GSE124483 Analysis

A0251614M
Guokai Chen

1. Study Design

Based on the summary of data set GSE124483 from platform GPL21877, this data set was described as the investigation of the role of NLRP3 in cardiac aging.

By knocking out of NLRP3, it states of inhibition of PI3K/AKT/MTOR pathway, increasing SIRT1 protein expression, reducing IGF-1 and leptin/adiponectin ratio levels. To check its conclusion, I choose to verify SIRT1 expression as a sample to clarify the function of NLRP3.

Beyond the original experiment, my report will check the up- and down- regulated gene by the contrast between young and old, NLRP3_KO and wild type mice to show the potential target gene of NLRP3 using volcano plot image. Besides that, according to the gene symbol provided by image, user can see how much difference in gene expression pattern and the number relation.

In a nutshell, this report will show the partial difference of gene expression pattern using heatmap, the number and fold change of differentially expressed gene using volcano image. According to the gene information provided by above image, I can also go back to raw data to double-check the gene expression difference.

Through above work, this report and code provided the visual form of gene expression difference and potential target gene group related to NLRP3 inflammasome to explore possible research target.

2. Preparation and description

1. File Description: Inside the CA1_ChenGuokai.zip file, there are 4 kinds of file.

   The R script(CA1_ChenGuokai) is used to reproduce this report result.

   The image files are the output of R script including heatmap, volcano image and boxplot.

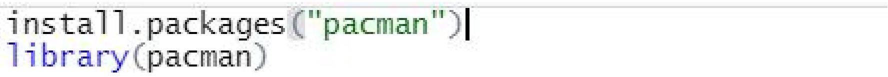
   The folder(Contrast_of_Age) is the republication of our study. But the contrast is changed into the constrast between old and young.

   The folder(GSE124483_RAW) is the cel file we will use.

2. Preparation before running R

   (1) Download zip file from github. Then extract it and set “CA1_ChenGuokai” as working directory. Also, users can download GSE124483 raw data from GEO by themselves via accession number GSE124483. Just make sure the folder with cel file is under the working directory with the name “GSE124483_RAW”.

   (2) Install “pacman” package as the following picture. Make sure first 26 rows code can run .



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

   (3) BiocManager and pacman are the most important packages. Make sure these two work.

3. Workflow and process method

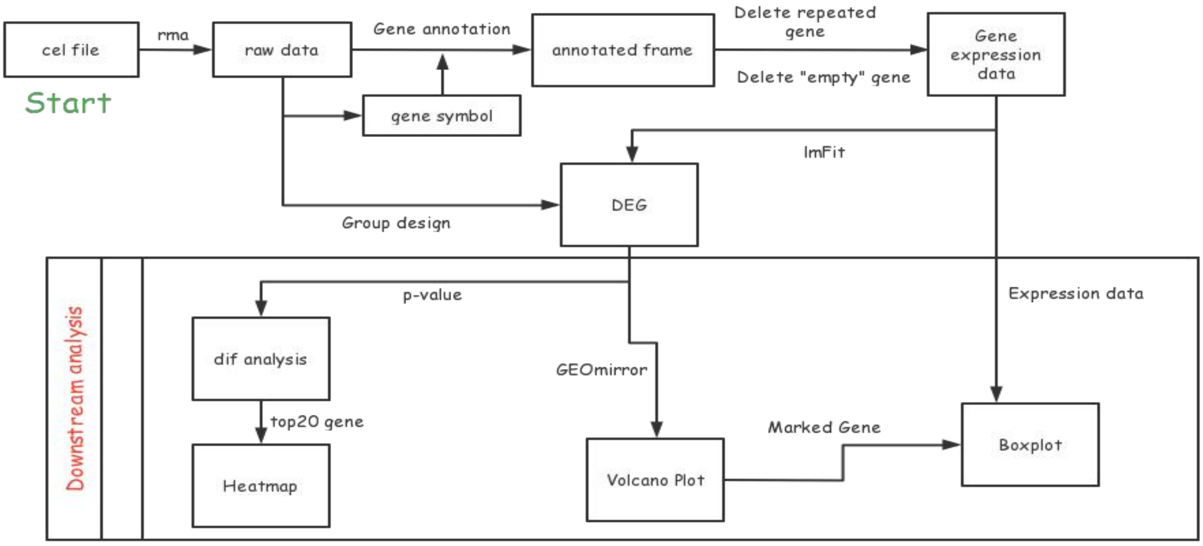
Started from cel file using rma function to normalize and get Biobase file.

Extract expression form by probe and probe annotation information to get the primary frame by gene name. After deleting “empty” gene which means no gene responds to the probe and repeated gene to get gene expression data frame. I need to mention the process method here is to keep the first repeated gene using logical flag. Also, we can use function “aggregate” to get the mean value of gene expression. Then through linear fit, 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 whose absolute value of logFC is lower than 1 to present the genes with high reliability and greater fold change. In other words, meaningful and reliable data. Then, I chose top20 gene with the least p-value to draft the 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 regulated 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.

Using the same workflow, I redesigned the contrast with young and old to explore the different phenomenon.

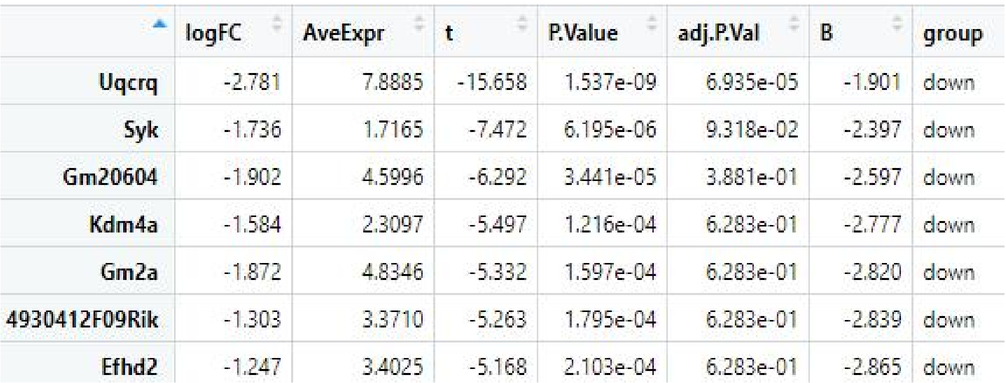


4. Output: Data frame and Image

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)

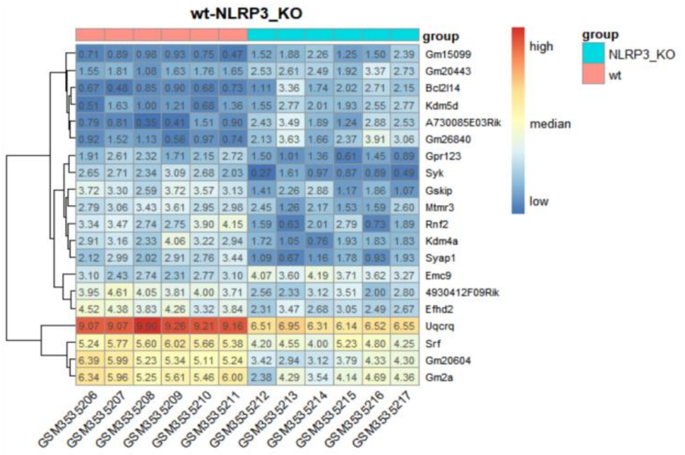
retu	heatmap of the lowest 20 p-value genes
huoshantu1	volcano plot with gene number of up&down
huoshantu2	volcano plot of whole gene with gene symbol



	logFC	AveExpr	t	P.Value	adj.P.Val	B	group
Uqcrq	-2.781	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.1. dif(P.Value<0.05&|logFC|>1). Top8 rows.

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



GSM3535206	GSM3535207	GSM3535208	GSM3535209	GSM3535210	GSM3535211	GSM3535212	GSM3535213	GSM3535214	GSM3535215	GSM3535216	GSM3535217	
0.71	0.89	0.98	0.93	0.75	0.47	1.52	1.88	2.26	1.25	1.50	2.39	Gm15099
1.55	1.81	1.08	1.63	1.76	1.65	2.53	2.61	2.49	1.92	3.37	2.73	Gm20443
0.67	0.48	0.85	0.90	0.68	0.73	1.11	3.36	1.74	2.02	2.71	2.15	Bcl2l14
0.51	1.63	1.00	1.21	0.68	1.36	1.55	2.77	2.01	1.93	2.55	2.77	Kdm5d
0.79	0.81	0.35	0.41	1.51	0.90	2.43	3.49	1.89	1.24	2.88	2.53	A730085E03Rik
0.92	1.52	1.13	0.56	0.97	0.74	2.13	3.63	1.66	2.37	3.91	3.06	Gm26840
1.91	2.61	2.32	1.71	2.15	2.72	1.50	1.01	1.36	0.61	1.45	0.89	Gpr123
2.65	2.71	2.34	3.09	2.68	2.03	0.27	1.61	0.97	0.87	0.89	0.49	Syk
3.72	3.30	2.59	3.72	3.57	3.13	1.41	2.26	2.88	1.17	1.86	1.07	Gskip
2.79	3.06	3.43	3.61	2.95	2.98	2.45	1.26	2.17	1.53	1.59	2.60	Mtmr3
3.34	3.47	2.74	2.75	3.90	4.15	1.59	0.63	2.01	2.79	0.73	1.89	Rnf2
2.91	3.16	2.33	4.06	3.22	2.94	1.72	1.05	0.76	1.93	1.83	1.83	Kdm4a
2.12	2.99	2.02	2.91	2.76	3.44	1.09	0.67	1.16	1.78	0.93	1.93	Syap1
3.10	2.43	2.74	2.31	2.77	3.10	4.07	3.60	4.19	3.71	3.62	3.27	Emc9
3.95	4.61	4.05	3.81	4.00	3.71	2.56	2.33	3.12	3.51	2.00	2.80	4930412F09Rik
4.52	4.38	3.83	4.26	3.32	3.84	2.31	3.47	2.68	3.05	2.49	2.67	Efhd2
9.07	9.07	9.90	9.26	9.21	9.16	6.51	6.95	6.31	6.14	6.52	6.55	Uqcrq
5.24	5.77	5.60	6.02	5.66	5.38	4.20	4.55	4.00	5.23	4.80	4.25	Srf
6.39	5.99	5.23	5.34	5.11	5.24	3.42	2.94	3.12	3.79	4.33	4.30	Gm20604
6.34	5.96	5.25	5.61	5.46	6.00	2.38	4.29	3.54	4.14	4.69	4.36	Gm2a

Fig.2.retu. Heatmap of least p-value gene.

From this step, I can choose to see the data with high fold change to explore the gene differentially expressed 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 chose the 20 gene with the least p-value. Making heatmap as below, from fig2, 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 NLRP3 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.

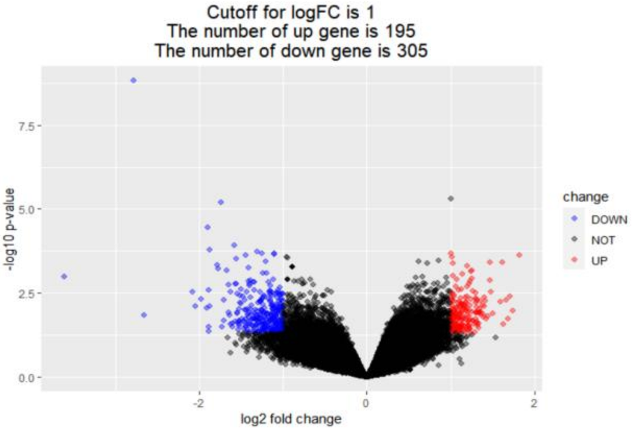


Fig.3.huoshantu1. Volcano plot showing the number of up&down gene.

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. These genes consist of the set related to gene NLRP3, which can provide theoretic support for following research about NLRP3 related pathways. 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.

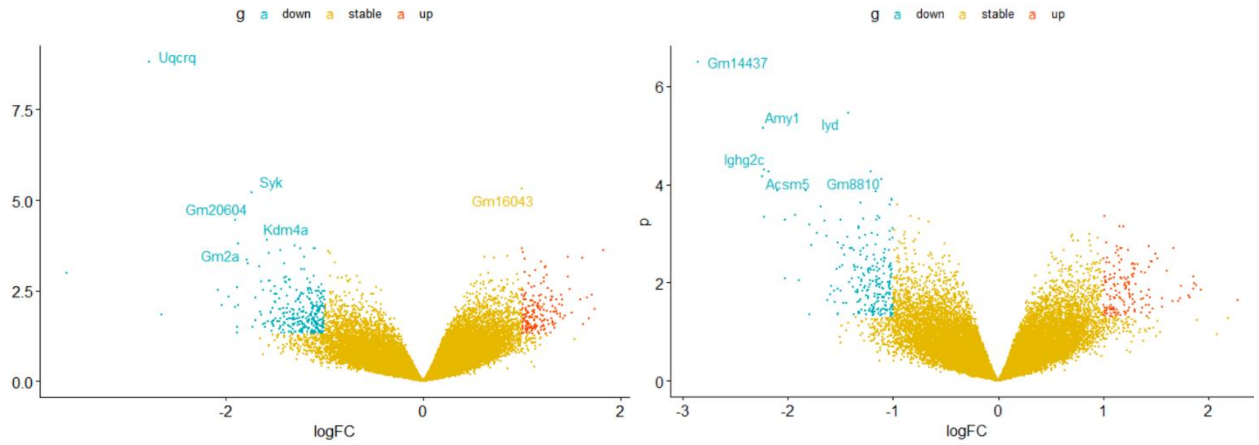


Fig.4.huoshantu2. Volcano image. Left one for genotype. Right one for age.

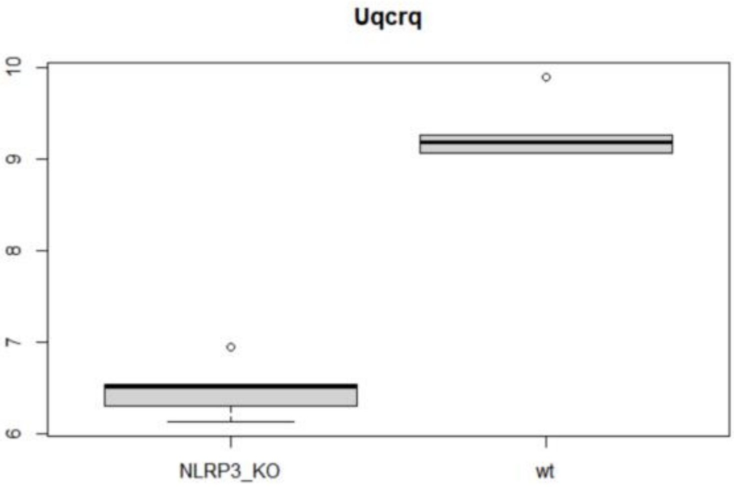


Fig.5.Boxplot of Uqcrq gene expression.

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.

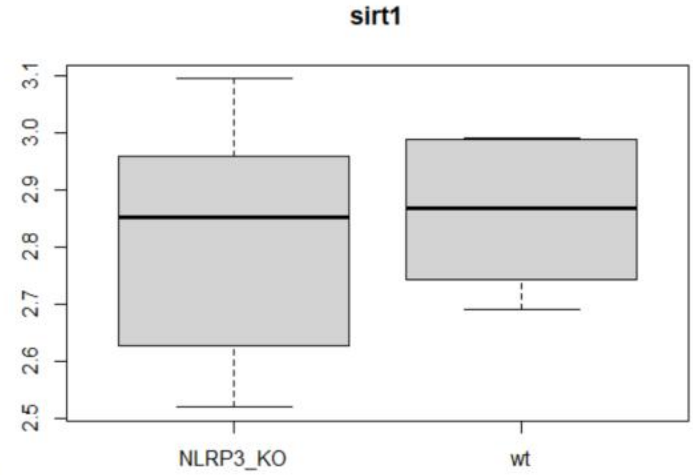


Fig.6. Boxplot of SIRT1

Then, when I go back to check the SIRT1 expression, I was surprised the no change shown in its expression which is totally against the description of this data set. This also provides a rising question. Is there another pathway NLRP3 uses to suppress the translation of SIRT1 protein? So our study also provides an interesting phenomenon which can make us have more profound notion of NLRP3 function.

5. 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. 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 coding gene rather than other kinds of genes like intron.

All in all, there are several advantage of our workflow.

1. By volcano plot image, we can see the whole pattern of gene regulated. Make the data frame visualized with overall image.
2. By heatmap, we can see the variance of gene. Comparing to volcano image, we can search for gene name in the image which facilitates us to trace back.
3. By my designed workflow, I retain the important variance during the process to increase the operability of my code. Uqcrq and SIRT1 just perform as my example to use my data.

6. Reference

getNetAffx function: https://blog.csdn.net/tommyhechina/article/details/80291987
gene symnol annotation strategy: http://www.bio-info-trainee.com/1586.html
Volcano image function resource: https://github.com/jmzeng1314