAN, Student No.: MB955114

**Data Source:**

GEO(Gene Expression Ominibus) database

Data link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32603

**Dataset:**

35069 gene expression profile of 248 breast cancer patients

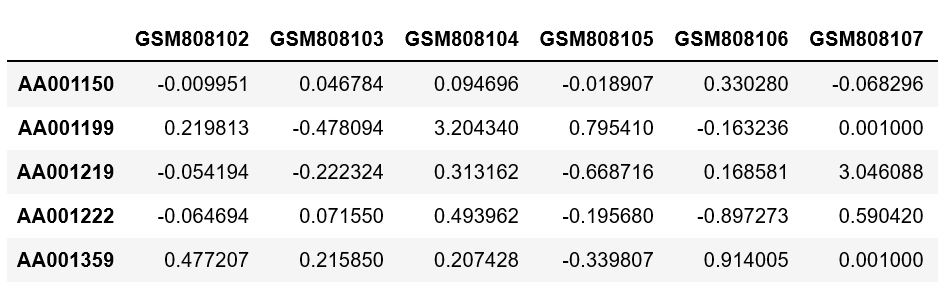
**Topic**:

Breast cancer is the leading cause of cancer-related deaths in women worldwide. Recurrence is major cause of death with poor treatment efficacy. Despite advances in early detection and comprehensive treatments for breast cancer, approximately 30% of patients with early-stage breast cancer still experience recurrent disease

From patients information, we can figure out patients recurrence information, then we tried to do different expression gene (DEG) research on the patients with breast cancer recurrence and without breast cancer recurrence

**Data Structure:**

The dataset is a data frame with 35069 rows(gene names) and 248 columns(patients). This dataset is already normalized.



**Data analysis:**

The dataset(No.GSE32603) includes over 35,000 gene expression of 248 breast cancer patients, in which 169 patients without recurrence and 79 patients with recurrence. We used numpy, pandas, os, matplotlib, seaborn and scipy for analysis.

Data loading and cleaning

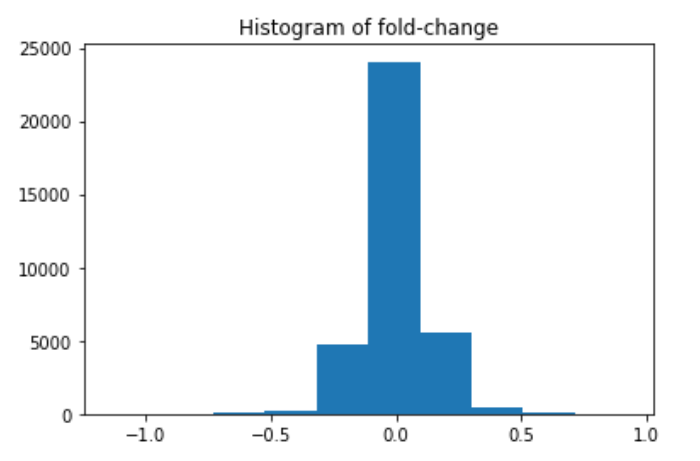
We first load the data sheet into Jupyter Notebook. Then use replace() command to We replace NA to a very small number 0.001. Then we made a box plot to see the gene expression distribution of the 248 samples.

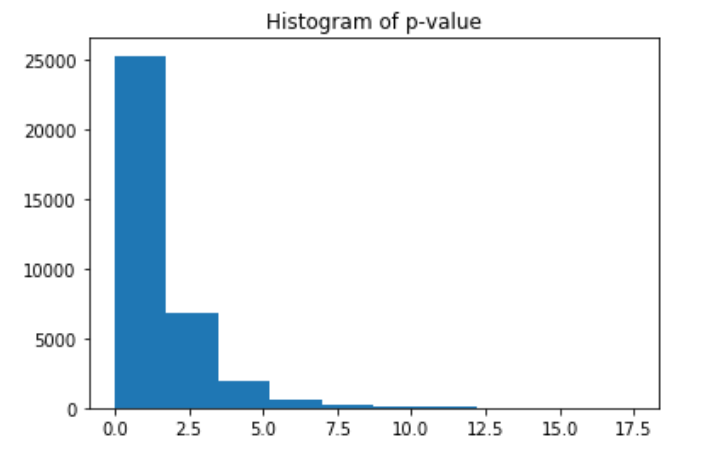
Patient data loading

We load the patient data sheet, which includes the group information (breast cancer recurrence and non-breast cancer recurrence). We use groupby() to identify the group number in the gene expression matrix.

Fold change and p-value histogram

We got the mean gene expression of every sample in two groups, then define the difference of means on each gene as fold change. Then we made a histogram of the fold change to see the distribution of the fold change. We also did T-test on these two groups and made a histogram of T test p-value. From these two graphs, we can find that the fold changes between these two groups are mainly between -0.5-0.5, which indicates that the fold changes of these two groups are not different very much.

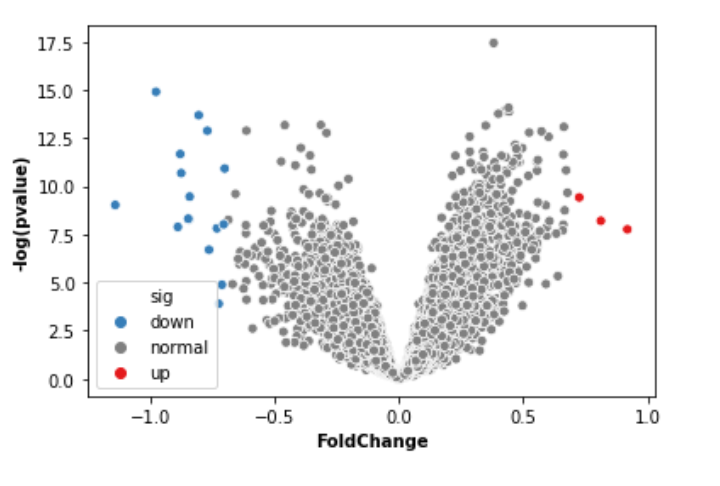


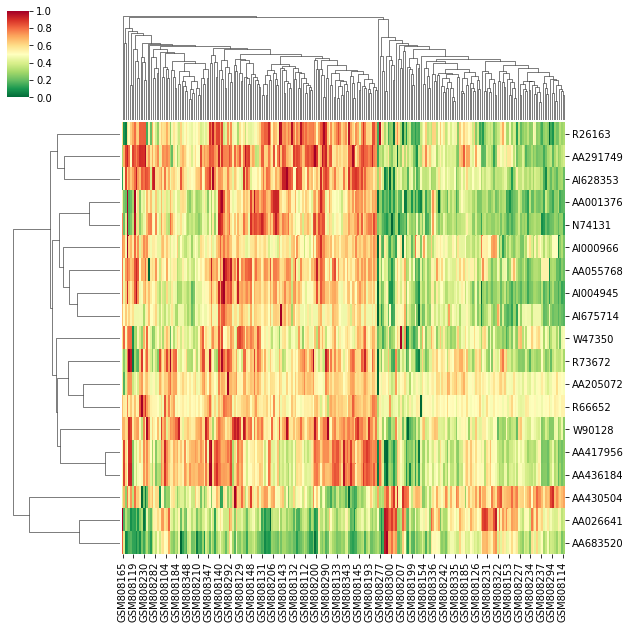


Volcano plot and Heat map

In DEG analysis, volcano plot is often used. Volcano plot is a type of scatter-plot that is used to quickly identify changes in large data sets composed of replicate data. It plots significance versus fold-change on the y and x axes, respectively. We set the threshold of fold change at 0.7 and p-value less than 0.05, the X axis is fold change and Y axis is –log(p-value), got volcano plot and find 19 different genes.

A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. We use heat map to represent the differences of these two groups.

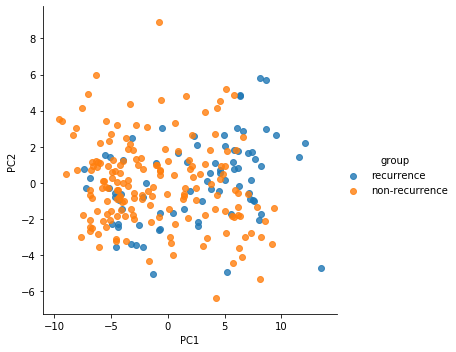




Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is a common method to reduce the dimensionality. It is usually done before a machine learning algorithm. PCA use fewest variables to explain the whole data set, giving convenient access to machine learning. First is scaling the data, calculate the mean of features of every sample, then minus mean on every sample. Next step is the covariance matrix on the gene expression. After we got the covariance matrix, we can do dimension reduction by finding feature values and vectors to figure out the dataset, then we calculate the rate of contribution on each feature vector. After sorting contribution rate, fine the features with the biggest contribution rates. The last, we can reduce the dimension of dataset. PCA is an unsupervised learning method and is similar to clustering. It finds patterns without reference to prior knowledge about whether the samples come from different treatment groups or have phenotypic differences.

In python, we can use sklearn package to do PCA analysis. We can use pcafit() to do dimension reduction after preparing the data. Then we use lmplot() to visualize the data. In this dataset we use, the PCA results is not very obvious, which means in the dataset we use, the gene expression of patients with breast cancer recurrence is similar with the gene expression of patients with non-breast cancer recurrence.



Next step

For further research, we can annotate these genes and find the related pathways, trying to explain breast cancer recurrence on gene level.

**Conclusions：**

Through comparing these two groups，We finally found 19 DEG between breast cancer recurrence and non-recurrence patients. Although the results of the whole research don’t indicate that both of these two gene patterns of breast cancer patients exist any obvious difference, we still think this experience gives us a angle and some research methods to explain and explore our topic, which makes us learn more skills to complete our project. It’s very beneficial to us. Maybe the answer of breast cancer recurrence hides among these 19 DEG and we need to explore more.

**Challenges：**

First, it is difficult to find the right data to totally match our topics and even though the data is relative to our topic，its size is too small to meet the requirements. So we spent some time to find our data. And the second challenge is that the raw data need to be cleaned and classified. The largest problem is that we need to learn more new libraries and meanwhile write our codes. Also, what is regrettable is that our results are not very significant.