# GSE124483 Analysis

A0251614M
Guokai Chen

## 1.My R script require following preparation

1. Download GSE124483 raw data. Then extract to work directory.
   (1) Download by (http)

| Supplementary file | Size | Download | File type/resource |
|---|---|---|---|
| GSE124483_RAW.tar | 236.7 Mb | (http)(custom) | TAR (of CEL, CHP) |

*Raw data provided as supplementary file*

*Processed data provided as supplementary file*

   (2) get work directory.

```
> getwd()
[1] "D:/bL5631/CA1.2/CA1.2ver"
```

   (3) Extract to work directory.


2. Install “pacman” package.

```
install.packages("pacman")
library(pacman)
```

3. Start to run R script.

## 2. Workflow and process method

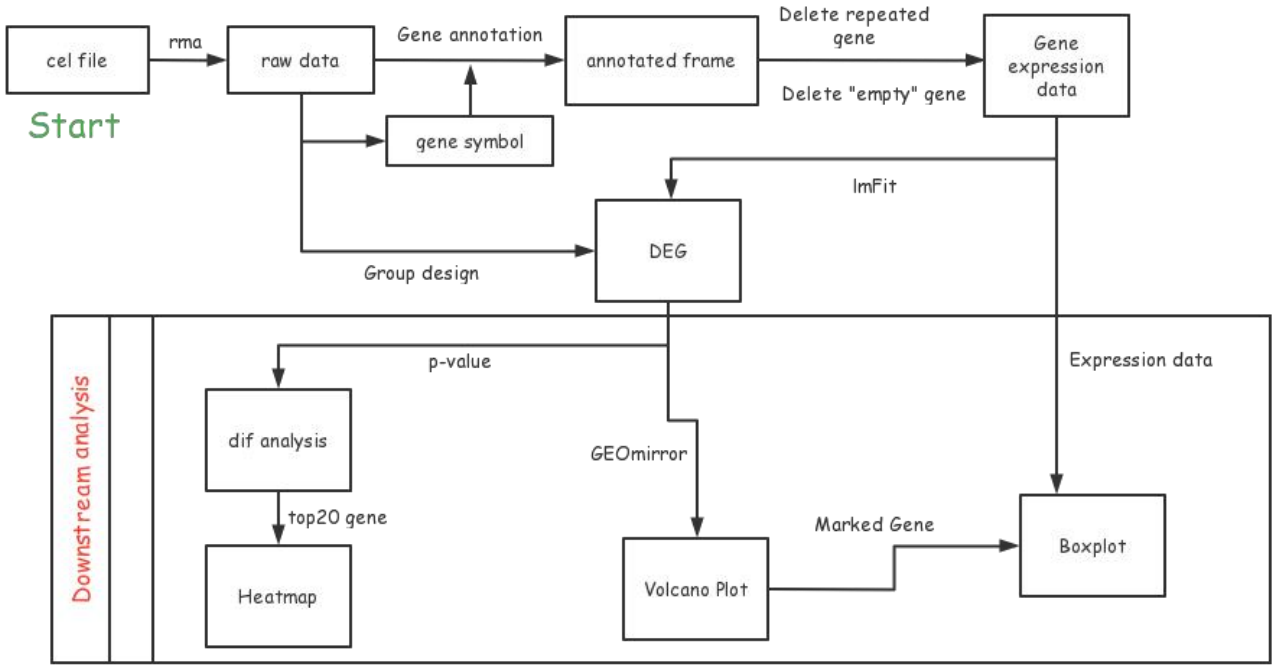
Started from cel file using rma function to normalize and get Biobase file.

Extract expression form by probe and probe annotation information to get the primary frame by gene name. After delete repeated gene and “empty” gene which means no gene responds to the probe to get gene expression data.

Then processing lmfit I founded differentially expression gene form. Then I did some downstream analysis.

First, I exclude gene whose p-value is higher than 0.05 and logFC is higher than 1 to present the genes with high reliability. Then, I chose top20 gene with the least p-value to make heatmap.

Next, I used GEOmirror and AnnoProbe package to draw volcano plot. There package provides 2 types of volcano plot. The first one provides information of number of up&down gene. The second one provides name of gene highly differentially expressed. And by the gene name from second volcano plot, I could go back to gene expression data to check the gene expression level of this “special” gene using boxplot.




## 3. Output: Data frame and Image

My output include:

| Variate Name | Description |
|---|---|
| exprdf | gene expression data according by probe |
| jiyinbiaodaliangbiao | gene expression data frame |
| Group | experiment group design |
| dif | differentially express gene analysis (p<0.05) |
| DEG | differentially express gene analysis (whole) |
| retu | heatmap of top20 p-value gene |
| huoshantu1 | volcano plot with gene number of up&down |
| huoshantu2 | volcano plot of whole gene with gene symbol |

| | logFC | AveExpr | t | P.Value | adj.P.Val | B | group |
|---|---|---|---|---|---|---|---|
| Uqcrq | -2.7812 | 7.8885 | -15.658 | 1.537e-09 | 6.935e-05 | -1.901 | down |
| Gm16043 | 0.9997 | 1.0585 | 7.656 | 4.812e-06 | 9.318e-02 | -2.371 | no_change |
| Syk | -1.7361 | 1.7165 | -7.472 | 6.195e-06 | 9.318e-02 | -2.397 | down |
| Gm20604 | -1.9015 | 4.5996 | -6.292 | 3.441e-05 | 3.881e-01 | -2.597 | down |
| Kdm4a | -1.5838 | 2.3097 | -5.497 | 1.216e-04 | 6.283e-01 | -2.777 | down |
| Gm2a | -1.8722 | 4.8346 | -5.332 | 1.597e-04 | 6.283e-01 | -2.820 | down |
| 4930412F09Rik | -1.3029 | 3.3710 | -5.263 | 1.795e-04 | 6.283e-01 | -2.839 | down |
| Efhd2 | -1.2473 | 3.4025 | -5.168 | 2.103e-04 | 6.283e-01 | -2.865 | down |

Fig.1. DEG (whole). top8 rows.

| | logFC | AveExpr | t | P.Value | adj.P.Val | B | group |
|---|---|---|---|---|---|---|---|
| Uqcrq | -2.781 | 7.8885 | -15.658 | 1.537e-09 | 6.935e-05 | -1.901 | down |
| Syk | -1.736 | 1.7165 | -7.472 | 6.195e-06 | 9.318e-02 | -2.397 | down |
| Gm20604 | -1.902 | 4.5996 | -6.292 | 3.441e-05 | 3.881e-01 | -2.597 | down |
| Kdm4a | -1.584 | 2.3097 | -5.497 | 1.216e-04 | 6.283e-01 | -2.777 | down |
| Gm2a | -1.872 | 4.8346 | -5.332 | 1.597e-04 | 6.283e-01 | -2.820 | down |
| 4930412F09Rik | -1.303 | 3.3710 | -5.263 | 1.795e-04 | 6.283e-01 | -2.839 | down |
| Efhd2 | -1.247 | 3.4025 | -5.168 | 2.103e-04 | 6.283e-01 | -2.865 | down |

Fig.2. dif(P.Value<0.05&logFC>1). top8 rows.

After normalizing the raw data and linear fitting, I had fig1 to show my gene expression data. But the more reliability, I selected data that fold change > 1 to define as difference. Meanwhile, excluded data that p-value > 0.05. So I got data frame in fig2 as reliable data to process.

From this step, I can choose to see the data with high fold change to check how their difference dramatically or the see the expression pattern of the most stable data. I gave my first curiosity to the data with low p-value. So I choose the 20 gene with the least p-value. Making heatmap as below, from fig3, I could see the overall expression pattern of them. Generally, gene of knocking out animal will downregulate due to the pathway of metabolism blocked. So I got interested in the upregulated gene shown on top 6 rows. They might be the alternative pathway of NLRP3 knocking out animal to supplement the damage of pathway. Besides, there is a very red gene named Uqcrq showing highly downregulated. It will be discussed later because it also stands out in volcano plot.

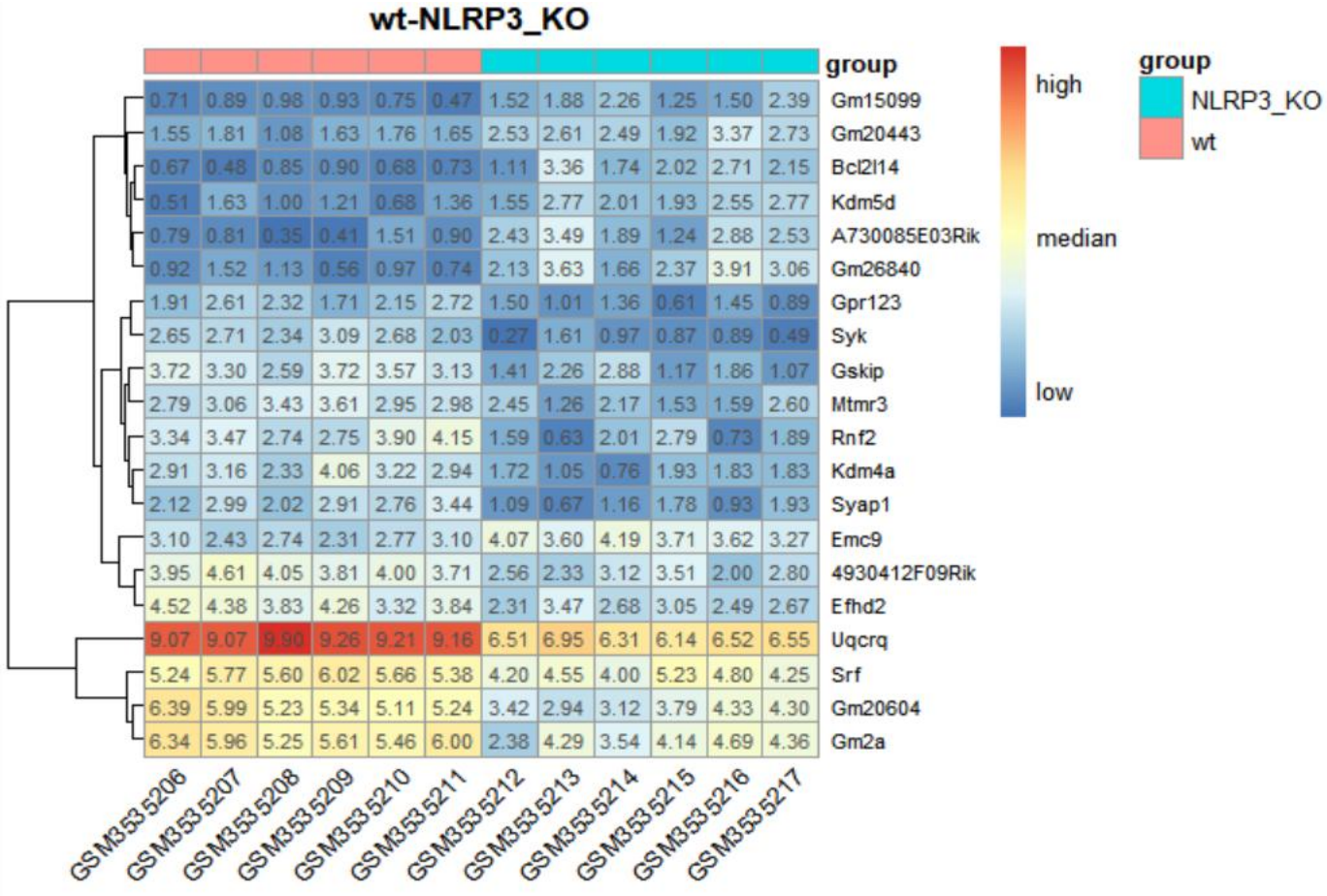



Fig.3.retu. Heatmap of least p-value gene.

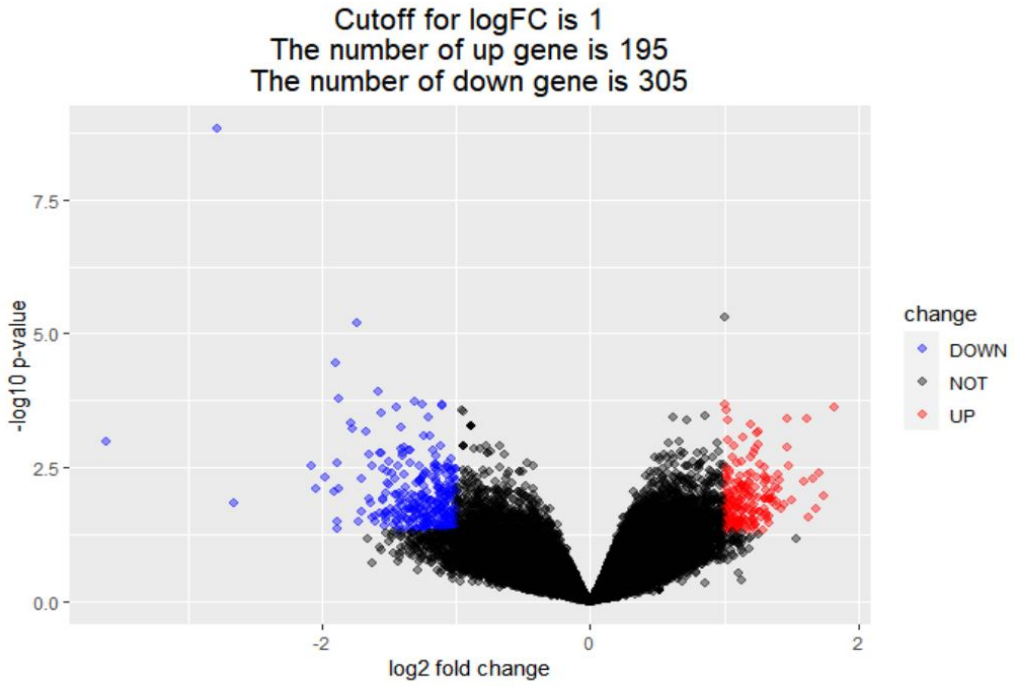



Fig.4.huoshantu1. Volcano plot showing the number of up&down gene.

In order to throw a light on the analysis, I applied a new package AnnoProbe to draw the volcano plot. From the first volcano image fig4, the title shows the number of up regulated gene 195 and downregulated gene 305. Apparently, most related gene was downregulated but this image can only give abstract observation of number of gene regulated. For more detailed information, I used the second volcano plot. It shows the name of genes with high change fold and high -log10 p-value. To my surprise, there is a very special gene Uqcrq. This might shows the close relationship with NLRP3 gene function or pathway in samples.

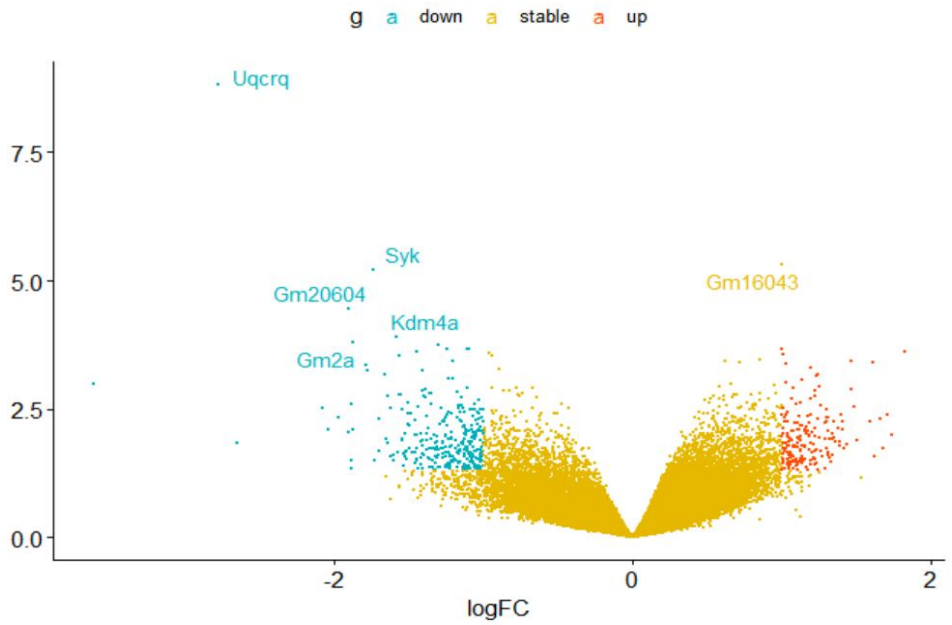



Fig.5.huoshantu2. Volcano plot showing name of the most “obvious” gene

Now we have already acquired the gene of interest and its name. I decided to go back to the original data of its gene expression level. Thus, I printed the boxplot. From this boxplot, the the data range of each group appears to be narrow and stable. That is to say, the analysis of our data processing is reliable. The same. We can apply this plot method to any gene we are interested in.

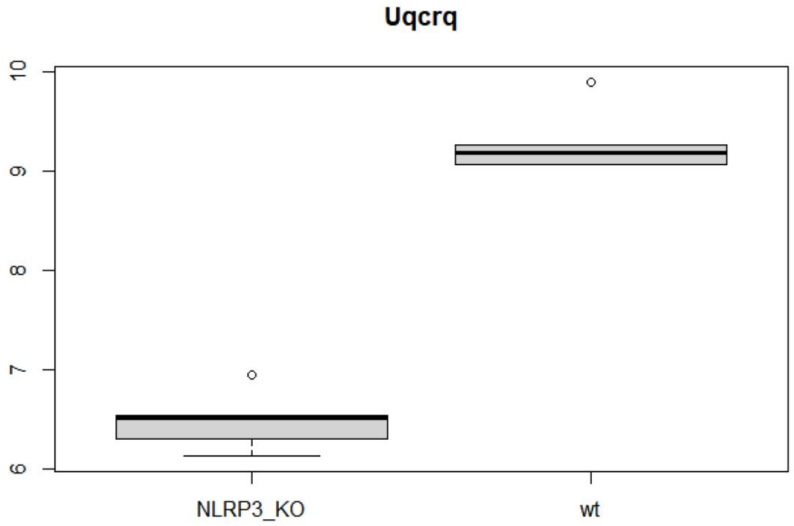



Fig.6.Boxplot of Uqcrq gene expression.

## 4. Summary

From the report, a workflow in identifying differentially expressed gene was founded.

This method can also apply to any other GSE biobase, but there are some points need to be noticed.

1. I applied function:”getNetAffx” in my workflow. But this function provided by oligo package can only be used in Exon ST and Gene ST microarray. For other data, user might need to find other way to get feature data.
2. In DEG, I defined group with p-value above 0.05 as “no change”. This threshold can be varied according to research requirement.
3. In the process of deleting repeated gene, I choose the method that only maintain the first one. It depends. There are some other method. First, choose gene symbol with the highest expression. Second, choose the mean expression of all repeated gene. Third. randomly choose one. Limited by complexity, I choose to keep the first one.
4. For some probe corresponding multiple genes for overlapping of genes, I choose the first symbol appear. In real research, we can screening gene symbol annotation column to get the coding gene only. Then keep gene functions rather than other kinds of genes.